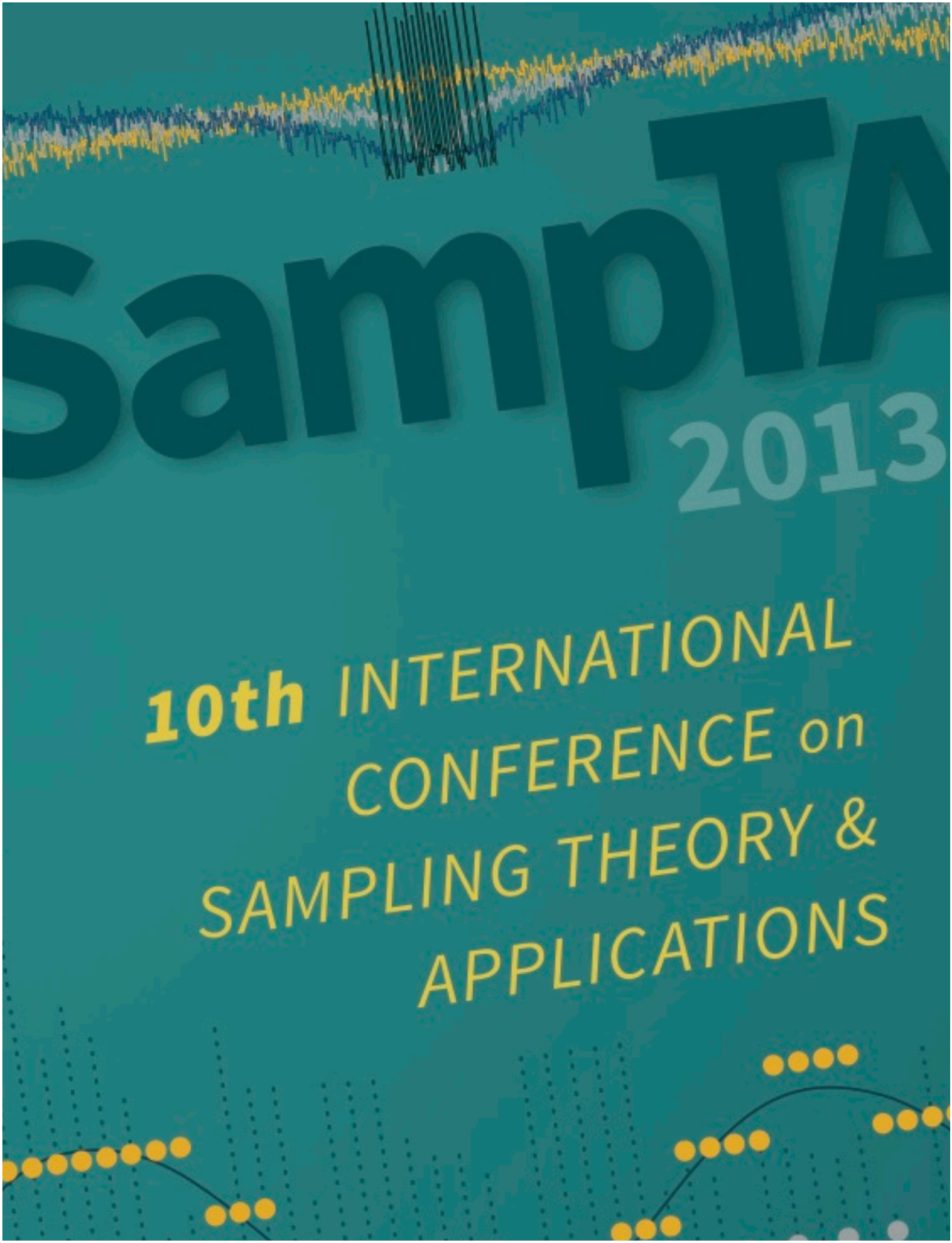

SampTA
2013
10th INTERNATIONAL
CONFERENCE on
SAMPLING THEORY &
APPLICATIONS

Foreword of the Proceedings of the 10th International Conference on Sampling Theory and Applications

The 10th International Conference on Sampling Theory and Applications (SampTA) took place from July 1 till July 5, 2013, in Bremen, Germany. SampTA'13 brought together mathematicians and engineers interested in sampling theory and its applications to related fields (such as signal and image processing, coding theory, control theory, complex analysis, harmonic analysis, differential equations) to exchange recent advances and to discuss open problems. Paper submissions were invited on any aspect of sampling theory and applications.

The renewed interest in sampling theory, in parts due to the emergence of Compressive Sensing, lead to increased numbers of paper submissions and registrations. All in all, 230 electrical engineers and mathematicians participated at SampTA'13 and we were happy to house them in the Guest House and in a new Residential College of Jacobs University, a fact that greatly enhanced the community spirit that SampTA conferences are known for.

As organizers we were very pleased with the outstanding scientific quality of the conference! This was due in part to the excellent plenary talks:

Robust subspace clustering	Emmanuel Candes
How to best sample a solution manifold?	Wolfgang Dahmen
Sampling theory and applications: developments in the last 20 years and future perspectives	Hans-Georg Feichtinger
Fast algorithms for sparse Fourier transform	Piotr Indyk
Seeing the invisible; predicting the unexpected	Michal Irani
Signal recognition and filter identification	Nikolai Nikolskii
Stemming the neural data deluge	Jan Rabaey
Event-driven sampling and continuous-time digital signal processing	Yannis Tsividis
Sampling and high-dimensional convex geometry	Roman Vershynin

The scientific quality of SampTA'13 is best illustrated through the papers that are published within these proceedings. They represent talks given in the following special sessions (with the invited session organizer listed on the right):

Advances in Compressive Sensing	Holger Rauhut and Joel Tropp
Circuit Design For Analog to Digital Converters	Yun Chiu
Finite Rate of Innovation	Chandra Seelamantula
Optical and RF Systems	Michael Gehm and Nathan Goodman
Sampling and Frame Theory	Peter Casazza, Bernhard Bodmann, and Matthew Fickus
Sampling and Geometry	Stephen Casey and Michael Robinson
Sampling and Learning	Albert Cohen
Sampling and Quantization	Holger Boche, Sinan Güntürk, and Özgür Yilmaz
Sampling for Imaging Science	Jalal Fadili and Gabriel Peyré
Sampling in Bio Imaging	Brigitte Forster, Hagai Kirshner, and Michael Unser
Super Resolution	Laurent Demanet

as well as posters and presentations that were organized into the following general sessions:

Compressed Sensing A and B
Time-Frequency Analysis
Harmonic Analysis A and B
Sampling of Bandlimited Functions A and B
Algorithms
Compressive Sensing and Applications
FFT and Related Algorithms

We would like to thank the SampTA'13 sponsors, namely, Jacobs University, DFG, IEEE Signal Processing Society, and EURASIP, and hope that you enjoy browsing through the SampTA'13 proceedings!

On behalf of the organization committees
Götz Pfander, General Chair Peter Oswald, Finance Chair Werner Henkel, Publications Chair

Committees

Local Organization

General Chair
Goetz Pfander (Jacobs University Bremen, Germany)

Finance Chair
Peter Oswald (Jacobs University Bremen, Germany)

Publication Chair
Werner Henkel (Jacobs University Bremen, Germany)

Members
Peter Maass (Bremen University)

Peter Massopust (Helmholtz Zentrum München, Germany)

Holger Rauhut (University of Bonn, Germany)

Technical Program Chairs

Yonina C. Eldar (Technion-Israel Institute of Technology, Israel)
Laurent Fesquet (TIMA, France)
Gitta Kutyniok (Technical University Berlin, Germany)
Pina Marziliano (Nanyang Technological University, Singapore)
Goetz Pfander (Jacobs University Bremen, Germany)
Bruno Torrésani (Aix-Marseille Université, France)

Steering Committee

Akram Aldroubi (Vanderbilt University, USA)
John Benedetto (University of Maryland, USA)
Paul Butzer (RWTH Aachen, Germany)
Hans Feichtinger (University of Vienna, Austria)
Paulo Ferreira (University of Aveiro, Portugal)
Karlheinz Groechenig (University of Vienna, Austria)
Rowland Higgins (Anglia Polytechnic University, Cambridge)
Abdul Jerri (Clarkson University)
Yuri Lyubarskii (Norwegian University of Science and Technology, Trondheim)
Farokh Marvasti (Sharif university of Technology, Iran)
Gerhard Schmeisser (Erlangen-Nürnberg University)
Bruno Torrésani (Aix-Marseille Université, France)
Michael Unser (EPFL, Switzerland)
Ahmed Zayed (DePaul University, USA)

Technical Program Committee

Technical Program Committee

Holger Boche	Technical University Munich	Germany
Bernhard Bodmann	University of Houston	USA
Pete Casazza	Unifversity of Missourri	USA
Stephen Casey	American University	USA
Yun Chiu	University of Texas at Dallas	USA
Albert Cohen	Universite Pierre et Marie Curie	France
Laurent Demanet	MIT	USA
Yonina Eldar	Technion-Israel Institute of Technology	Israel
Jalal Fadili	GREYC CNRS UMR 6072, ensicaen	France
Laurent Fesquet	TIMA	France
Matthew Fickus	AF Institute of Technology	USA
Brigitte Forster	University of Passau	Germany
Michael Gehm	University of Arizona	USA
Nathan Goodman	University of Oklahoma	USA
Sinan Gunturk	New York University	USA
Werner Henkel	Jacobs University Bremen	Germany
Hagai Kirshner	EPFL	Switzerland
Gitta Kutyniok	Technical University Berlin	Germany
Nicolas Marchand	GIPSA-lab	France
Pina Marziliano	Nanyang Technological University	Singapore
Peter Massopust	Helmholtz Zentrum München	Germany
Peter Oswald	Jacobs University Bremen	Germany
Gabriel Peyré	CNRS and Université Paris-Dauphine	France
Goetz Pfander	Jacobs University Bremen	Germany
Holger Rauhut	University of Bonn	Germany
Michael Robinson	American University	USA
Chandra Seelamantula	Indian Institute of Science	India
Bruno Torrésani	Aix-Marseille Université	France
Joel Tropp	California Insitute of Technology	USA
Michael Unser	EPFL	Switzerland
Ozgur Yilmaz	University of British Columbia	Canada

Table of Contents

10th international conference on Sampling Theory and Applications

Compressed Sensing

Overcoming the coherence barrier in compressed sensing
Ben Adcock (Purdue University, USA), Anders Hansen (University of Cambridge, USA), Clarice Poon (University of Cambridge, United Kingdom), Bogdan Roman (University of Cambridge, United Kingdom) 1

On construction and analysis of sparse matrices and expander graphs with applications to CS
Bubacarr Bah (École Polytechnique Fédérale de Lausanne (EPFL), Switzerland), Jared Tanner (University of Oxford, United Kingdom) 5

OMP with Highly Coherent Dictionaries
Raja Giryes (Technion, Israel), Michael Elad (Technion, Israel) 9

Recovery of cosparse signals with Gaussian measurements
Maryia Kabanava (Hausdorff Center for Mathematics, Germany), Holger Rauhut (University of Bonn, Germany) 13

q-ary compressive sensing
Youssef Mroueh (MIT-IIT, USA), Lorenzo Rosasco (DIBRIS, Unige and LCSL - MIT, IIT, USA) 17

Low-rank Tensor Recovery via Iterative Hard Thresholding
Holger Rauhut (University of Bonn, Germany), Reinhold Schneider (Technische Universität Berlin, Germany), Zeljka Stojanac (Universität Bonn, Germany) 21

Time-Frequency Analysis

(Non-)Density Properties of Discrete Gabor Multipliers
Dominik Bayer (Acoustics Research Institute, Austria), Peter Balazs (Austrian Academy of Sciences, Austria) 25

Estimation of frequency modulations on wideband signals; applications to audio signal analysis
Harold Omer (Aix Marseille Université, France), Bruno Torrésani (Aix-Marseille Université, France) 29

Gabor dual windows using convex optimization
Nathanaël Perraudin (Austrian Academy of Sciences, Switzerland), Nicki Holighaus (Austrian Academy of Sciences, Austria), Peter Soendergaard (Austrian Academy of Sciences, Austria), Peter Balazs (Austrian Academy of Sciences, Austria) 33

Sparse Finite Gabor Frames for Operator Sampling
Goetz Pfander (Jacobs University Bremen, Germany), David Walnut (George Mason University, USA) 37

Optimal wavelet reconstructions from Fourier samples via generalized sampling
Clarice Poon (University of Cambridge, United Kingdom), Anders Hansen (University of Cambridge, USA), Ben Adcock (Purdue University, USA) 41

Wavelet Signs: A New Tool for Signal Analysis
Martin Storath (Ecole Polytechnique Federale de Lausanne, Switzerland), Laurent Demaret (HelmholtzZentrum München, Germany), Peter Massopust (Helmholtz Zentrum München, Germany) 45

Optical and RF Systems

Measurement Structures and Constraints in Compressive RF Systems
Nathan A Goodman (University of Oklahoma, USA) 49

Calibration—An open challenge in creating practical computational- and compressive-sensing systems
Michael Gehm (University of Arizona, USA) 53
Compressive CFAR Radar Processing
Laura Anitori (TNO, The Netherlands), Arian Maleki (Rice University, USA), Wim Lambertus van Rossum (TNO, The Netherlands), Matern Otten (TNO, The Netherlands), Richard Baraniuk (Rice University, USA) 57
Sampling Techniques for Improved Algorithmic Efficiency in Electromagnetic Sensing
Kyle R Krueger (Georgia Institute of Technology, USA), James H McClellan (Georgia Institute of Technology, USA), Waymond R Scott, Jr. (Georgia Institute of Technology, USA) 61
Coding and sampling for compressive tomography
David Brady (Duke University, USA) 65
Challenges in Optical Compressive Imaging and Some Solutions
Adrian Stern (Ben-Gurion University of the Negev, Israel), Yair Rivenson (Ben-Gurion University of the Negev, Israel), Yitzhak August (Ben-Gurion University of the Negev, Israel) 69

Sampling and Frame Theory

Balayage and short time Fourier transform frames
Enrico Au-Yeung (University of British Columbia, Canada), John Benedetto (University of Maryland, USA) 73
Fundamental Limits of Phase Retrieval
Afonso Bandeira (Princeton University, USA), Jameson Cahill (University of Missouri, USA), Dustin G. Mixon (Air Force Institute of Technology, USA), Aaron Nelson (Air Force Institute of Technology, USA) 77
On transformations between Gabor frames and wavelet frames
Ole Christensen (Technical University of Denmark, Denmark), Say Goh (National University of Singapore, Singapore) 81
Perfect Preconditioning of Frames by a Diagonal Operator
Gitta Kutyniok (Technical University Berlin, Germany), Kasso Okoudjou (University of Maryland, USA), Friedrich Philipp (Technische Universität Berlin, Germany) 85
Characterizing completions of finite frames
Matthew Fickus (AF Institute of Technology, USA), Miriam Poteet (Air Force Institute of Technology, USA) 89
A note on scalable frames
Jameson Cahill (University of Missouri, USA), Xuemei Chen (University of Maryland, College Park, USA) 93

Sampling and Quantization

Finite-power spectral analytic framework for quantized sampled signals
Truong Thao Nguyen (City College of New York, CUNY, USA) 97
Non-Convex Decoding for Sigma Delta Quantized Compressed Sensing
Evan Chou (New York University, USA) 101
Quantized Iterative Hard Thresholding: Bridging 1bit and HighResolution Quantized Compressed Sensing
Laurent Jacques (University of Louvain, Belgium), Kévin Degraux (Université Catholique Louvain, Belgium), Christophe De Vleeschouwer (UCL, ? Be) 105
Sigma-Delta quantization of sub-Gaussian compressed sensing measurements
Felix Krahmer (University of Göttingen, Germany), Rayan Saab (Duke University, USA), Ozgur Yilmaz (University of British Columbia, Canada) 109

Stable Recovery with Analysis Decomposable Priors
Jalal Fadili (GREYC CNRS UMR 6072, ensicaen, France), Gabriel Peyré (CNRS and Université Paris-Dauphine, France), Samuel Vaiter (CNRS, CEREMADE, Université Paris-Dauphine, France), Charles-Alban Deledalle (CNRS-Université Bordeaux 1, France), Joseph Salmon (CNRS-Télécom ParisTech, France) 113

Finite Rate of Innovation

FRI-based Sub-Nyquist Sampling and Beamforming in Ultrasound and Radar
Tanya Chernyakova (The Technion, IIT, Israel), Omer Bar-Ilan (Technion - Israel Institude of Technology, Israel), Yonina C. Eldar (Technion-Israel Institute of Technology, Israel) 117
Robust Spike Train Recovery from Noisy Data by Structured Low Rank Approximation
Laurent Condat (GIPSA-lab, France), Akira Hirabayashi (Yamaguchi University, Japan) 121
Multichannel ECG Analysis using VPW-FRI
Amrish Nair (Nanyang Technological University, Singapore), Pina Marziliano (Nanyang Technological University, Singapore), Frank Quick (Qualcomm Inc., USA), Ronald Crochiere (Qualcomm Inc., USA), Gilles Baechler (EPFL, Switzerland) 125
Recovery of bilevel causal signals with finite rate of innovation using positive sampling kernels
Gayatri Ramesh (Applying, USA), Elie Atallah (University of Central Florida, USA), Qiyu Sun (University of Central Florida, USA) 129
Approximate FRI with Arbitrary Kernels
Jose Antonio Uriguen (Imperial College of London, Spain), Pier Luigi Dragotti (Imperial College London, United Kingdom), Thierry Blu (Chinese University of Hong Kong, Hong Kong) 133
Algebraic signal sampling, Gibbs phenomenon and Prony-type systems
Dmitry Batenkov (Weizmann Institute of Science, Israel), Yosef Yomdin (Weizmann Institute of Science, Israel) 137

Super Resolution

Super-resolution via superset selection and pruning
Laurent Demanet (MIT, USA), Deanna Needell (Claremont McKenna College, USA), Nam Nguyen (Massachusetts Institute of Technology, USA) 141
Support detection in super-resolution
Carlos Fernandez-Granda (Stanford University, USA) 145
Using Correlated Subset Structure for Compressive Sensing Recovery
Deanna Needell (Claremont McKenna College, USA), Atul Divekar (Alcatel-Lucent, USA) 149
Sub-Wavelength Coherent Diffractive Imaging based on Sparsity
Yoav Shechtman (Technion, Israel), Alexander Szameit (Technion, Israel), Eliahyu Osherovich (Technion, Israel), Pavel Sidorenko (Technion, Israel), Elad Bullkich (Technion - Israel Institute of Technology, Israel), Hod Dana (Technion, Israel), Shy Shoham (Technion, Israel), Irad Yavneh (Technion, Israel), Michael Zibulevsky (Technion - Israel Institute of Technology, Israel), Oren Cohen (Technion, Israel), Yonina C. Eldar (Technion-Israel Institute of Technology, Israel), Mordechai Segev (Technion, Israel) 153
Robust Polyhedral Regularization
Samuel Vaiter (CNRS, CEREMADE, Université Paris-Dauphine, France), Gabriel Peyré (CNRS and Université Paris-Dauphine, France), Jalal Fadili (GREYC CNRS UMR 6072, ensicaen, France) 156

Sampling and Learning

On the Performance of Adaptive Sensing for Sparse Signal Inference
Rui Castro (Eindhoven University of Technology, The Netherlands) 160
Reconstruction of solutions to the Helmholtz equation from punctual measurements
Gilles Chardon (Austrian Academy of Sciences, Austria), Albert Cohen (Universite Pierre et Marie Curie, France), Laurent Daudet (Université Paris Diderot, France) 164
A priori convergence of the Generalized Empirical Interpolation Method
Yvon Maday (UPMC Univ Paris VI - Laboratoire Jacques-Louis Lions, France), Olga Mula (UPMC Univ Paris VI - Laboratoire Jacques-Louis Lions, France), Turinici Gabriel (CEREMADE, Univ Paris Dauphine, France) 168
Test-size Reduction Using Sparse Factor Analysis
Divyanshu Vats (Rice University, USA), Christoph Studer (Rice University, USA), Richard Baraniuk (Rice University, USA) 172

Poster Session I

Special Frames
Daniel Abreu (Acoustic Research Institute, Austria) 176
Variation and approximation for Mellin-type operators
Laura Angeloni (University of Perugia, Italy), Gianluca Vinti (University of Perugia, Italy) 178
iterative methods for random sampling recovery and compressed sensing recovery
Masomeh Azghani (Sharif University of Technology, Iran), Farokh Marvasti (Sharif university of Technology, Iran) 182
A Review of the Invertibility of Frame Multipliers
Peter Balazs (Austrian Academy of Sciences, Austria), Diana Stoeva (Austrian Academy of Sciences, Austria) 186
Hybrid Regularization and Sparse Reconstruction of Imaging Mass Spectrometry Data
Andreas Bartels (University of Bremen, Germany) 189
Level crossing sampling of strongly monoHölder functions
Brigitte Bidegaray-Fesquet (CNRS, France), Marianne Clausel (University Joseph Fourier, France) 193
MAP Estimators for Self-Similar Sparse Stochastic Models
Emrah Bostan (Ecole Polytechnique Fédérale de Lausanne, Switzerland), Julien Fageot (EPFL, Switzerland), Ulugbek S. Kamilov (EPFL, Switzerland), Michael Unser (EPFL, Switzerland) 197
From variable density sampling to continuous sampling using Markov chains
Nicolas Chauffert (CEA, Neurospin Center, Parietal Team., France), Philippe Ciuciu (LNAO, France), Pierre Armand Weiss (ITAV USR 3505, France), Fabrice Gamboa (Université Paul Sabatier (Toulouse III), France) 200
A Comparison of Reconstruction Methods for Compressed Sensing of the Photoplethysmogram
Nicholas Conn (Rochester Institute of Technology, USA), David Borkholder (Rochester Institute of Technology, USA) 204
Generalized sampling in U-invariant subspaces
Héctor Fernández-Morales (Universidad Carlos III de Madrid, Spain), Antonio García (Universidad Carlos III de Madrid, Spain), Miguel Hernández-Medina (Universidad Politécnica de Madrid, Spain) 208
Iterative Hard Thresholding with Near Optimal Projection for Signal Recovery
Raja Giryes (Technion, Israel), Michael Elad (Technion, Israel) 212
The Design of Non-redundant Directional Wavelet Filter Bank Using 1-D Neville Filters
Youngmi Hur (Johns Hopkins University, USA), Fang Zheng (Johns Hopkins University, USA) 216
Sparse Approximation of Ion-Mobility Spectrometry Profiles by Minutely Shifted Discrete B-splines
Masaru Kamada (Ibaraki University, Japan), Masakazu Ohno (Ibaraki University, Japan) 220

Tracking Dynamic Sparse Signals with Kalman Filters: Framework and Improved Inference
Evripidis Karseras (Imperial College London, United Kingdom), Kin K. K. Leung (Imperial College, United Kingdom), Wei Dai (Imperial College, United Kingdom) 224
The Variation Detracting Property of some Shannon Sampling Series and their Derivatives
Andi Kivinukk (Tallinn University, Estonia), Tarmo Metsmägi (Tallinn University, Estonia) 228
Jointly filtering and regularizing seismic data using space-varying FIR filters
Apostolos Kontakis (Delft University of Technology, The Netherlands), Xander Campman (Shell Global Solutions International B. V., The Netherlands), Geert Leus (Delft University of Technology, The Netherlands), Zijian Tang (Shell Global Solutions International B. V., The Netherlands), Mike Danilouchkine (Shell Global Solutions International B. V., The Netherlands) 232
Non-uniform sampling pattern recognition based on atomic decomposition
Tugdual Le Pelleter (TIMA Laboratory, France), Taha Beyrouthy (TIMA Laboratory, France), Robin Rolland (CIME Nanotech, France), Agnès Bonvilain (TIMA Laboratory, France), Laurent Fesquet (TIMA Laboratory, France) 236
Particle Filter Acceleration Using Multiscale Sampling Methods
Yaniv Shmueli (Tel-Aviv University, Israel), Gil Shabat (Tel-Aviv University, Israel), Amit Bermanis (Tel-Aviv University, Israel), Amir Averbuch (Tel Aviv University, Israel) 240
Analysis of Multistage Sampling Rate Conversion for Potential Optimal Factorization
Zhengmao Ye (Southern University, USA), Habib Mohamadian (Southern University, USA) 244
Sparse 2D Fast Fourier Transform
Andre Rauh (University of Delaware, USA), Gonzalo Arce (University of Delaware, USA) 248
GESPAR: Efficient Sparse Phase Retrieval with Application to Optics
Yoav Shechtman (Technion, Israel), Amir Beck (The Technion, Israel), Yonina C. Eldar (Technion-Israel Institute of Technology, Israel) 252

Compressed Sensing

Sparse Signal Reconstruction from Phase-only Measurements
Petros T Boufounos (Mitsubishi Electric Research Laboratories, USA) 256
Optimal Sampling Rates in Infinite-Dimensional Compressed Sensing
Mitra Fatemi (EPFL, Switzerland), Loic Baboulaz (Imperial College, United Kingdom), Martin Vetterli (EPFL, Switzerland) 260
Deterministic Binary Sequences for Modulated Wideband Converter
Lu Gan (Brunel University, United Kingdom), Wang Huali (ICE, PLA UST, P.R. China) 264

Harmonic Analysis

Fractional Prolate Spheroidal Wave Functions
Ahmed Zayed (DePaul University, USA) 268
Absolute Convergence of the Series of Fourier-Haar Coefficients
Boris Golubov (Moscow Institute of Physics and Technologies, Russia), Sergey Volosivets (Saratov State University, Russia) 271
Mellin analysis and exponential sampling. Part I: Mellin fractional integrals
Paul Butzer (RWTH Aachen, Germany), Carlo Bardaro (University of Perugia, Italy), Ilaria Mantellini (University of Perugia, Italy) 274
Mellin analysis and exponential sampling. Part II: Mellin differential operators and sampling
Paul Butzer (RWTH Aachen, Germany), Carlo Bardaro (University of Perugia, Italy), Ilaria Mantellini (University of Perugia, Italy) 277

Sampling in Bio Imaging

Optimisation and control of sampling rate in localisation microscopy
Seamus Holden (Swiss Federal Institute of Technology (EPFL), Switzerland), Thomas Pengo (Center for Genomic Regulation, Spain), Suliana Manley (EPFL, Switzerland) 281
Fast Maximum Likelihood High-density Low-SNR Super-resolution Localization Microscopy
Kyung Sang Kim (KAIST, Korea), Junhong Min (KAIST, Korea), Lina Carlini (EPFL, Switzerland), Michael Unser (EPFL, Switzerland), Suliana Manley (EPFL, Switzerland), Daejong Jeon (KAIST, Korea), Jong Chul Ye (KAIST, Korea) 285
Analogies and differences in optical and mathematical systems and approaches
Bettina Heise (CDL MS-MACH/ ZONA, FLLL, Johannes Kepler University Linz, Austria), Stefan Schausberger (JKU Linz, Austria), Martin Reinhardt (TU Bergakademie Freiberg, Germany), David Stifter (JKU Linz, Austria) 289

Sampling and Geometry

The Nyquist theorem for cellular sheaves
Michael Robinson (American University, USA) 293
Frames of eigenspaces and localization of signal components
José Luis Romero (University of Vienna, Austria), Monika Doerfler (University of Vienna, Austria) 297
A Lie group approach to diffusive wavelets
Swanhild Bernstein (TU Bergakademie Freiberg, Germany) 301
Shannon Sampling and Parseval Frames on Compact Manifolds
Isaac Pesenson (Temple University and CCP, USA) 305
Signal Analysis with Frame Theory and Persistent Homology
Mijail Guillemard (Technische Universität Berlin, Germany), Holger Boche (Technical University Munich, Germany), Gitta Kutyniok (Technical University Berlin, Germany), Friedrich Philipp (Technische Universität Berlin, Germany) 309
Signal Adaptive Frame Theory
Stephen D. Casey (American University, USA) 313

Sampling for Imaging Science

Joint reconstruction of misaligned images from incomplete measurements for cardiac MRI
Gilles Puy (EPFL, Switzerland), Gabriele Bonanno (University of Lausanne, Switzerland), Matthias Stuber (University of Lausanne, Switzerland), Pierre Vandergheynst (EPFL, Switzerland) 317
Localization of point sources in wave fields from boundary measurements using new sensing principle
Zafer Dogan (EPFL, Switzerland), Ivana Jovanovic (EPFL Lausanne, Switzerland), Thierry Blu (Chinese University of Hong Kong, Hong Kong), Dimitri Van De Ville (Ecole Polytechnique Fédérale de Lausanne, Switzerland) 321
Compressive Acquisition of Sparse Deflectometric Maps
Prasad Sudhakar (Universite Catholique de Louvain, Belgium), Laurent Jacques (University of Louvain, Belgium), Adriana Gonzalez Gonzalez (Université Catholique Louvain, Belgium), Xavier Dubois (Lambda-X SA, Belgium), Philippe Antoine (Lambda-X SA, Belgium), Luc Joannes (Lambda-X SA, Belgium) 325
Fourier-Laguerre transform, convolution and wavelets on the ball
Jason McEwen (University College London, United Kingdom), Boris Leistedt (University College London, United Kingdom) 329
Truncation Error in Image Interpolation
Loic Simon (· ENSICAEN Ecole Nationale Superieure d'Ingenieurs de Caen, France) 333

Sampling of Bandlimited Functions

Identification of Rational Transfer Functions from Sampled Data
Hagai Kirshner (EPFL, Switzerland), John Paul Ward (Ecole Polytechnique Federale de Lausanne (EPFL), Switzerland), Michael Unser (EPFL, Switzerland) 337
Reconstruction of Signals from Highly Aliased Multichannel Samples by Generalized Matching Pursuit
Ali Ozbek (Schlumberger Cambridge Research, United Kingdom), Massimiliano Vassallo (WesternGeco London Technology Centre, United Kingdom), Kemal Ozdemir (WesternGeco Oslo Technology Center, Norway), Dirk-Jan van Manen (Schlumberger Cambridge Research, United Kingdom), Kurt Eggenberger (Schlumberger, USA) 340
Joint Signal Sampling and Detection
Mirek Pawlak (University of Manitoba, Canada) 344
On Optimal Sampling Trajectories for Mobile Sensing
Jayakrishnan Unnikrishnan (EPFL, Switzerland), Martin Vetterli (EPFL, Switzerland) 348
Phase Retrieval via Structured Modulations in Paley-Wiener Spaces
Fanny Yang (UC Berkeley, USA), Volker Pohl (Technische Universität München, Germany), Holger Boche (Technical University Munich, Germany) 352

Advances in Compressive Sensing

Energy-aware adaptive bi-Lipschitz embeddings
Bubacarr Bah (École Polytechnique Fédérale de Lausanne (EPFL), Switzerland), Volkan Cevher (Ecole Polytechnique Federale de Lausanne, Switzerland), Ali Sadeghian (Sharif University of Technology, Iran) 356
Randomized Singular Value Projection
Stephen Becker (IBM T. J. Watson Research Center, Yorktown Heights, New York, USA), Volkan Cevher (Ecole Polytechnique Federale de Lausanne, Switzerland), Anastasios Kyrillidis (EPFL, Switzerland) 360
On Sparsity Averaging
Rafael Carrillo (EPFL, Switzerland), Jason McEwen (University College London, United Kingdom), Yves Wiaux (EPFL, Switzerland) 364
Conditions for Dual Certificate Existence in Semidefinite Rank-1 Matrix Recovery
Paul Hand (Massachusetts Institute of Technology, USA) 368
The restricted isometry property for random convolutions
Felix Krahmer (University of Göttingen, Germany), Shahar Mendelson (Technion, Israel), Holger Rauhut (University of Bonn, Germany) 372

Algorithms

Optimal Interpolation Laws for Stable AR(1) Processes
Arash Amini (EPFL, Switzerland) 376
Hierarchical Tucker Tensor Optimization - Applications to Tensor Completion
Curt Da Silva (University of British Columbia, Canada), Felix J. Herrmann (the University of British Columbia, Canada) 380
Estimation of large data sets on the basis of sparse sampling
Anatoli Torokhti (University of South Australia, Australia), Phil Howlett (University of South Australia, Australia), Hamid Laga (University of South Australia, Australia) 384
Analysis of Hierarchical Image Alignment with Descent Methods
Elif Vural (Ecole Polytechnique Federale de Lausanne, Switzerland), Pascal Frossard (EPFL, Switzerland) 388

Spectrum Reconstruction from Sub-Nyquist Sampling of Stationary Wideband Signals
Deborah Cohen (Technion - Israel Institute of Technology, Israel), Yonina C. Eldar (Technion-Israel Institute of Technology, Israel) 392

Poster Session II

Multivariate sampling Kantorovich operators: approximation and applications to civil engineering
Federico Cluni (University of Perugia, Italy), Danilo Costarelli (University of Roma 3, Italy), Anna Maria Minotti (University of Perugia, Italy), Gianluca Vinti (University of Perugia, Italy) 396
On the Number of Degrees of Freedom of Band-Limited Functions
Tatiana Levitina (TU Braunschweig, Germany) 400
Tracing Sound Objects in Audio Textures
Monika Doerfler (University of Vienna, Austria), Ewa Matusiak (Vienna, Austria) 404
An Uncertainty Principle for Discrete Signals
Sangnam Nam (Aix-Marseille Université, France) 408
Efficient Simulation of Continuous Time Digital Signal Processing RF Systems
Alin Ratiu (CEA, France), Dominique Morche (CEA Leti, France), Arnaud Arias (CEA Leti, France), Bruno Allard (INSA Lyon, France), Xuefang Lin-Shi (INSA Lyon, France), Jacques Verdier (Institut National des Sciences Appliquées, France) 412
Shift-Variance and Cyclostationarity of Linear Periodically Shift-Variant Systems
Bashir Sadeghi (Eastern Mediterranean University, Turkey), Runyi Yu (Eastern Mediterranean University, Turkey) 416
Constructive sampling for patch-based embedding
Moshe Salhov (Tel Aviv University, Israel), Guy Wolf (Tel Aviv University, Israel), Amit Bermanis (Tel-Aviv University, Israel), Amir Averbuch (Tel Aviv University, Israel) 420
The Constrained Earth Mover Distance Model, with Applications to Compressive Sensing
Ludwig Schmidt (MIT, USA), Chinmay Hegde (MIT, USA), Piotr Indyk (MIT, USA) 424
Orlicz Modulation Spaces
Catherine Schnackers (RWTH Aachen University, Germany), Hartmut Führ (RWTH Aachen University, Germany) 428
Binary Reduced Row Echelon Form Approach for Subspace Segmentation
Ali Sekmen (Tennessee State University, USA), Akram Aldroubi (Vanderbilt University, USA) 432
Missing Entries Matrix Approximation and Completion
Gil Shabat (Tel-Aviv University, Israel), Yaniv Shmueli (Tel-Aviv University, Israel), Amir Averbuch (Tel Aviv University, Israel) 436
Using Affinity Perturbations to Detect Web Traffic Anomalies
Yaniv Shmueli (Tel-Aviv University, Israel), Tuomo Sipola (University of Jyväskylä, Finland), Gil Shabat (Tel-Aviv University, Israel), Amir Averbuch (Tel Aviv University, Israel) 440
Finite Rate of Innovation Signals: Quantization Analysis with Resistor-Capacitor Acquisition Filter
Srikanth Tenneti (California Institute of Technology, USA), Animesh Kumar (Indian Institute of Technology Bombay, India), Abhay Karandikar (IIT Bombay, India) 444
Tangent space estimation bounds for smooth manifolds
Hemant Tyagi (ETH Zurich, Switzerland), Elif Vural (Ecole Polytechnique Federale de Lausanne, Switzerland), Pascal Frossard (EPFL, Switzerland) 448
A null space property approach to compressed sensing with frames
Rongrong Wang (University of Maryland, USA), Xuemei Chen (University of Maryland, College Park, USA), Haichao Wang (University of California, Davis, USA) 452
Irregular Sampling of the Radon Transform of Bandlimited Functions
Thomas Wiese (Technische Universität München, Germany), Laurent Demaret (HelmholtzZentrum München, Germany) 456
Spline-based frames for image restoration
Valery Zheludev (Tel Aviv University, Israel), Pekka Neittaanmäki (University of Jyväskylä, Finland), Amir Averbuch (Tel Aviv University, Israel) 460

On the Noise-Resilience of OMP with BASC-Based Low Coherence Sensing Matrices
Henning Zörlein (Ulm University, Germany), Dejan Lazich (Ulm University, Germany), Martin Bossert (Ulm University, Germany) 464
Tight frames in spiral sampling
Somantika Datta (University of Idaho, USA), Enrico Au-Yeung (University of British Columbia, Canada) 468

Advances in Compressive Sensing

Local coherence sampling for stable sparse recovery
Felix Krahmer (University of Göttingen, Germany), Holger Rauhut (University of Bonn, Germany), Rachel Ward (University of Texas, USA) 472
Structured-signal recovery from single-bit measurements
Yaniv Plan (University of Michigan, USA) 477
Dictionary Identification Results for K-SVD with Sparsity Parameter 1
Karin Schnass (University of Sassari, Italy) 481

Harmonic Analysis

Measure-based diffusion kernel methods
Amit Bermanis (Tel-Aviv University, Israel), Guy Wolf (Tel Aviv University, Israel), Amir Averbuch (Tel Aviv University, Israel) 485
Spectral properties of dual frames
Felix Krahmer (University of Göttingen, Germany), Gitta Kutyniok (Technical University Berlin, Germany), Jakob Lemvig (Technical University of Denmark, Denmark) 489

Compressive Sensing and Applications

Sparse Recovery with Fusion Frames via RIP
Ulas Ayaz (University of Bonn, Germany), Holger Rauhut (University of Bonn, Germany) 493
Blind Sensor Calibration in Sparse Recovery Using Convex Optimization
Cagdas Bilen (INRIA Rennes, France), Gilles Puy (EPFL, Switzerland), Rémi Gribonval (INRIA, France), Laurent Daudet (Université Paris Diderot, France) 497
Sampling by blocks of measurements in Compressed Sensing
Claire Boyer (Université Paul Sabatier, France), Jérémie Bigot (ISAE, France), Pierre Armand Weiss (ITAV USR 3505, France) 501
Travelling salesman-based variable density sampling
Nicolas Chauffert (CEA, Neurospin Center, Parietal Team., France), Philippe Ciuciu (LNAO, France), Jonas Kahn (Laboratoire Painlevé, CNRS, France), Pierre Armand Weiss (ITAV USR 3505, France) 505
Incremental Sparse Bayesian Learning for Parameter Estimation of Superimposed Signals
Dmitriy Shutin (German Aerospace Center (DLR), Germany), Wei Wang (German Aerospace Center (DLR), Germany), Jost Thomas (German Aerospace Center (DLR), Germany) 509
Sparse MIMO Radar with Random Sensor Arrays and Kerdock Codes
Thomas Strohmer (University of California, Davis, USA), Haichao Wang (University of California, Davis, USA) 513

Sampling of Bandlimited Functions

Sampling and Reconstruction of Bandlimited BMO-Functions
Holger Boche (Technical University Munich, Germany), Ullrich J Mönich (Massachusetts Institute of Technology, USA) 517
Bandlimited Signal Reconstruction From the Distribution of Unknown Sampling Locations
Animesh Kumar (Indian Institute of Technology Bombay, India) 521
Sampling aspects of approximately time-limited multiband and bandpass signals
Joseph Lakey (New Mexico State University, USA), Jeffrey Hogan (University of Newcastle, Australia) 525
Recovery of Bandlimited Signal Based on Nonuniform Derivative Sampling
Dominik Rzepka (AGH University of Science and Technology, Poland), Marek Miskowicz (AGH University of Science and Technology, Poland), Anna Gryboś (AGH University of Science and Technology, Poland), Dariusz Koscielnik (AGH University of Science and Technology, Poland) 529
Approximation by Shannon sampling operators in terms of an averaged modulus of smoothness
Gert Tamberg (Tallinn University of Technology, Estonia), Andi Kivinukk (Tallinn University, Estonia) 533

Circuit Design for Analog to Digital Converters

Digital Calibration of SAR ADC
Yun Chiu (University of Texas at Dallas, USA) 537
Trend of High-Speed SAR ADC towards RF Sampling
Mike Shuo-Wei Chen (University of Southern California, USA) 541
Multi-Step Switching Methods for SAR ADCs
Ying-Zu Lin (Novatek Inc., Taiwan), Ya-Ting Shyu (National Cheng Kung University, Taiwan), Che-Hsun Kuo (National Cheng Kung University, Taiwan), Guan-Ying Huang (National Cheng Kung University, Taiwan), Chun-Cheng Liu (MediaTek Inc., Taiwan), Soon-Jyh Chang (NCKU, Taiwan) 545
On the use of redundancy in successive approximation A/D converters
Boris Murmann (Stanford University, USA) 549
Design Considerations of Ultra-Low-Voltage Self-Calibrated SAR ADC
Hai Huang (UESTC, P.R. China), Xiaoyang Wang (UESTC, P.R. China), Qiang Li (University of Electronic Science and Technology of China, P.R. China) 553

FFT and Related Algorithms

Phase retrieval using time and Fourier magnitude measurements
Martin Ehler (University of Vienna, Germany), Stefan Kunis (University of Osnabrück, Germany) 557
Fast Ewald summation under 2d- and 1d-periodic boundary conditions based on NFFTs
Franziska Nestler (Chemnitz University of Technology, Germany), Daniel Potts (Chemnitz University, Germany) 561
A sparse Prony FFT
Daniel Potts (Chemnitz University, Germany), Stefan Kunis (University of Osnabrück, Germany), Sabine Heider (University of Osnabrück, Germany), Michael Veit (Chemnitz University of Technology, Germany) 565
Taylor and rank-1 lattice based nonequispaced fast Fourier transform
Toni Volkmer (Chemnitz University of Technology, Germany) 569
Decoupling of Fourier Reconstruction System for Shifts of Several Signals
Yosef Yomdin (Weizmann Institute of Science, Israel), Dmitry Batenkov (Weizmann Institute of Science, Israel), Niv Sarig (Nova Measuring Instruments, Israel) 573

Overcoming the coherence barrier in compressed sensing

Ben Adcock
Department of Mathematics
Purdue University

Anders C. Hansen
DAMTP
University of Cambridge

Clarice Poon
DAMTP
University of Cambridge

Bogdan Roman
Computer Laboratory
University of Cambridge

***Abstract*—We introduce a mathematical framework that bridges a substantial gap between compressed sensing theory and its current use in applications. Although completely general, one of the principal applications for our framework is the Magnetic Resonance Imaging (MRI) problem. Our theory provides an explanation for the abundance of numerical evidence demonstrating the advantage of so-called variable density sampling strategies in compressive MRI. Another important conclusion of our theory is that the success of compressed sensing is resolution dependent. At low resolutions, there is little advantage over classical linear reconstruction. However, the situation changes dramatically once the resolution is increased, in which case compressed sensing can and will offer significant benefits.**

I. Introduction

In this paper we present a new mathematical framework for compressed sensing (CS). Our framework generalizes the three traditional pillars of CS—namely, *sparsity*, *incoherence* and *uniform random subsampling*—to three new concepts: *asymptotic sparsity*, *asymptotic incoherence* and *multilevel random subsampling*. As we explain, asymptotic sparsity and asymptotic incoherence are more representative of real-world problems—e.g. imaging—than the usual assumptions of sparsity and incoherence.

Our second contribution is an analysis of an intriguing effect that occurs in asymptotically sparse and asymptotically incoherent problems. Namely, *the success of CS is resolution dependent*. As suggested by their names, asymptotic incoherence and asymptotic sparsity are only truly witnessed for reasonably large problem sizes. When the problem size is small, there is consequently little to be gained from CS over classical linear reconstruction techniques. However, once the resolution of the problem is sufficiently large, CS can and will offer a substantial advantage.

The phenomenon has two important consequences for practitioners seeking to use CS in applications:

(i) Consider a CS experiment where the sampling device, the object to be recovered, the sampling strategy and subsampling percentage are all fixed, but the resolution is allowed to vary. Resolution dependence means that a CS reconstruction done at high resolutions will give much higher quality when compared to full sampling than one done at a low resolution. Hence a practitioner working at low resolution may well conclude that CS imparts limited benefits. However, a markedly different conclusion would be reached if the same experiment were to be performed at higher resolution.

(ii) Suppose we conduct a similar experiment, but we now use the same total number of samples (instead of the same percentage) at low resolution as we take at high resolution. Intriguingly, the above result still holds: namely, the higher resolution reconstruction will give substantially better results. This is true because the multilevel random sampling strategy successfully exploits asymptotic sparsity and asymptotic incoherence. Thus, with the same total number of measurements, CS with multilevel sampling works as a *resolution enhancer*: it recovers fine details of an image in a way that is not possible with the lower resolution reconstruction.

Such resolution dependence suggests the following advisory. It is critical that simulations with CS be carried out with a careful understanding of the influence of the problem resolution. Naïve simulations with standard, low-resolution test images may very well lead to incorrect conclusions about the efficacy of CS as a practical tool.

An important application of our work is the MRI problem. This served as one of the original motivations for CS, and continues to be a topic of substantial research. Some of the earliest work on this problem—in particular, the research of Lustig et al. [1], [2]—demonstrated that the standard random sampling strategies of CS theory lead to substandard reconstructions. This is due to a phenomenon known as the *coherence barrier*.

On the other hand, random sampling according to some nonuniform density was shown empirically to lead to substantially improved reconstruction quality. It is now standard in MR applications to sample in this way [1]–[3]. However, whilst MRI is now viewed as a successful application area for CS, a mathematical theory addressing these sampling strategies is largely lacking. Despite some recent work [4], a substantial gap remains between the standard theorems of CS and its implementation in such problems (see [5] for a detailed discussion). Our framework bridges this gap. In particular, we provide a mathematical foundation for CS for such problems, and gives credence to the abundance of empirical studies demonstrating the success of variable density sampling in overcoming the coherence barrier.

Whilst the MRI problem will serve as our main application, we stress that our theory is general in that it holds for almost arbitrary sampling and sparsity systems. Moreover, standard CS results, in particular those of Candès & Plan [6], are specific instances of our main results.

For brevity, we shall provide only the most salient aspects

of our framework. A substantially more detailed discussion can be found in [5]. We shall also only consider the finite-dimensional case. However, we remark that everything that follows can be extended to infinite-dimensional signals in separable Hilbert spaces [5]. This generalizes the theory of infinite-dimensional CS introduced in [7].

II. Background

A. Compressed sensing

A typical setup in CS is as follows. Let $\{\psi_j\}_{j=1}^N$ and $\{\varphi_j\}_{j=1}^N$ be two orthonormal bases of $\mathbb{C}^N$, the *sampling* and *sparsity* bases respectively, and let

$$U = (u_{ij})_{i,j=1}^N \in \mathbb{C}^{N\times N}, \qquad u_{ij} = \langle \varphi_j, \psi_i \rangle.$$

Note that U is an isometry. The *coherence* of U is given by

$$\mu(U) = \max_{i,j=1,\dots,N} |u_{ij}|^2 \in [N^{-1}, 1], \tag{1}$$

and we say that U is *perfectly* incoherent if $\mu(U) = N^{-1}$.

Let $f \in \mathbb{C}^N$ be s-sparse in the basis $\{\varphi_j\}_{j=1}^N$. In other words, $f = \sum_{j=1}^N x_j \varphi_j$, and the vector $x = (x_j)_{j=1}^N \in \mathbb{C}^N$ satisfies $|\mathrm{supp}(x)| \leq s$, where

$$\mathrm{supp}(x) = \{j : x_j \neq 0\}.$$

Suppose now we have access to the samples

$$\hat{f}_j = \langle f, \psi_j \rangle, \quad j = 1, \dots, N,$$

and let $\Omega \subseteq \{1, \dots, N\}$ be of cardinality m and chosen uniformly at random. According to a result of Candès & Plan [6] and Adcock & Hansen [7], f can be recovered exactly with probability exceeding $1 - \epsilon$ from the subset of measurements $\{\hat{f}_j : j \in \Omega\}$, provided

$$m \gtrsim \mu(U) \cdot N \cdot s \cdot \left(1 + \log(\epsilon^{-1})\right) \cdot \log N. \tag{2}$$

In practice, recovery is achieved by solving the convex optimization problem:

$$\min_{\eta \in \mathbb{C}^N} \|\eta\|_{l^1} \text{ subject to } P_\Omega U \eta = P_\Omega \hat{f}, \tag{3}$$

where $\hat{f} = (\hat{f}_1, \dots, \hat{f}_N)^\top$, and $P_\Omega \in \mathbb{C}^{N\times N}$ is the diagonal projection matrix with j^{th} entry 1 if $j \in \Omega$ and zero otherwise.

B. The coherence barrier

The estimate (2) shows that the number of measurements m is, up to a log factor, on the order of the sparsity s, provided the coherence $\mu(U) = \mathcal{O}\left(N^{-1}\right)$. This is the case, for example, when U is the DFT matrix; a problem which was studied in some of the first papers on CS [8].

On the other hand, when $\mu(U)$ is large, one cannot expect to reconstruct an s-sparse vector f from highly subsampled measurements, regardless of the recovery algorithm employed [6]. We refer to this as the *coherence barrier*.

The MRI problem gives an important instance of this barrier. If $\{\varphi_j\}_{j=1}^N$ is a discrete wavelet basis and $\{\psi_j\}_{j=1}^N$ corresponds to the rows of the $N \times N$ discrete Fourier transform (DFT) matrix, then the matrix $U = \mathrm{DFT} \cdot \mathrm{DWT}^{-1}$ satisfies $\mu(U) = \mathcal{O}(1)$ for any N [4], [9]. Hence, although signals and images are typically sparse in wavelet bases, they cannot be recovered from highly subsampled measurements using the standard CS algorithm.

III. New concepts

We now introduce our new framework that overcomes the aforementioned coherence barrier. We first require the following three new concepts.

A. Asymptotic incoherence

Consider the above example. It is known that, whilst the global coherence $\mu(U)$ is $\mathcal{O}(1)$, the coherence decreases as either the Fourier frequency or wavelet scale increases. We refer to this property as *asymptotic incoherence*:

Definition 1. *Let $U \in \mathbb{C}^{N\times N}$ be an isometry. Then U is asymptotically incoherent if*

$$\lim_{\substack{K,N\to\infty \\ K<N}} \mu(P_K^\perp U) = \lim_{\substack{K,N\to\infty \\ K<N}} \mu(U P_K^\perp) = 0, \tag{4}$$

where $P_K^\perp : \mathbb{C}^{N\times N}$ is the projection matrix corresponding to the index set $\{K+1, \dots, N\}$.

Note that, for the wavelet example discussed above, one has $\mu(P_K^\perp U), \mu(U P_K^\perp) = \mathcal{O}\left(K^{-1}\right)$ [9] for all large N.

B. Multilevel sampling

When U is asymptotically incoherent a different subsampling strategy should be used instead of standard random sampling. High coherence in the first few rows of U means that we cannot subsample in this region without risking losing important information about the signal to be recovered. Hence we fully sample these rows. However, once outside of this region, where the coherence is less, we are free to subsample. Therefore, instead of sampling uniformly at random, we now consider the following *multilevel* random sampling scheme:

Definition 2. *Let $r \in \mathbb{N}$, $\mathbf{N} = (N_1, \dots, N_r) \in \mathbb{N}^r$ with $1 \leq N_1 < \dots < N_r$, $\mathbf{m} = (m_1, \dots, m_r) \in \mathbb{N}^r$, with $m_k \leq N_k - N_{k-1}$, $k = 1, \dots, r$, and suppose that*

$$\Omega_k \subseteq \{N_{k-1}+1, \dots, N_k\}, \quad |\Omega_k| = m_k, \quad k = 1, \dots, r,$$

are chosen uniformly at random, where $N_0 = 0$. We refer to the set $\Omega = \Omega_{\mathbf{N},\mathbf{m}} := \Omega_1 \cup \dots \cup \Omega_r$ as an $(\mathbf{N}, \mathbf{m})$-multilevel sampling scheme.

Note that similar sampling strategies are found in most empirical studies on compressive MRI [1]–[3].

C. Asymptotic sparsity in levels

Having introduced the new sampling strategy for asymptotically incoherent problems, we now consider the following question: what is an appropriate signal model for such a sampling strategy? In the case of incoherence and uniform random subsampling, sparsity is an appropriate model. However, in this new setting we require a somewhat different notion.

To explain this, let $x = (x_j)_{j=1}^N$ be vector of coefficients of a signal f in the basis $\{\varphi_j\}_{j=1}^N$. Suppose that x was very sparse

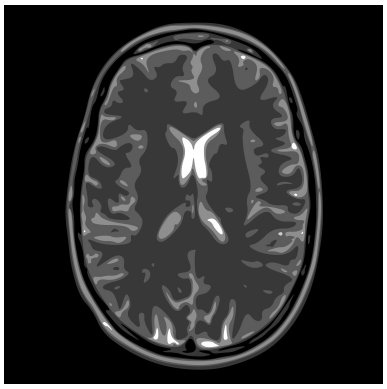


Fig. 1. The GPLU phantom.

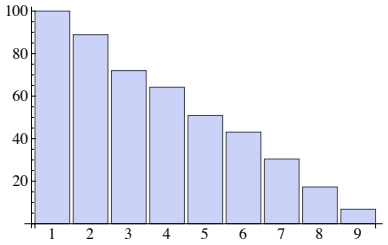


Fig. 2. The percentage of Haar wavelet coefficients at each scale for the image in Fig. 1 which are greater than 10^{-3} in magnitude.

in its entries $j = 1, \ldots, M_1$. Since the matrix U is highly coherent in its corresponding rows, there is no way we can exploit this sparsity to achieve subsampling. High coherence forces us to sample fully the first M_1 rows of U, otherwise we risk missing critical information about x.

This means that there is nothing to be gained from high sparsity of x in its first few entries. However, we can expect to achieve subsampling if the sparsity patten of x matches the incoherence pattern of the matrix U. We therefore consider:

Definition 3. *For $r \in \mathbb{N}$ let $\mathbf{M} = (M_1, \ldots, M_r) \in \mathbb{N}^r$ with $1 \leq M_1 < \ldots < M_r$ and $\mathbf{s} = (s_1, \ldots, s_r) \in \mathbb{N}^r$, with $s_k \leq M_k - M_{k-1}$, $k = 1, \ldots, r$, where $M_0 = 0$. We say that $x \in \mathbb{C}^N$, where $N = M_r$, is $(\mathbf{s}, \mathbf{M})$-sparse if, for each $k = 1, \ldots, r$, the quantity $\Delta_k := \mathrm{supp}(x) \cap \{M_{k-1} + 1, \ldots, M_k\}$ satisfies $|\Delta_k| \leq s_k$.*

In other words, we allow x to be split up into r levels, each with a different amount of sparsity. If the sparsity ratios $s_k/(M_k - M_{k-1})$ decrease with k, then we refer to x as being *asymptotically sparse in levels*.

As we shall see, signals possessing this sparsity pattern are ideally suited to multilevel sampling schemes. Roughly speaking, the concomitance of asymptotic sparsity and asymptotic incoherence means that the number of measurements m_k required in each band Ω_k is determined primarily by the sparsity of f in the corresponding band Δ_k times by a small asymptotic coherence factor.

This leads to the question: is asymptotic sparsity in levels a realistic signal model? The answer is emphatically yes. Most images possess exactly this type of sparsity structure. To illustrate, in Fig. 2 we plot the percentage of significant wavelet coefficients at each scale for the image given in Fig. 1. Note that this image is the analytic phantom introduced by Guerquin–Kern, Lejeune, Pruessmann and Unser in [10]. As is evident, there is little sparsity at coarse scales, but sparsity rapidly increases with refinement.

IV. Main result

For brevity, we shall only address the two-level case (the multilevel case is described in [5]). Thus, we consider signals with a two-level sparsity structure, with the first part being nonsparse, and the second part sparse, and a two-level sampling strategy that corresponds to full sampling in the first rows, and uniform random subsampling in the remaining rows.

Write $\mu_K = \mu(P_K^{\perp} U)$. We now have:

Theorem 4. *Let $U \in \mathbb{C}^{N \times N}$ be an isometry and $x \in \mathbb{C}^N$ be $(\mathbf{s}, \mathbf{M})$-sparse, where $r = 2$, $\mathbf{s} = (s_1, s_2)$ and $\mathbf{M} = (M_1, M_2)$ with $s_1 = M_1$ and $M_2 = N$. Suppose that*

$$\|P_{N_1}^{\perp} U P_{M_1}\| \leq \gamma / \sqrt{M_1}, \tag{5}$$

for some $1 \leq N_1 \leq N$ and $\gamma \in (0, 2/5]$, and that $\gamma \leq s_2 \sqrt{\mu_{N_1}}$. For $\epsilon > 0$, let $m \in \mathbb{N}$ satisfy

$$m \gtrsim (N - N_1) \cdot (\log((s_1 + s_2)\epsilon^{-1}) + 1) \cdot \mu_{N_1} \cdot s_2 \cdot \log(N).$$

Let $\Omega = \Omega_{\mathbf{N},\mathbf{m}}$ be a two-level sampling scheme, where $\mathbf{N} = (N_1, N_2)$ and $\mathbf{m} = (m_1, m_2)$ with $N_2 = N$, $m_1 = N_1$ and $m_2 = m$, and suppose that $\xi \in \mathbb{C}^N$ is a minimizer of (3), where $\hat{f} = Ux$. Then, with probability exceeding $1 - \epsilon$, ξ is unique and $\xi = x$.

Note that if f is not exactly $(\mathbf{s}, \mathbf{M})$-sparse, and if the measurements $\hat{f} = Ux + z$ are corrupted by noise z satisfying $\|z\| \leq \delta$, then one can also prove that under essentially the same conditions the minimization

$$\inf_{\eta \in \mathcal{H}} \|\eta\|_{l^1} \text{ subject to } \|P_\Omega U \eta - y\| \leq \delta. \tag{6}$$

recovers f exactly, up to an error depending only on δ and the error $\sigma_{\mathbf{s},\mathbf{M}}(f)$ of the best approximation of x by an $(\mathbf{s}, \mathbf{M})$-sparse vector. We refer to [5] for details.

A. Discussion

Theorem 4 shows that asymptotic incoherence and two-level sampling overcomes the coherence barrier for two-level sparse signals. To see this, we note:

(i) The condition $\|P_{N_1}^{\perp} U P_{M_1}\| \leq 2/(5\sqrt{M_1})$ (which is always satisfied for some N_1, since U is an isometry) implies that fully sampling the first N_1 measurements allows one to recover the first M_1 coefficients of f.

(ii) To recover the remaining s_2 coefficients we require, up to log factors, an additional $m_2 \gtrsim (N - N_1) \cdot \mu_{N_1} \cdot s_2$ measurements, taken randomly.

Let us explain how this relates to the MRI problem. With Fourier samples and wavelets as the sparsity system, (i) gives

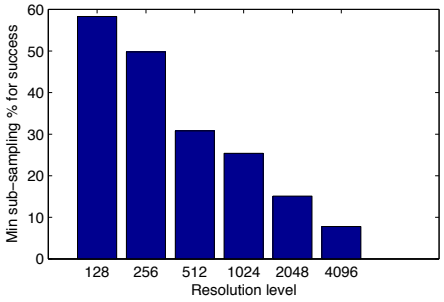


Fig. 3. The minimum subsampling percentage p.

that we recover the nonsparse part of the signal with $N_1 \approx M_1$ measurements. The fact that $N_1 \approx M_1$ in this case was shown in [11]. Since $\mu_{N_1} = \mathcal{O}\left(N_1^{-1}\right)$, (ii) gives that an additional $m_2 \gtrsim s_2$ measurements are required to recover the sparse part of the signal. Hence this result is nearly optimal for signals with two-level asymptotic sparsity. Namely, the full and the sparse parts of the signal are recovered using (up to constants and log factors) optimal numbers of measurements.

We remark that it is not necessary to know the sparsity structure, i.e. the values $\mathbf{s}$ and $\mathbf{M}$, of the image f in order to implement the multilevel sampling technique. Given a multilevel scheme $\Omega = \Omega_{\mathbf{N},\mathbf{m}}$, the result of [5] governing $(\mathbf{s},\mathbf{M})$-compressible signals shows that f will be recovered exactly up to an error on the order of $\sigma_{\mathbf{s},\mathbf{M}}(f)$, where $\mathbf{s}$ and $\mathbf{M}$ are determined implicitly by $\mathbf{N}$, $\mathbf{m}$ and the conditions of the theorem. Of course, some *a priori* knowledge of $\mathbf{s}$ and $\mathbf{M}$ will greatly assist in selecting the parameters $\mathbf{N}$ and $\mathbf{m}$ so as to get the best recovery results. However, this is not strictly necessary for implementing the method.

V. Resolution dependence and numerical results

As explained, natural images are not sparse at coarse wavelet scales, nor is there substantial asymptotic incoherence. Hence, regardless of how we choose to recover f, there is little possibility for substantial subsampling when the problem resolution is low. On the other hand, asymptotic incoherence and asymptotic sparsity both kick in when the resolution increases. Multilevel sampling allows us to exploit these properties, and by doing so we achieve far greater subsampling.

To illustrate this, consider the reconstruction of the 1D function $f(t) = \mathrm{e}^{-t}\chi_{[0.2,0.8]}(t)$, $t \in [0,1]$, from its Fourier samples using Haar wavelets. We use a two-level scheme with $p/2\%$ fixed samples and $p/2\%$ random samples, where p is the total subsampling percentage, and search for the smallest value of p such that the two-level sampling scheme succeeds: namely, it gives an error smaller than that obtained by taking all possible samples of f.

In Fig. 3 we plot p against the resolution N. The difference between low resolution ($N = 128$) and high resolution ($N = 4096$) is clear and dramatic. We conclude that the success of the reconstruction is highly *resolution dependent*.

Now consider a different experiment, where the total number of measurements is fixed and equal to $512^2 = 262144$,

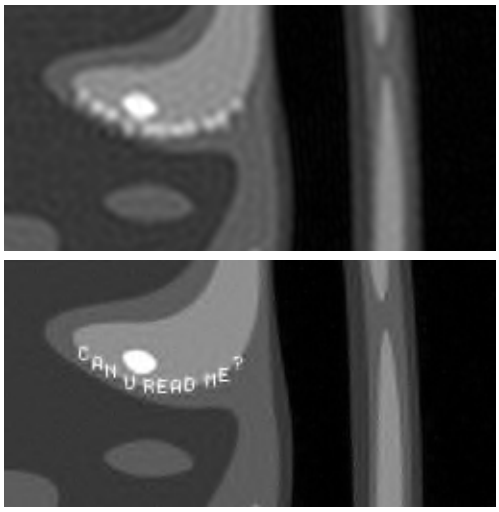


Fig. 4. The reconstruction of the 2048×2048 GPLU phantom (Fig. 1) from 512^2 Fourier samples. Top: linear reconstruction using the first 512^2 Fourier samples and zero padding elsewhere. Bottom: multilevel random CS reconstruction. Note that standard uniform random sampling CS would give an extremely poor reconstruction in this case, due to the $\mathcal{O}(1)$ global coherence.

but the sampling pattern is allowed to vary. In Fig. 4 we display a segment of the reconstruction. For the purposes of comparison, artificial fine details were added to the image to be recovered. As is clear, CS with multilevel sampling acts a *resolution enhancer*. By sampling higher in the Fourier spectrum, one recovers fine details of the image whilst taking the same number of measurements.

For further numerical examples and discussion, see [5].

References

[1] M. Lustig, D. L. Donoho, and J. M. Pauly, "Sparse MRI: the application of compressed sensing for rapid MRI imaging," *Magn. Reson. Imaging*, vol. 58, no. 6, pp. 1182–1195, 2007.

[2] M. Lustig, D. L. Donoho, J. M. Santos, and J. M. Pauly, "Compressed Sensing MRI," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 72–82, March 2008.

[3] G. Puy, P. Vandergheynst, and Y. Wiaux, "On variable density compressive sampling," *IEEE Signal Process. Letters*, vol. 18, pp. 595–598, 2011.

[4] F. Krahmer and R. Ward, "Compressive imaging: stable and robust recovery from variable density frequency samples," *Preprint*, 2012.

[5] B. Adcock, A. C. Hansen, C. Poon, and B. Roman, "Breaking the coherence barrier: asymptotic incoherence and asymptotic sparsity in compressed sensing," *Arxiv 1302.0561*, 2013.

[6] E. J. Candès and Y. Plan, "A probabilistic and RIPless theory of compressed sensing," *IEEE Trans. Inform. Theory*, vol. 57, no. 11, pp. 7235–7254, 2011.

[7] B. Adcock and A. C. Hansen, "Generalized sampling and infinite-dimensional compressed sensing," *Technical report NA2011/02, DAMTP, University of Cambridge*, 2011.

[8] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inform. Theory*, vol. 52, no. 2, pp. 489–509, 2006.

[9] E. J. Candès and J. Romberg, "Sparsity and incoherence in compressive sampling," *Inverse Problems*, vol. 23, no. 3, pp. 969–985, 2007.

[10] M. Guerquin-Kern, L. Lejeune, K. P. Pruessmann, and M. Unser, "Realistic analytical phantoms for parallel Magnetic Resonance Imaging," *IEEE Trans. Med. Imaging*, vol. 31, no. 3, pp. 626–636, 2012.

[11] B. Adcock, A. C. Hansen, and C. Poon, "On optimal wavelet reconstructions from Fourier samples: linearity and universality of the stable sampling rate," *Technical report NA2012/07, DAMTP, University of Cambridge*, 2012.

On construction and analysis of sparse random matrices and expander graphs with applications to compressed sensing.

Bubacarr Bah
Laboratory for Information and Inference Systems
École Polytechnique Fédérale de Lausanne
Lausanne, Switzerland
Email: bubacarr.bah@epfl.ch

Jared Tanner
Mathematics Institute and Exeter College
University of Oxford
Oxford, United Kingdom
Email: tanner@maths.ox.ac.uk

***Abstract*—We revisit the probabilistic construction of sparse random matrices where each column has a fixed number of nonzeros whose row indices are drawn uniformly at random. These matrices have a one-to-one correspondence with the adjacency matrices of lossless expander graphs. We present tail bounds on the probability that the cardinality of the *set of neighbors* for these graphs will be less than the expected value. The bounds are derived through the analysis of collisions in unions of sets using a *dyadic splitting* technique. This analysis led to the derivation of better constants that allow for quantitative theorems on existence of lossless expander graphs and hence the sparse random matrices we consider and also quantitative compressed sensing sampling theorems when using sparse non mean-zero measurement matrices.**

I. Introduction

Sparse matrices are particularly useful in applied and computational mathematics because of their low storage complexity and fast implementation as compared to dense matrices. Of late, significant progress has been made to incorporate sparse matrices in compressed sensing, with [1], [2], [3], [4] giving both theoretical performance guarantees and also exhibiting numerical results that shows sparse matrices coming from expander graphs can be as good sensing matrices as their dense counterparts. In fact, Blanchard and Tanner [5] recently demonstrated in a GPU implementation how well these type of matrices do compared to dense Gaussian and Discrete Cosine Transform matrices even with very small fixed number of nonzeros per column (as considered here).

In this manuscript we consider random sparse matrices that are adjacency matrices of lossless expander graphs. Expander graphs are highly connected graphs with very sparse adjacency matrices, a precise definition of a lossless expander graph is given in Definition 1.

Definition 1: $G(U,V,E)$ is a lossless (k,d,ϵ)-expander if it is a bipartite graph with $|U| = N$ left vertices, $|V| = n$ right vertices and has a regular left degree d, such that any $X \subset U$ with $|X| \leq k$ has a set of neighbors $\Gamma(X) \subset V$ with $|\Gamma(X)| \geq (1-\epsilon)\,d|X|$ neighbors.

Note that these graphs are *lossless* because $\epsilon \ll 1$, they are also referred to as *unbalanced expanders* in the literature because $n \ll N$ and a (k,d,ϵ)-lossless expander graph has an *expansion* of $(1-\epsilon)\,d$. Such graphs have been well studied in theoretical computer science and mathematics and have many applications. Probabilistic constructions of such graphs using random left-regular bipartite graphs with optimal parameters exist but deterministic constructions only achieve sub-optimal parameters, see [6] or [7] for a more detailed survey.

Using a novel technique of *dyadic splitting of sets*, this work derives quantitative guarantees on the probabilistic construction of these graphs in the form of a bound on the tail probability of the size of the *set of neighbors*, $\Gamma(X)$ for a given $X \subset U$, of a randomly generated left-degree bipartite graph. Moreover, this tail bound proves a bound on the tail probability of the *expansion* of the graph, $|\Gamma(X)|/|X|$. In addition, we derive the first phase transitions showing regions in parameter space that depicting when a left-regular bipartite graph with a given set of parameters is guaranteed to be a lossless expander with high probability. Similar results in terms of the adjacency matrices of these graphs is also presented. Another contribution of this work is the derivation of sampling theorems comparing performance guarantees for some of the algorithms proposed for compressed sensing using such sparse matrices as well as the more traditional ℓ_1 minimization compressed sensing formulation. It also provides phase transitions of ℓ_1 minimization performance guarantees for such sparse matrices compared to what ℓ_2 restricted isometry constants (RIC_2) analysis yields for Gaussian matrices.

II. Tail Bound

Our main result is the presentation of formulae for the expected cardinality of the *set of neighbors* of (k,d,ϵ)-lossless expander graphs and the sparse non-mean zero matrices from these graphs. Based on this, we present a tail bound on the probability that this cardinality will be less than the expected value. We start by defining the class of matrices we consider and a key concept of a *set of neighbors* used in our derivation.

Definition 2: Let A be an $n \times N$ matrix with d nonzeros in each column. We refer to A as a random a) sparse expander

(SE) if every nonzero has value 1 and b) sparse signed expander (SSE) if every nonzero has value from $\{-1, 1\}$.

The support set of the d nonzeros per column of these matrices are drawn uniformly at random and independently for each column. An SE matrix is an adjacency matrix of (k, d, ϵ)-lossless expander graph while an SSE matrix have random sign patterns in the nonzeros of an adjacency matrix of a (k, d, ϵ)-lossless expander graph. If A is either an SE or SSE it will have only d nonzeros per column and since we fix $d \ll n$, A is therefore extremely sparse.

We formally define the *set of neighbors* in both graph theory and linear algebra notation to aid translation between the terminology of the two communities. Denote A_S as a submatrix of A composed of columns of A indexed by the set S with $|S| = s$.

Definition 3: Consider a bipartite graph $G(U, V, E)$ where E is the set of edges and $e_{ij} = (x_i, y_j)$ is the edge that connects vertex x_i to vertex y_j. For a given subset of left vertices $S \subset U$ its set of neighbors $\Gamma(S) \subset V$ is defined as $\Gamma(S) := \{y_j | x_i \in S \text{ and } e_{ij} \in E\}$. In terms of the adjacency matrix, A, of $G(U, V, E)$ the set of neighbors of A_S denoted by A_s, is the set of rows with at least one nonzero.

Henceforth, we will only use the linear algebra notation A_s which is equivalent to $\Gamma(S)$. Note that $|A_s|$ is a random variable depending on the draw of the set of columns, S, for each fixed A. Therefore, we can ask what is the probability that $|A_s|$ is not greater than a_s, in particular where a_s is smaller than the expected value of $|A_s|$. This is the question that Theorem 4 attempts to answers.

Theorem 4 (Theorem 1.6, [8]): For fixed s, n, N, d and $d \leq a_s < \infty$, let an $n \times N$ matrix, A be drawn from either of the classes of matrices defined in Definition 2, then

$$\text{Prob}\left(|A_s| \leq a_s\right) < p_{max}(s, d) \cdot e^{[n \cdot \Psi(a_s, \ldots, a_1)]} \qquad (1)$$

where $p_{max}(s, d) = \frac{2}{25\sqrt{2\pi s^3 d^3}}$, and for random variables $a_s, \ldots, a_2$ and $a_1 := d$, $\Psi\left(a_s, \ldots, a_1\right)$ is given by

$$\frac{1}{n}\Bigg[3s\log\left(5d\right) + \sum_{i=1}^{\lceil s/2 \rceil} \frac{s}{2i}\left(\left(n - a_i\right) \cdot \mathrm{H}\left(\frac{a_{2i} - a_i}{n - a_i}\right) + a_i \cdot \mathrm{H}\left(\frac{a_{2i} - a_i}{a_i}\right) - n \cdot \mathrm{H}\left(\frac{a_i}{n}\right)\right)\Bigg],$$

where $\mathrm{H}(\cdot)$ is the Shannon entropy function of base e logarithm. Consequently:

1) if no restriction is imposed on a_s then the a_i for $i > 1$ take on the expected values of $|A_s|$, which are given by $\hat{a}_{2i} = \hat{a}_i\left(2 - \frac{\hat{a}_i}{n}\right)$ for $i = 1, 2, 4, \ldots, \lceil s/2 \rceil$;
2) else if a_s is restricted to be less than $\hat{a}_s$, then the a_i for $i > 1$ are the unique solutions to the following polynomial system $a_{2i}^3 - 2a_i a_{2i}^2 + 2a_i^2 a_{2i} - a_i^2 a_{4i} = 0$ for $i = 1, 2, \ldots, \lceil s/4 \rceil$ with $a_{2i} \geq a_i$ for each i.

Theorem 4 gives a bound on the probability that the cardinality of a union of k sets each with d elements is less than a_k. Figure 1 shows plots of values of $|A_k|$ (size of set of neighbors) for different k taken over 500 realizations (in blue), superimposed on these plots is the empirical mean values of $|A_k|$ over the 500 runs (in red) and the $\hat{a}_k$ in green.

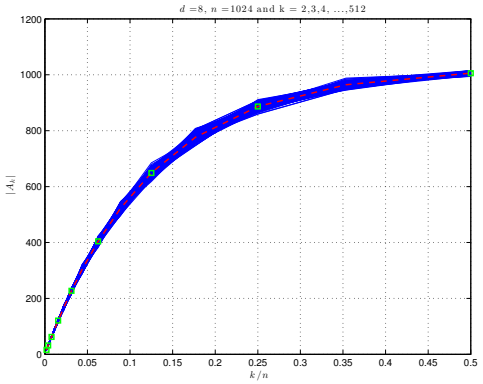


Fig. 1. For fixed $d = 8$ and $n = 2^{10}$, over 500 realizations, plots (in blue) of the cardinalities of the index sets of nonzeros in a given number of set sizes, k. The dotted red curve is mean of the simulations and the green squares are the $\hat{a}_k$.

Furthermore, simulations illustrate that the $\hat{a}_k$ are the expected values of the cardinalities of the union of k sets, $|A_k|$, as shown in Figure 2, where we show the relative error between $\hat{a}_k$ and the empirical mean values of the $|A_k|$, denoted by $\bar{a}_k$, realized over 500 runs, to be less than 10^{-3}.

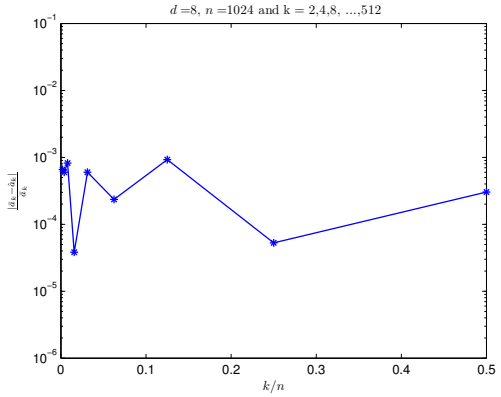


Fig. 2. For fixed $d = 8$ and $n = 2^{10}$, over 500 realizations, plots of the relative error between the mean values of a_k (referred to as $\bar{a}_k$) and the $\hat{a}_k$.

III. Sampling Theorems

We now use Theorem 4 with the ℓ_1-norm restricted isometry property (RIP-1), introduced by Berinde et. al. in [1], to deduce the corollaries that follow which are about the probabilistic construction of expander graphs, the matrices we consider, and sampling theorems of some selected compressed sensing algorithms. Firstly, using only the expansion property of these graphs we can draw the following corollary from Theorem 4.

Corollary 5: For fixed s, n, N, d and $0 < \epsilon < 1/2$, let an $n \times N$ matrix, A be drawn from the class of matrices defined in Definition 2, then

$$\text{Prob}\left(\|A_S x\|_1 \leq (1 - 2\epsilon)d\|x\|_1\right) < p_{max}(s, d) \cdot e^{[n \cdot \Psi(s, d, \epsilon)]},$$

where $\Psi(s,d,\epsilon) = \Psi(a_s,\ldots,a_1)$ with $a_s = (1-\epsilon)ds$.

Theorem 4 and Corollary 5 allow us to calculate s, n, N, d, ϵ where the probability of the probabilistic constructions in Definition 2 not being a (s,d,ϵ)-lossless expander is exponentially small. Using Corollary 5 and the RIP-1 results in [1] we derived a bound for the probability that a random draw of a matrix with d 1s or $\pm$1s in each column fails to satisfy the lower bound of the RIP-1 constant (RIC_1) and hence fails to come from the class of matrices given in Definition 2, for details see [8]. From this bound we deduce the following corollary which is a sampling theorem on the existence of lossless expander graphs.

Corollary 6: Consider $0 < \epsilon < 1/2$ and d fixed. If A is drawn from the class of matrices in Definition 2 and any k-sparse x with $(k,n,N) \to \infty$ while $k/n \to \rho \in (0,1)$ and $n/N \to \delta \in (0,1)$ then for $\rho < (1-\gamma)\rho^{exp}(\delta;d,\epsilon)$ and $\gamma > 0$

$$\text{Prob}\left(\|Ax\|_1 \geq (1-2\epsilon)d\|x\|_1\right) \to 1 \qquad (2)$$

exponentially in n, where $\rho^{exp}(\delta;d,\epsilon)$ is the largest limiting value of k/n for which $\mathrm{H}\left(\frac{k}{N}\right) + \frac{n}{N}\Psi(k,d,\epsilon) = 0$.

For each fixed $0 < \epsilon < 1/2$ and each fixed d, $\rho^{exp}(\delta;d,\epsilon)$ in Corollary 6 is a function of δ and a phase transition function in the (δ,ρ) plane. Below the curve of $\rho^{exp}(\delta;d,\epsilon)$ the probability in (2) goes to one exponentially in n as the problem size grows. That is if A is drawn at random with d 1s or d $\pm$ 1s in each column and having parameters (k,n,N) that fall below the curve of $\rho^{exp}(\delta;d,\epsilon)$ then we say it is from the class of matrices in Definition 2 with probability approaching one exponentially in n. In terms of $|\Gamma(X)|$ for $X \subset U$ and $|X| \leq k$, Corollary 6 say that the probability $|\Gamma(X)| \geq (1-\epsilon)dk$ goes to one exponentially in n if the parameters of our graph lies in the region below $\rho^{exp}(\delta;d,\epsilon)$. This implies that if we draw a random bipartite graphs that has parameters in the region below the curve of $\rho^{exp}(\delta;d,\epsilon)$ then with probability approaching one exponentially in n that graph is a (k,d,ϵ)-lossless expander.

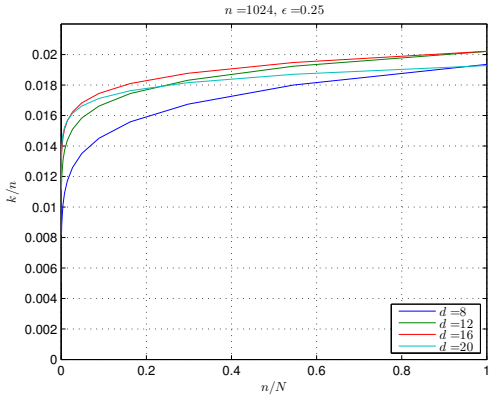


Fig. 3. Phase transition plots of $\rho^{exp}(\delta;d,\epsilon)$ for fixed $\epsilon = 1/6$ and $n = 2^{10}$ with d varied.

Figure 3 shows a plot of what $\rho^{exp}(\delta;d,\epsilon)$ converge to for different values of d with ϵ and n fixed. It is interesting to note how increasing d increases the phase transition up to a point then it decreases the phase transition. Essentially beyond $d = 16$ there is inconsequential gain in increasing d. This vindicates the use of small d in most of the numerical simulations involving the class of matrices considered here. Note the vanishing sparsity as the problem size (k,n,N) grows while d is fixed to a small value of 8.

Corollary 6 can also be arrived at based on similar probabilistic constructions of expander graphs first proven by Pinsker in [9] with more recent proofs in [10], [6]. To put our results in perspective, we compare them to the phase transitions derived from the constants from the construction in [10], shown in Figure 4.

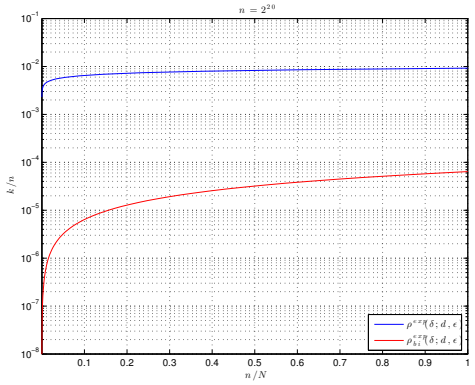


Fig. 4. A comparison of ρ^{exp} in Corollary 6 to ρ_{bi}^{exp} derived from the alternative construction proven in [10].

Furthermore, for moderate values of ϵ this allows us to make quantitative sampling theorems for some compressed sensing reconstruction algorithms. As usual in compressed sensing, in addition to ℓ_1-minimization quite a few *combinatorial* greedy algorithms have been proposed for these sparse non-mean zero matrices. These algorithms iteratively locates and eliminate large (in magnitude) components of the vector, [1]. They include Sequential Sparse Matching Pursuit (SSMP), see [11]; and Expander Recovery (ER), see [3]. Besides, theoretical guarantees have been given for ℓ_1 recovery and some of the greedy algorithms including SSMP and ER. Base on these theoretical guarantees, we derived sampling theorems and present here phase transition curves which are plots of phase transition functions $\rho^{alg}(\delta;d,\epsilon)$ of algorithms such that for $k/n \to \rho < (1-\gamma)\rho^{alg}(\delta;d,\epsilon)$, $\gamma > 0$, a given algorithm is guaranteed to recovery all k-sparse signals with overwhelming probability approaching one exponentially in n.

Figure 5 compares the phase transition of thee above mentioned algorithms. Remarkably, for ER recovery is guaranteed for a larger portion of the (δ,ρ) plane than is guaranteed by the theory for ℓ_1-minimization using sparse matrices; however, ℓ_1-minimization has a larger recovery region than does SSMP. Figure 6 shows a comparison of the phase transition of ℓ_1-minimization as presented by Blanchard et. al. in [12] for dense Gaussian matrices based on RIC_2 analysis and the phase transition we derived here for the sparse binary matrices coming from lossless expander based on RIC_1 analysis. This

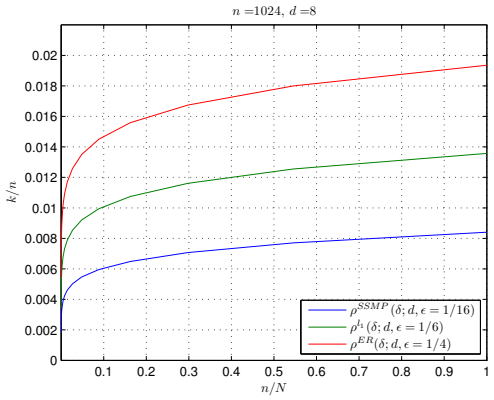


Fig. 5. Phase transition curves $\rho^{alg}(\delta; d, \epsilon)$ computed over finite values of $\delta \in (0, 1)$ with d fixed and the different ϵ values for each algorithm - 1/4, 1/6 and 1/16 for ER, ℓ_1 and SSMP respectively.

shows a significant difference between the two with sparse matrices having better performance guarantees. However, these improved recovery guarantees are likely more due to the closer match of the method of analysis than to the efficacy of sparse matrices over dense matrices.

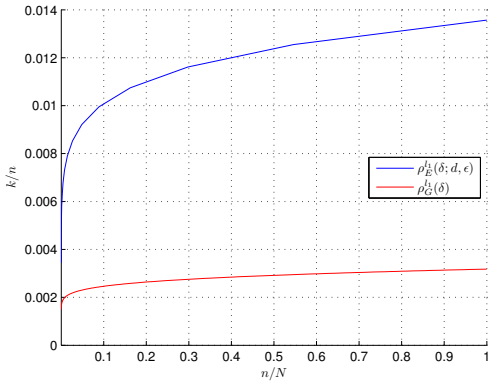


Fig. 6. Phase transition plots of ℓ_1, $\rho_G^{\ell_1}(\delta)$, for Gaussian matrices derived using RIC_2 and $\rho_E^{\ell_1}(\delta; d, \epsilon)$ for adjacency matrices of expander graphs with $n = 1024$, $d = 8$, and $\epsilon = 1/6$.

IV. Sketch of Main Proof

Due to space constraints the details of the proofs are skipped and the interested reader is referred to [8]. It is however important to briefly describe the key innovations in the derivation of the main result, Theorem 4.

For one fixed set of columns of A, denoted A_S, the probability in (1) can be understood as the cardinality of the unions of nonzeros in the columns. Our analysis of this probability follows from a nested unions of subsets using a *dyadic splitting* technique. Given a starting set of columns we recursively split the number of columns from this set and the resulting sets into two sets of cardinality of the ceiling and floor of the cardinality of their union until a level when the cardinalities are at most two. Resulting from this type of splitting is a regular binary tree where the size of each child is either the ceiling or the floor of the size of it's parent set. The probability of interest becomes a product of the probabilities involving all the children from the dyadic splitting of A_s. The proof therefore reduces to upper bounding this product.

Furthermore, in the binary tree resulting from our dyadic splitting scheme the number of columns in the two children of a parent node is the ceiling and the floor of half of the number of columns of the parent node. At each level of the split the number of columns of the children of that level differ by one. The enumeration of these two quantities at each level of the splitting process is necessary in the computation of the bound in (1). This led to another novel technical result in our derivation, i.e. *dyadic splitting lemma* (Lemma 2.5 in [8]).

V. Conclusions

This work derived bounds on the tail probability of the cardinality of the *set of neighbours* of expander graphs resulting into better order constants than the standard probabilistic construction. Using this bound and RIC_1 analysis, we deduce sampling theorems for the existence of expander graphs and their adjacency matrices. The derivation of the tail bound used a novel technique of *dyadic set splitting*. We also compared quantitatively, performance guarantees of compressed sensing algorithms which show greater phase transitions for ER than ℓ_1-minimization which in turn is greater than SSMP. A comparison of ℓ_1-minimization for dense and sparse matrices shows a higher phase transition for sparse matrices.

References

[1] R. Berinde, A. Gilbert, P. Indyk, H. Karloff, and M. Strauss, "Combining geometry and combinatorics: A unified approach to sparse signal recovery," in *Communication, Control, and Computing, 2008 46th Annual Allerton Conference on*. IEEE, 2008, pp. 798–805.

[2] R. Berinde and P. Indyk, "Sparse recovery using sparse random matrices," *preprint*, 2008.

[3] S. Jafarpour, W. Xu, B. Hassibi, and R. Calderbank, "Efficient and robust compressed sensing using optimized expander graphs," *Information Theory, IEEE Transactions on*, vol. 55, no. 9, pp. 4299–4308, 2009.

[4] W. Xu and B. Hassibi, "Further results on performance analysis for compressive sensing using expander graphs," in *Signals, Systems and Computers, 2007. ACSSC 2007. Conference Record of the Forty-First Asilomar Conference on*. IEEE, 2007, pp. 621–625.

[5] J. Blanchard and J. Tanner, "GPU accelerated greedy algorithms for compressed sensing," *Preprint*, 2012.

[6] M. Capalbo, O. Reingold, S. Vadhan, and A. Wigderson, "Randomness conductors and constant-degree lossless expanders," in *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*. ACM, 2002, pp. 659–668.

[7] S. Hoory, N. Linial, and A. Wigderson, "Expander graphs and their applications," *Bulletin of the American Mathematical Society*, vol. 43, no. 4, pp. 439–562, 2006.

[8] B. Bah and J. Tanner, "Vanishingly sparse matrices and expander graphs, with application to compressed sensing," *arXiv preprint arXiv:1207.3094*, 2012.

[9] M. Pinsker, "On the complexity of a concentrator," in *7th annual teletraffic conference*, 1973, p. 318.

[10] R. Berinde, "Advances in sparse signal recovery methods," Master's thesis, Massachusetts Institute of Technology, 2009.

[11] R. Berinde and P. Indyk, "Sequential sparse matching pursuit," in *Communication, Control, and Computing, 2009. Allerton 2009. 47th Annual Allerton Conference on*. IEEE, 2009, pp. 36–43.

[12] J. Blanchard, C. Cartis, J. Tanner, and A. Thompson, "Phase transitions for greedy sparse approximation algorithms," *Applied and Computational Harmonic Analysis*, vol. 30, no. 2, pp. 188–203, 2011.

OMP with Highly Coherent Dictionaries

Raja Giryes and Michael Elad
Computer Science Department
Technion - IIT 32000, Haifa, ISRAEL
Email: [raja,elad]@cs.technion.ac.il

***Abstract*—Recovering signals that has a sparse representation from a given set of linear measurements has been a major topic of research in recent years. Most of the work dealing with this subject focus on the reconstruction of the signal's representation as the means to recover the signal itself. This approach forces the dictionary to be of low-coherence and with no linear dependencies between its columns. Recently, a series of contributions show that such dependencies can be allowed by aiming at recovering the signal itself. However, most of these recent works consider the analysis framework, and only few discuss the synthesis model. This paper studies the synthesis and introduces a new mutual coherence definition for signal recovery, showing that a modified version of OMP can recover sparsely represented signals of a dictionary with very high correlations between pairs of columns. We show how the derived results apply to the plain OMP.**

I. Introduction

Much attention has been given to the problem of recovering a sparse signal from a given set of linear measurements in the recent decade. In the basic setup, an unknown signal $\mathbf{x} \in \mathbb{R}^d$ passes through a given linear transformation $\mathbf{M} \in \mathbb{R}^{m \times d}$ (including $m < d$) with an additive noise $\mathbf{e} \in \mathbb{R}^m$, providing $\mathbf{y} = \mathbf{M}\mathbf{x} + \mathbf{e}$. The signal $\mathbf{x}$ is assumed to have a k-sparse representation $\boldsymbol{\alpha} \in \mathbb{R}^n$ under a given dictionary $\mathbf{D} \in \mathbb{R}^{d \times n}$, i.e. $\mathbf{x} = \mathbf{D}\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}$ has at most k non-zero entries. Most existing work dealing with the problem of estimating $\mathbf{x}$ from $\mathbf{y}$ focuses on the recovery of the signal's representation, assuming that this would lead to the desired signal recovery. This approach forces $\mathbf{D}$ to be incoherent, and in particular, with no linear dependencies between small groups of its atoms.

Recently, a series of papers have shown that such dependencies can be permitted by aiming at estimating the signal itself [1], [2], [3], [4], [5]. Indeed, in [3], [4], [5] it is even suggested that such linear dependencies should be encouraged. However, these contributions consider the "analysis" framework. A first clue that this is not unique to the analysis model but rather applicable also to the "synthesis" appears in [1]. Though its results are for signals from the analysis model, the recovery conditions rely on the $\mathbf{D}$-RIP, a synthesis model property.

The work reported in [6] is different and daring, as it addresses the synthesis model, presenting a variation of CoSaMP that targets the recovery of the signal directly. In their theoretical study, they use the $\mathbf{D}$-RIP to analyze the algorithm's performance assuming the existence of an efficient near-optimal projection scheme, like in [4]. However, the availability of such a projection is questionable in the general case. Another recent work that exploits the $\mathbf{D}$-RIP in the context of the synthesis is [7], proposing stable signal recovery conditions for the basic synthesis ℓ_0-minimization problem.

It is interesting to note that in [6] it is observed that orthogonal matching pursuit (OMP) [8], though not backed up theoretically, achieves some success in recovering signals in the presence of high coherence in the dictionary. In this work we make the first steps to explain this behaviour. We propose a slightly modified version of OMP, OMP$_{\epsilon,2}$, and analyze its performance in the noiseless case ($\mathbf{e} = 0$). Instead of using the $\mathbf{D}$-RIP, we rely on a new property of $\mathbf{M}$ and $\mathbf{D}$: The ϵ-coherence μ_ϵ, which generalizes the definition of the regular coherence μ. Using this definition we show that if $k \leq \frac{1}{2}(1+\frac{1}{\mu_\epsilon}) - O(\epsilon)$ then the OMP$_{\epsilon,2}$ signal recovery error is $O(\epsilon)$. This result implies that OMP$_{\epsilon,2}$ achieves an almost exact reconstruction in the case of very high correlations within pairs of dictionary columns. We draw also the connection between OMP and OMP$_{\epsilon,2}$. Note that our conditions do not include the need for an efficient projection, as needed in [6] .

The organization of this paper is as follows. Section II introduces the ϵ-coherence along with other new definitions. In Section III a modified version of OMP is introduced to support high correlation between pairs of columns. In Section IV the algorithm is analyzed using the ϵ-coherence providing some performance guarantees for the noiseless case. In Section V the derived results are demonstrated empirically.

II. New Coherence Definition

We start with some notation. The largest singular value of a matrix $\mathbf{M}$ is denoted by $\sigma_{\mathbf{M}}$. The i-th column/element of a matrix/vector $\mathbf{D}$/$\mathbf{x}$ is denoted by $\mathbf{d}_i$/$\mathbf{x}_i$, and the sub-matrix/vector with the entries of the support set T by $\mathbf{D}_T$/$\boldsymbol{\alpha}_T$. By abuse of notation, $\boldsymbol{\alpha}_T$ corresponds both to the sub-vector with these entries alone and to the zero padded one. We denote by $\mathbf{W}_\mathbf{D}$ a diagonal matrix that contains the norms of the columns of $\mathbf{D}$ on its diagonal, i.e, $\mathbf{W}_{i,i} = \|\mathbf{d}_i\|_2$.

We turn to introduce some definitions which serve as building blocks in our proposed algorithm and theoretical study. As in [9] the columns of $\mathbf{MD}$ are assumed to be normalized, since if this is not the case a simple scaling can be applied.

Definition 2.1 (ϵ-coherence): Let $0 \leq \epsilon < 1$, $\mathbf{M}$ be a fixed measurement matrix and $\mathbf{D}$ be a fixed dictionary. The ϵ-coherence $\mu_\epsilon(\mathbf{M}, \mathbf{D})$ is defined as

$$\mu_\epsilon(\mathbf{M}, \mathbf{D}) = \max_{1 \leq i < j \leq n} |\langle \mathbf{M}\mathbf{d}_i, \mathbf{M}\mathbf{d}_j \rangle| \quad s.t. \quad \frac{|\langle \mathbf{d}_i, \mathbf{d}_j \rangle|^2}{\|\mathbf{d}_i\|_2^2 \|\mathbf{d}_j\|_2^2} < 1 - \epsilon^2. \tag{1}$$

For calculating $\mu_\epsilon(\mathbf{M},\mathbf{D})$, one may compute the Gram matrices $\mathbf{G_{MD}} = \mathbf{D}^*\mathbf{M}^*\mathbf{MD}$ and $\mathbf{G_D} = \mathbf{W_D^{-1}}\mathbf{D}^*\mathbf{DW_D^{-1}}$. The ϵ-coherence is simply the value of the largest off-diagonal element in absolute value in $\mathbf{G_{MD}}$, corresponding to an entry in $\mathbf{G_D}$ that is smaller in its absolute value than $\sqrt{1-\epsilon^2}$. Note that for $\mathbf{D}=\mathbf{I}$, the ϵ-coherence coincides with the regular coherence $\mu(\mathbf{M})$ and we have $\mu_\epsilon(\mathbf{M},\mathbf{I}) = \mu(\mathbf{M})$. When it is clear to which $\mathbf{M}$ and $\mathbf{D}$ we refer, we use simply μ_ϵ.

Definition 2.2 (ϵ-independent support set): Let $0 \le \epsilon < 1$, $\mathbf{D}$ be a fixed dictionary. A support set T is ϵ-independent with respect to a dictionary $\mathbf{D}$ if $\forall i \neq j \in T$, $\frac{|\langle \mathbf{d}_i, \mathbf{d}_j\rangle|^2}{\|\mathbf{d}_i\|_2^2\|\mathbf{d}_j\|_2^2} < 1-\epsilon^2$.

Definition 2.3 (ϵ-closure): Let $0 \le \epsilon < 1$ and $\mathbf{D}$ be a fixed dictionary. The ϵ-closure of a given support T is defined as $clos_{\epsilon,2}(T) = \{i | \exists j \in T,\ \frac{|\langle \mathbf{d}_i, \mathbf{d}_j\rangle|^2}{\|\mathbf{d}_i\|_2^2\|\mathbf{d}_j\|_2^2} \ge 1-\epsilon^2\}$.

The ϵ-closure of a support T extends it to include each column in $\mathbf{D}$ which is "ϵ-correlated" with elements included in T. Obviously, $T \subseteq clos_{\epsilon,2}(T)$. Note that the last two definitions are related to a given dictionary $\mathbf{D}$. If $\mathbf{D}$ is clear from the context, it is omitted.

III. ϵ-Orthogonal Matching Pursuit

In order to treat the ϵ dependencies in a dictionary we propose the ϵ-orthogonal matching pursuit (OMP$_{\epsilon,2}$) presented in Algorithm 1, which is a modification of OMP [8]. OMP$_{\epsilon,2}$ is the same as the regular OMP but with the addition of the ϵ-closure step. The methods coincide if $\epsilon = 0$ as OMP's orthogonality property guarantees not selecting the same vector twice and thus the ϵ-closure step in OMP$_{\epsilon,2}$ has no effect.

IV. Algorithm Recovery Guarantees

We start with the following Lemma.

Lemma 4.1: Let $\mathbf{x} = \mathbf{D}\boldsymbol{\alpha}$, T be the support of $\boldsymbol{\alpha}$, $\tilde{T}$ be a support set such that $T \subseteq clos_{\epsilon,2}(\tilde{T})$, $\beta_i = \frac{\langle \mathbf{d}_i, \mathbf{d}_{F(i,\mathbf{D}_T)}\rangle}{\left\|\mathbf{d}_{F(i,\mathbf{D}_T)}\right\|_2^2}$ and $\tilde{i} = F(i,\mathbf{D}_T)$ is a function of i such that $\frac{\left|\langle \mathbf{d}_i, \mathbf{d}_{\tilde{i}}\rangle\right|^2}{\|\mathbf{d}_i\|_2^2\left\|\mathbf{d}_{\tilde{i}}\right\|_2^2} \ge 1-\epsilon^2$. If there are several possible $\tilde{i}$ for a given i, choose any one of those and proceed. For the construction

$$\tilde{\mathbf{x}} = \sum_{i\in T\cap\tilde{T}} \mathbf{d}_i\boldsymbol{\alpha}_i + \sum_{i\in T\setminus\tilde{T}} \beta_i \mathbf{d}_{F(i,\mathbf{D}_T)}\boldsymbol{\alpha}_i, \tag{2}$$

we have

$$\|\mathbf{x}-\tilde{\mathbf{x}}\|_2^2 \le \left\|\mathbf{W}_{\mathbf{D}_T}\boldsymbol{\alpha}_{T\setminus\tilde{T}}\right\|_1^2 \epsilon^2. \tag{3}$$

Proof: Note that $\mathbf{x}-\tilde{\mathbf{x}} = \sum_{i\in T\setminus\tilde{T}} \left(\mathbf{d}_i - \beta_i\mathbf{d}_{F(i,\mathbf{D}_T)}\right)\boldsymbol{\alpha}_i$ and $\left\|\mathbf{d}_i - \beta_i\mathbf{d}_{F(i,\mathbf{D}_T)}\right\|_2^2 = \|\mathbf{d}_i\|_2^2\left(1 - \frac{\left|\langle \mathbf{d}_i, \mathbf{d}_{\tilde{i}}\rangle\right|^2}{\|\mathbf{d}_i\|_2^2\left\|\mathbf{d}_{\tilde{i}}\right\|_2^2}\right) \le \|\mathbf{d}_i\|_2^2\,\epsilon^2$. The Cauchy-Schwartz inequality with some arithmetics gives

$$\begin{aligned}\|\mathbf{x}-\tilde{\mathbf{x}}\|_2^2 &= \left\|\sum_{i\in T\setminus\tilde{T}} \left(\mathbf{d}_i - \beta_i\mathbf{d}_{F(i,\mathbf{D}_T)}\right)\boldsymbol{\alpha}_i\right\|_2^2 \qquad (4)\\ &= \sum_{i,j\in T\setminus\tilde{T}} \left(\mathbf{d}_i - \beta_i\mathbf{d}_{F(i,\mathbf{D}_T)}\right)^* \left(\mathbf{d}_j - \beta_j\mathbf{d}_{F(j,\mathbf{D}_T)}\right)\boldsymbol{\alpha}_i\boldsymbol{\alpha}_j\\ &\le \sum_{i\in T\setminus\tilde{T}} \epsilon^2\|\mathbf{d}_i\|_2^2\,\boldsymbol{\alpha}_i^2 + \sum_{i\neq j\in T\setminus\tilde{T}} \epsilon^2\|\mathbf{d}_i\|_2\|\mathbf{d}_j\|_2\,\boldsymbol{\alpha}_i\boldsymbol{\alpha}_j.\end{aligned}$$

Algorithm 1 ϵ-Orthogonal Matching Pursuit

Require: $k, \mathbf{M}, \mathbf{D}, \mathbf{y}$ where $\mathbf{y} = \mathbf{Mx}+\mathbf{e}$, $\mathbf{x} = \mathbf{D}\boldsymbol{\alpha}$, $\|\boldsymbol{\alpha}\|_0 \le k$ and $\mathbf{e}$ is an additive noise.

Ensure: $\hat{\mathbf{x}}_{\text{OMP}_{\epsilon,2}}$: k-sparse approximation of $\mathbf{x}$.

Initialize estimate $\hat{\mathbf{x}}^0 = \mathbf{0}$, residual $\mathbf{r}^0 = \mathbf{y}$, support $\hat{T}^0 = \check{T}^0 = \emptyset$ and set $t = 0$.

while $t \le k$ **do**

$t = t+1$.

New support element: $i^t = \text{argmax}_{i\notin\check{T}^{t-1}} |\mathbf{d}_i^*\mathbf{M}^*(\mathbf{r}^{t-1})|$.

Extend support: $\hat{T}^t = \hat{T}^{t-1}\cup\{i^t\}$.

Calculate a new estimate: $\hat{\mathbf{x}}^t_{\text{OMP}_{\epsilon,2}} = \mathbf{D}_{\hat{T}^t}(\mathbf{MD}_{\hat{T}^t})^\dagger\mathbf{y}$.

Calculate a new residual: $\mathbf{r}^t = \mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^t_{\text{OMP}_{\epsilon,2}}$.

Support ϵ-closure: $\check{T}^t = clos_{\epsilon,2}(\hat{T}^t)$.

end while

Form the final solution $\hat{\mathbf{x}}_{\text{OMP}_{\epsilon,2}} = \hat{\mathbf{x}}^k_{\text{OMP}_{\epsilon,2}}$.

By the definitions of the ℓ_1-norm and $\mathbf{W}_{\mathbf{D}_T}$ we have that the rhs (right-hand-side) of (4) is equal to the rhs of (3). □

Theorem 4.2: Let $0 \le \epsilon < 1$, $\mathbf{M}$ be a fixed measurement matrix, $\mathbf{D}$ be a fixed dictionary with ϵ-coherence $\mu_\epsilon = \mu_\epsilon(\mathbf{M},\mathbf{D})$ and $\mathbf{y} = \mathbf{Mx}$ be a set of measurements of $\mathbf{x} = \mathbf{D}\boldsymbol{\alpha}$ where $\boldsymbol{\alpha}$ is supported on T and $|T| = k$. Let $\tilde{T} \subseteq T$ be an ϵ-independent set such that $T \subseteq clos_{\epsilon,2}(\tilde{T})$ and $\tilde{\mathbf{x}} = \mathbf{D}\tilde{\boldsymbol{\alpha}}$ is constructed according to (2) such that $\tilde{\boldsymbol{\alpha}}$ is supported on $\tilde{T}$. If

$$k < \frac{1}{2}\left(1+\frac{1}{\mu_\epsilon}\right) - \frac{2\|\mathbf{W}_\mathbf{D}\tilde{\boldsymbol{\alpha}}\|_1 + \left\|\mathbf{W}_\mathbf{D}\boldsymbol{\alpha}_{T\setminus\tilde{T}}\right\|_1}{|\tilde{\boldsymbol{\alpha}}_{\min}|\,\mu_\epsilon}\sigma_\mathbf{M}\epsilon, \tag{5}$$

where $\tilde{\boldsymbol{\alpha}}_{\min}$ is the minimal non-zero entry in absolute value of $\tilde{\boldsymbol{\alpha}}$, then after k iterations at most, $\hat{\mathbf{x}}_{\text{OMP}_{\epsilon,2}}$ satisfies

$$\left\|\hat{\mathbf{x}}_{\text{OMP}_{\epsilon,2}} - \mathbf{x}\right\|_2^2 \le \left\|\mathbf{W}_{\mathbf{D}_{T\setminus\tilde{T}}}\boldsymbol{\alpha}_{T\setminus\tilde{T}}\right\|_1^2\epsilon^2 + \|\mathbf{W}_\mathbf{D}\tilde{\boldsymbol{\alpha}}\|_1^2\,\epsilon^2. \tag{6}$$

In particular, if T is an ϵ-independent set then $\boldsymbol{\alpha} = \tilde{\boldsymbol{\alpha}}$ and

$$\left\|\hat{\mathbf{x}}_{\text{OMP}_{\epsilon,2}} - \mathbf{x}\right\|_2^2 \le \|\mathbf{W}_\mathbf{D}\boldsymbol{\alpha}\|_1^2\,\epsilon^2. \tag{7}$$

Before proceeding we comment on the role of ϵ and $\tilde{T}$ in the theorem. If two columns are ϵ-correlated and we use the regular coherence μ, the condition in (5) cannot be met. The use of ϵ-coherence allows us to ignore these correlations and have a reduced coherence value. Thus, the value of ϵ determines the level of correlations the algorithm can handle. Condition (5) bounds this level by $\frac{\frac{1}{2}(\mu_\epsilon+1)-k\mu_\epsilon}{2\|\mathbf{W}_\mathbf{D}\tilde{\boldsymbol{\alpha}}\|_1+\left\|\mathbf{W}_\mathbf{D}\boldsymbol{\alpha}_{T\setminus\tilde{T}}\right\|_1}\frac{|\tilde{\boldsymbol{\alpha}}_{\min}|}{\sigma_\mathbf{M}}$. Remark that as ϵ approaches zero the value of μ_ϵ approaches μ_0, a mutual coherence of $\mathbf{D}$ that ignores the dependent columns.

The set $\tilde{T}$ is needed in the theorem because the columns of $\mathbf{D}_T$, which span $\mathbf{x}$, might be ϵ-correlated or even dependent. To avoid that, we select the maximal subset of T which is ϵ-independent and still includes T in its ϵ-closure. The construction of such a maximal subset is easy. We start by initializing $\tilde{T} = T$, and then sequentially for each index $i \in \tilde{T}$ update $\tilde{T} = \tilde{T}\setminus clos_{\epsilon,2}(\{i\})$. The resulting subset $\tilde{T}$ is guaranteed to be ϵ-independent and have $T \subseteq clos_{\epsilon,2}(\tilde{T})$.

The following key Lemma is used in the Theorem's proof.

Lemma 4.3: Under the same setup of Theorem 4.2, we have

$$\tilde{T} \subseteq \check{T}^k = clos_{\epsilon,2}(\hat{T}^k). \tag{8}$$

Proof: We prove by induction on the iteration $t \leq |\tilde{T}| = \tilde{k}$ that either $\tilde{T} \subseteq \check{T}^t$ or $\exists i \in \tilde{T}$ such that $i \in \check{T}^t$ and $i \notin \check{T}^{t-1}$. Since the induction guarantees that in each iteration a new element from $\tilde{T}$ is included in $\check{T}^t$, after $k \geq \tilde{k}$ iterations (8) holds.

The basis of the induction is $t = 1$. Define $\bar{T} = clos_{\epsilon,2}(\tilde{T})$. The basis holds if in the first iteration we select an element from $\bar{T}$. This is true due to the fact that $\forall i, j \in T$, $i \in clos_{\epsilon,2}(\{j\})$ iff $j \in clos_{\epsilon,2}(\{i\})$. Thus, we need to require

$$\max_{i\in\bar{T}} |\mathbf{d}_i^*\mathbf{M}^*\mathbf{y}| > \max_{i\in\bar{T}^C} |\mathbf{d}_i^*\mathbf{M}^*\mathbf{y}|. \tag{9}$$

First note that $\mathbf{y} = \mathbf{M}\tilde{\mathbf{x}} + \mathbf{M}(\mathbf{x} - \tilde{\mathbf{x}})$. Thus, using the triangle inequality, the Cauchy-Shwartz inequality and the facts that the ℓ_2-norm is multiplicative and $\|\mathbf{M}\mathbf{d}_i\|_2 = 1$, (9) holds if

$$\max_{i\in\bar{T}} |\mathbf{d}_i^*\mathbf{M}^*\mathbf{M}\tilde{\mathbf{x}}| > \max_{i\in\bar{T}^C} |\mathbf{d}_i^*\mathbf{M}^*\mathbf{M}\tilde{\mathbf{x}}| + 2\|\mathbf{M}(\mathbf{x} - \tilde{\mathbf{x}})\|_2. \tag{10}$$

In order to check when the last happens we shall bound its lhs (left-hand-side) from below and its rhs from above.

Assuming w.l.o.g that the index of the largest entry in $\tilde{\boldsymbol{\alpha}}$ is 1, we have for the lhs of (10)

$$\begin{aligned}\max_{i\in\bar{T}} |\mathbf{d}_i^*\mathbf{M}^*\mathbf{M}\tilde{\mathbf{x}}| &\geq |\mathbf{d}_1^*\mathbf{M}^*\mathbf{M}\tilde{\mathbf{x}}| = |\sum_{l\in\tilde{T}} \mathbf{d}_1^*\mathbf{M}^*\mathbf{M}\mathbf{d}_l\tilde{\boldsymbol{\alpha}}_l| \quad (11)\\ &\geq |\mathbf{d}_1^*\mathbf{M}^*\mathbf{M}\mathbf{d}_1\tilde{\boldsymbol{\alpha}}_1| - \sum_{l\in\tilde{T}, l\neq 1} |\mathbf{d}_1^*\mathbf{M}^*\mathbf{M}\mathbf{d}_l\tilde{\boldsymbol{\alpha}}_l| \\ &\geq \tilde{\boldsymbol{\alpha}}_1 - \mu_\epsilon \sum_{l\in\tilde{T}, l\neq 1} |\tilde{\boldsymbol{\alpha}}_l| = (1 - (\tilde{k} - 1)\mu)\,|\tilde{\boldsymbol{\alpha}}_1|,\end{aligned}$$

where the first inequality is due to the triangle inequality; the second is due to the fact that $\|\mathbf{M}\mathbf{d}_i\|_2 = 1$, the definition of μ_ϵ and the Cauchy-Schwartz inequality; and the last is because $\tilde{\boldsymbol{\alpha}}_1$ is the largest element in $\tilde{\boldsymbol{\alpha}}$ and $|\tilde{T}| = \tilde{k}$.

We turn now to bound the rhs of (10) from above. Using the same considerations, we have

$$\begin{aligned}\max_{i\in\bar{T}^C} |\mathbf{d}_i^*\mathbf{M}^*\mathbf{M}\mathbf{x}| &= \max_{i\in\bar{T}^C} |\sum_{l\in\tilde{T}} \mathbf{d}_i^*\mathbf{M}^*\mathbf{M}\mathbf{d}_l\tilde{\boldsymbol{\alpha}}_l| \quad (12)\\ \leq \max_{i\in\bar{T}^C} \sum_{l\in\tilde{T}} |\mathbf{d}_i^*\mathbf{M}^*\mathbf{M}\mathbf{d}_l\tilde{\boldsymbol{\alpha}}_l| &\leq \sum_{l\in\tilde{T}} \mu_\epsilon\,|\tilde{\boldsymbol{\alpha}}_l| \leq |\tilde{\boldsymbol{\alpha}}_1|\,\tilde{k}\mu_\epsilon.\end{aligned}$$

Plugging (11) and (12) into (10) and then using Lemma 4.1 with the fact that $\|\mathbf{M}\|_2 = \sigma_{\mathbf{M}}$ gives us the condition

$$\tilde{k} < \frac{1}{2}(1 + \frac{1}{\mu_\epsilon}) - \frac{\sigma_{\mathbf{M}}}{\mu_\epsilon\tilde{\boldsymbol{\alpha}}_1} \left\|\mathbf{W}_{\mathbf{D}_T}\boldsymbol{\alpha}_{T\setminus\tilde{T}}\right\|_1 \epsilon, \tag{13}$$

for selecting an element from $\bar{T}$ in the first iteration.

Having the induction basis proven, we turn to the induction step. Assume that the induction assumption holds till iteration $t-1$. We need to prove that it holds also in the t-th iteration. Let $\bar{T}^t = clos_{\epsilon,2}(\tilde{T} \setminus \check{T}^{t-1}))$. This set includes the ϵ-closure of elements in $\tilde{T}$ for which an element was not selected in the previous iterations. For proving the induction step it is enough to show that in the t-th iteration we select an index from $\bar{T}^t$:

$$\max_{i\in\bar{T}^t} \left|\mathbf{d}_i^*\mathbf{M}^*\mathbf{r}^{t-1}\right| > \max_{i\in(\bar{T}^t)^C\setminus\check{T}^{t-1}} \left|\mathbf{d}_i^*\mathbf{M}^*\mathbf{r}^{t-1}\right|. \tag{14}$$

On the rhs we do not check the maximum over elements in $\check{T}^{t-1}$ because $\text{OMP}_{\epsilon,2}$ excludes these indices in the step of selecting a new element. As in the basis of the induction, in order to check when (14) holds we shall bound its lhs from below and its rhs from above. Let $\tilde{\mathbf{x}}^{t-1} = \sum_{i\in\tilde{T}\setminus\hat{T}^{t-1}} \mathbf{d}_i\tilde{\boldsymbol{\alpha}}_i + \sum_{i\in\hat{T}^{t-1}} \beta_i\mathbf{d}_{F(i,D_{\tilde{T}})}\tilde{\boldsymbol{\alpha}}_i$ be constructed as in (2) where we use the fact that $\tilde{\boldsymbol{\alpha}}$ is supported on $\tilde{T}$. Denoting $\tilde{\mathbf{r}}^{t-1} = (\mathbf{I} - \mathbf{M}\mathbf{D}_{\hat{T}^{t-1}}(\mathbf{M}\mathbf{D}_{\hat{T}^{t-1}})^\dagger)\mathbf{M}\tilde{\mathbf{x}}^{t-1}$ and using a similar argument like in (10) we have that (14) holds if

$$\begin{aligned}\max_{i\in\bar{T}^t} |\mathbf{d}_i^*\mathbf{M}^*\tilde{\mathbf{r}}^{t-1}| \;&> \max_{i\in(\bar{T}^t)^C\setminus\check{T}^{t-1}} |\mathbf{d}_i^*\mathbf{M}^*\tilde{\mathbf{r}}^{t-1}| \quad (15)\\ &\quad + 2\left\|\tilde{\mathbf{r}}^{t-1} - \mathbf{r}^{t-1}\right\|_2.\end{aligned}$$

Notice that $\tilde{\mathbf{r}}^{t-1}$ is supported on $\hat{T}^{t-1}\cup(\tilde{T}\setminus\check{T}^{t-1})$, i.e., $\tilde{\mathbf{r}}^{t-1} = \mathbf{M}\mathbf{D}_{\hat{T}^{t-1}\cup(\tilde{T}\setminus\check{T}^{t-1})}\tilde{\boldsymbol{\alpha}}^{\mathbf{r}^{t-1}}$, and $\tilde{\boldsymbol{\alpha}}^{\mathbf{r}^{t-1}}_{\tilde{T}\setminus\check{T}^{t-1}} = \tilde{\boldsymbol{\alpha}}_{\tilde{T}\setminus\check{T}^{t-1}}$.

We want to show that the index of the maximal coefficient (in absolute value) of $\tilde{\mathbf{r}}^{t-1}$ belongs to $\tilde{T} \setminus \check{T}^{t-1}$ and hence we will be able to use almost the same derivation of the basis of the induction. We prove it by contradiction. Assume that the maximum is achieved for $i \in \hat{T}^{t-1}$. By the orthogonality property of the residual it is easy to see that $\mathbf{d}_i^*\mathbf{M}^*\tilde{\mathbf{r}}^{t-1} = 0$. Using similar considerations as in (11) we have $0 = \left|\mathbf{d}_i^*\mathbf{M}^*\tilde{\mathbf{r}}^{t-1}\right| \geq (1-(\tilde{k}-1)\mu_\epsilon)\left|\tilde{\boldsymbol{\alpha}}_i^{\mathbf{r}^{t-1}}\right|$ which implies $\tilde{k} \geq 1 + \frac{1}{\mu_\epsilon}$ and we get a contradiction to (5).

Let w.l.o.g. t be the maximal coefficient in $\tilde{\boldsymbol{\alpha}}_{i_t}^{\mathbf{r}^{t-1}}$. By the above observations $t \in \tilde{T} \setminus \check{T}^{t-1}$ and $\tilde{\boldsymbol{\alpha}}_t^{\mathbf{r}^{t-1}} = \tilde{\boldsymbol{\alpha}}_t$. Applying the same steps as in (11) and (12), we have

$$\begin{aligned}\max_{i\in\bar{T}^t} \left|\mathbf{d}_i^*\mathbf{M}^*\tilde{\mathbf{r}}^{t-1}\right| &\geq (1 - \mu_\epsilon(\tilde{k} - 1))\,|\tilde{\boldsymbol{\alpha}}_t|, \quad (16)\\ \max_{i\in(\bar{T}^t)^C\setminus\check{T}^{t-1}} \left|\mathbf{d}_i^*\mathbf{M}^*\tilde{\mathbf{r}}^{t-1}\right| &\leq \mu_\epsilon\tilde{k}\,|\tilde{\boldsymbol{\alpha}}_t|.\end{aligned}$$

Using norm inequalities and the projection property that implies $\left\|\mathbf{I} - \mathbf{M}\mathbf{D}_{\hat{T}^{t-1}}(\mathbf{M}\mathbf{D}_{\hat{T}^{t-1}})^\dagger\right\|_2 \leq 1$, we have

$$\begin{aligned}\left\|\tilde{\mathbf{r}}^{t-1} - \mathbf{r}^{t-1}\right\|_2 &\leq \left\|\mathbf{M}(\tilde{\mathbf{x}}^{t-1} - \mathbf{x})\right\|_2 \leq \sigma_{\mathbf{M}}\left\|\tilde{\mathbf{x}}^{t-1} - \mathbf{x}\right\|_2 \quad (17)\\ &\leq \sigma_{\mathbf{M}}\left\|\tilde{\mathbf{x}}^{t-1} - \tilde{\mathbf{x}}\right\|_2 + \sigma_{\mathbf{M}}\left\|\tilde{\mathbf{x}} - \mathbf{x}\right\|_2\end{aligned}$$

Using Lemma 4.1 with (17) and then combining it with (15) and (16) results with the condition

$$\tilde{k} < \frac{1}{2} + \frac{1}{2\mu_\epsilon} - \frac{\sigma_{\mathbf{M}}\epsilon}{|\tilde{\boldsymbol{\alpha}}_t|\,\mu_\epsilon}(\|\mathbf{W}_{\mathbf{D}}\tilde{\boldsymbol{\alpha}}_{\tilde{T}^{t-1}}\|_1 + \left\|\mathbf{W}_{\mathbf{D}}\boldsymbol{\alpha}_{T\setminus\tilde{T}}\right\|_1). \tag{18}$$

The proof ends by noticing that (18) is implied by (5). □

Proof of Theorem 4.2: Note that $\hat{\mathbf{x}}_{\text{OMP}_{\epsilon,2}} = \mathbf{D}_{\hat{T}^k}(\mathbf{M}\mathbf{D}_{\hat{T}^k})^\dagger\mathbf{y}$ and $\mathbf{y} = \mathbf{M}\mathbf{x}$. Using some basic algebraic steps we have

$$\begin{aligned}\left\|\hat{\mathbf{x}}_{\text{OMP}_{\epsilon,2}} - \mathbf{x}\right\|_2 &= \left\|\mathbf{D}_{\hat{T}^k}(\mathbf{M}\mathbf{D}_{\hat{T}^k})^\dagger\mathbf{M}\mathbf{x} - \mathbf{x}\right\|_2 \quad (19)\\ &= \left\|(\mathbf{D}_{\hat{T}^k}(\mathbf{M}\mathbf{D}_{\hat{T}^k})^\dagger\mathbf{M} - \mathbf{I})(\mathbf{I} - \mathbf{D}_{\hat{T}^k}\mathbf{D}_{\hat{T}^k}^\dagger)\mathbf{x}\right\|_2 \\ &\leq \left\|(\mathbf{I} - \mathbf{D}_{\hat{T}^k}\mathbf{D}_{\hat{T}^k}^\dagger)\mathbf{x}\right\|_2,\end{aligned}$$

where the last inequality is due to the fact that $\mathbf{D}_{\hat{T}^k}(\mathbf{M}\mathbf{D}_{\hat{T}^k})^\dagger\mathbf{M} - \mathbf{I}$ is a projection operator and thus its operator norm is smaller or equal to 1. Splitting $\mathbf{x}$ into $\tilde{\mathbf{x}}$

and $\mathbf{x}-\tilde{\mathbf{x}}$, and then using the triangle inequality and the fact that $\mathbf{I}-\mathbf{D}_{\hat{T}^k}\mathbf{D}^{\dagger}_{\hat{T}^k}$ is a projection with (19) give

$$\|\hat{\mathbf{x}}_{\text{OMP}}-\mathbf{x}\|_2 \leq \left\|(\mathbf{I}-\mathbf{D}_{\hat{T}^k}\mathbf{D}^{\dagger}_{\hat{T}^k})\tilde{\mathbf{x}}\right\|_2 + \|\mathbf{x}-\tilde{\mathbf{x}}\|_2 . \quad (20)$$

By Lemma 4.3, after k iterations (8) holds. Thus, Lemma 4.1 implies the existence of a vector $\hat{\mathbf{z}}^k$, with a representation supported on $\hat{T}^k$, satisfying $\left\|\tilde{\mathbf{x}}-\hat{\mathbf{z}}^k\right\|_2 \leq \left\|\mathbf{W}_{\mathbf{D}_{\tilde{T}}}\tilde{\boldsymbol{\alpha}}\right\|_1 \epsilon$. This and projection properties yield for the first element in the rhs

$$\left\|(\mathbf{I}-\mathbf{D}_{\hat{T}^k}\mathbf{D}^{\dagger}_{\hat{T}^k})\tilde{\mathbf{x}}\right\|_2 \leq \left\|\tilde{\mathbf{x}}-\hat{\mathbf{z}}^k\right\|_2 \leq \left\|\mathbf{W}_{\mathbf{D}_{\tilde{T}}}\tilde{\boldsymbol{\alpha}}\right\|_1 \epsilon. \quad (21)$$

For the second element we have using Lemma 4.1

$$\|\mathbf{x}-\tilde{\mathbf{x}}\|_2 \leq \left\|\mathbf{W}_{\mathbf{D}_{T\setminus\tilde{T}}}\boldsymbol{\alpha}_{T\setminus\tilde{T}}\right\|_1 \epsilon. \quad (22)$$

Plugging (22) and (21) in (20) results with (6). Notice that if T is an ϵ-independent set then $T=\tilde{T}$ and (7) follows immediately from (6) because the first term in its rhs vanishes and in the second one $\mathbf{W}_{\mathbf{D}_T}\boldsymbol{\alpha}_T = \mathbf{W}_{\mathbf{D}}\boldsymbol{\alpha}$ since $\boldsymbol{\alpha}_{T^C}=0$. □

Remark 4.4: Theorem 4.2 can be easily extended to the noisy case using the proof technique in [9].

Remark 4.5: If for a certain vector $\mathbf{x}$ supported on T, we get $\left|\check{T}^k\right| \leq d$ then the condition in (5) in Theorem 4.2 implies a perfect recovery by using a simple twist in $\text{OMP}_{\epsilon,2}$, setting $\check{\mathbf{x}}_{\text{OMP}_{\epsilon,2}} = \mathbf{D}_{\check{T}^k}(\mathbf{M}\mathbf{D}_{\check{T}^k})^{\dagger}\mathbf{y}$. Due to uniqueness conditions, in this case $\check{\mathbf{x}}_{\text{OMP}_{\epsilon,2}} = \mathbf{x}$. It can be easily shown that $|clos_{2\epsilon,2}(T)| \leq d$ is a sufficient condition for this to happen.

Remark 4.6: From the previous remark we conclude that if for any T such that $|T| \leq k$ we have $|clos_{2\epsilon,2}(T)| \leq d$ then the algorithm provides us always with a perfect recovery.

Remark 4.7: Theorem 4.2 applies also to the regular OMP if $\frac{|\langle \mathbf{d}_i,\mathbf{d}_j\rangle|^2}{\|\mathbf{d}_i\|_2^2\|\mathbf{d}_j\|_2^2} < 1-\epsilon^2$ implies $|\langle \mathbf{M}\mathbf{d}_i,\mathbf{M}\mathbf{d}_j\rangle|^2 < 1-\epsilon^2$. The latter property guarantees that in the step of selecting a new element, OMP does not choose an index from $\check{T}^t$. For a formal proof, the induction step in Lemma 4.3 needs to be modified showing that an element from $\check{T}^t$ is not chosen.

V. Numerical Simulation

We turn to check numerically the recovery performance of OMP and $\text{OMP}_{\epsilon,2}$ for sparse signals under a dictionary that contains pairs of correlated columns. We generate a dictionary $\mathbf{D}=[\mathbf{D}^1,\mathbf{D}^2]$ where $\mathbf{D}^1,\mathbf{D}^2 \in \mathbb{R}^{d\times d}$, $d=1000$, $\mathbf{D}^1$ contains sparse columns with 2 non-zero entries which are ± 1 with probability 0.5 like in [7] and $\mathbf{D}^2$ is constructed such that each of its columns $\mathbf{d}_i^2$ is ϵ-correlated to the corresponding column $\mathbf{d}_i^1$. Each entry of the measurement matrix $\mathbf{M} \in \mathbb{R}^{m\times d}$ is distributed according to a normal Gaussian distribution, where $m{=}\lfloor \gamma d \rfloor$ and γ is the sampling rate – a value in the range $(0,1]$. We set k to be $\lfloor \rho m \rfloor$ ($\rho \ll 1$) and measure the recovery rate of the representation $\boldsymbol{\alpha}$ and the signal $\mathbf{x}$ for various values of $\gamma \in \{0.1, 0.2, \ldots, 0.9\}$ and $\rho \in \{0.02, 0.04, \ldots, 0.2\}$.

Figure 1 presents the recovery performance over 100 realizations per each parameter setting. We use the observation in Remark 4.5 and present the recovery rate of $\text{OMP}_{\epsilon,2}$ for both $\hat{\mathbf{x}}_{\text{OMP}_{\epsilon,2}}$ and $\check{\mathbf{x}}_{\text{OMP}_{\epsilon,2}}$. As expected from Theorem 4.2, for the first we do not get a perfect recovery but only an error

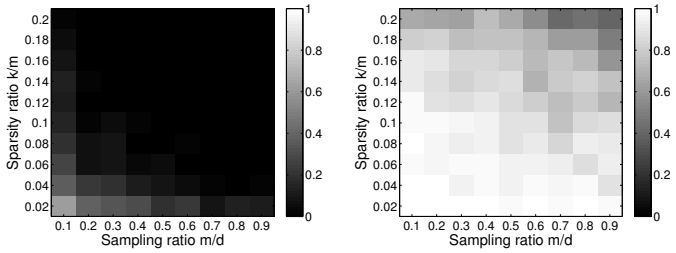


Fig. 1. $\text{OMP}_{\epsilon,2}$ recovery rate for $\hat{\mathbf{x}}_{\text{OMP}_{\epsilon,2}}$ (left) and $\check{\mathbf{x}}_{\text{OMP}_{\epsilon,2}}$ (right) for the synthetic experiment described in Section V. Color attribute: fraction of realizations in which a perfect recovery is achieved.

of an order of ϵ (due to lack of space we do not present the recovery error). However, as observed in Remark 4.5 when we take an ϵ-closure on the achieved support we get an almost perfect recovery. As high correlations between columns in $\mathbf{D}$, indeed imply high correlations between columns in $\mathbf{MD}$ in the common case, the recovery performance we present for $\text{OMP}_{\epsilon,2}$ are the same as for OMP as predicted in Remark 4.7. This provides a partial explanation for the reason that OMP achieves recovery in the experiments in [6].

VI. Conclusion

In this paper we have proposed a variant of the OMP algorithm – the ϵ-OMP ($\text{OMP}_{\epsilon,2}$) – for recovering signals with sparse representations under dictionaries with pairs of highly correlated columns. We have shown, both theoretically and empirically, that $\text{OMP}_{\epsilon,2}$ succeeds in recovering such signals and that the same holds for OMP. These results are a first step for explaining its success for coherent dictionaries.

Acknowledgment

R. Giryes thanks the Azrieli Foundation for the Azrieli Fellowship. This research was supported by European Community's FP7- ERC program, grant agreement no. 320649.

References

[1] E. J. Candès, Y. C. Eldar, D. Needell, and P. Randall, "Compressed sensing with coherent and redundant dictionaries," *Appl. Comput. Harmon. Anal*, vol. 31, no. 1, pp. 59 – 73, 2011.

[2] S. Nam, M. Davies, M. Elad, and R. Gribonval, "The cosparse analysis model and algorithms," *Appl. Comput. Harmon. Anal.*, vol. 34, no. 1, pp. 30 – 56, 2013.

[3] T. Peleg and M. Elad, "Performance guarantees of the thresholding algorithm for the cosparse analysis model," *IEEE Trans. on Information Theory*, vol. 59, no. 3, pp. 1832–1845, Mar. 2013.

[4] R. Giryes, S. Nam, M. Elad, R. Gribonval, and M. E. Davies, "Greedy-like algorithms for the cosparse analysis model," *to appear in the Special Issue in Linear Algebra and its Applications on Sparse Approximate Solution of Linear Systems*, 2013.

[5] R. Rubinstein, T. Peleg, and M. Elad, "Analysis K-SVD: A dictionary-learning algorithm for the analysis sparse model," *IEEE Trans. on Signal Processing*, vol. 61, no. 3, pp. 661–677, Feb. 2013.

[6] M. A. Davenport, D. Needell, and M. B. Wakin, "Signal space CoSaMP for sparse recovery with redundant dictionaries," *CoRR*, vol. abs/1208.0353, 2012.

[7] R. Giryes and M. Elad, "Can we allow linear dependencies in the dictionary in the synthesis framework?" in *ICASSP 2013*.

[8] S. Chen, S. A. Billings, and W. Luo, "Orthogonal least squares methods and their application to non-linear system identification," *International Journal of Control*, vol. 50, no. 5, pp. 1873–1896, 1989.

[9] D. L. Donoho, M. Elad, and V. N. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Trans. Inf. Theory*, vol. 52, no. 1, pp. 6–18, Jan. 2006.

Recovery of cosparse signals with Gaussian measurements

Holger Rauhut
Hausdorff Center for Mathematics
Endenicher Allee 60
53115 Bonn, Germany
Email: rauhut@hcm.uni-bonn.de

Maryia Kabanava
Hausdorff Center for Mathematics
Endenicher Allee 60
53115 Bonn, Germany
Email: maryia.kabanava@hcm.uni-bonn.de

***Abstract*—This paper provides theoretical guarantees for the recovery of signals from undersampled measurements based on ℓ_1-analysis regularization. We provide both nonuniform and stable uniform recovery guarantees for Gaussian random measurement matrices when the rows of the analysis operator form a frame. The nonuniform result relies on a recovery condition via tangent cones and the case of uniform recovery is based on an analysis version of the null space property.**

I. Introduction

Compressed sensing is a recent field of mathematical signal processing that exploits the sparsity of a signal in order to reconstruct it from incomplete and possibly corrupted measurements. A signal $x \in \mathbb{R}^d$ is sparse, if the number of non-zero entries of x, denoted by $\|x\|_0$, is relatively small. The information about $x \in \mathbb{R}^d$ is provided by $m \ll d$ linear measurements

$$y = Mx + \varepsilon, \tag{1}$$

where $M \in \mathbb{R}^{m\times d}$ is a measurement matrix and ε corresponds to noise. Since this system is underdetermined it is impossible to recover x from y without additional information.

The most common approach for recovering x is to use regularization. This leads to an optimization problem of the form

$$\min_{z\in\mathbb{R}^d} \|Mz - y\|_2^2 + \lambda R(z).$$

The second term penalizes large values of $R(z)$ and reflects our prior knowledge on the signal to be recovered. In case of noiseless observations $\varepsilon = 0$ we rather use

$$\min_{z\in\mathbb{R}^d} R(z) \quad \text{subject to} \quad Mz = y.$$

The *analysis sparsity prior* assumes that x is sparse in some transform domain, that is, given an analysis operator $\Omega \in \mathbb{R}^{p\times d}$, the vector Ωx is sparse. Such operators can be generated by the discrete Fourier transform, the finite difference operator (related to total variation), wavelet [11], [17], [19] or curvelet transforms [3]. Then the signal is reconstructed by solving

$$\min_{z\in\mathbb{R}^d} \|\Omega z\|_1 \quad \text{subject to} \quad Mz = y. \tag{P_1}$$

Problem (P_1) often appears in image processing [2], [5]. Theoretical guarantees for the successful recovery of x via (P_1) were studied in [4], [7], [10], [13], [14], [20]. In the present paper we assume that the analysis operator is given by a frame. Put formally, let $\{\omega_i\}_{i=1}^p$, $\omega_i \in \mathbb{R}^d$, be a frame, i.e., there exist positive constants A, $B > 0$ such that for all $x \in \mathbb{R}^d$

$$A\|x\|_2^2 \le \sum_{i=1}^p |\langle \omega_i, x\rangle|^2 \le B\|x\|_2^2.$$

Its elements are collected as rows of the matrix $\Omega \in \mathbb{R}^{p\times d}$. The analysis representation of a signal x is given by the vector $\Omega x = \{\langle \omega_i, x\rangle\}_{i=1}^p \in \mathbb{R}^p$. Cosparsity is then defined as follows.

Definition 1: Let $x \in \mathbb{R}^d$, $\Omega \in \mathbb{R}^{p\times d}$ and $s = \|\Omega x\|_0$. The cosparsity of x with respect to Ω is defined as

$$l := p - s. \tag{2}$$

The index set of the zero entries of Ωx is called the cosupport of x. If x is l-cosparse, then Ωx is s-sparse with $l = p - s$. From Definition 1 it follows, that if Λ is the cosupport of x, then

$$\langle \omega_j, x\rangle = 0, \quad \forall j \in \Lambda.$$

Hence, the set of l-cosparse signals can be written as $\cup_{\#\Lambda=l} W_\Lambda$, where W_Λ is the orthogonal complement of the linear span of $\{\omega_j : j \in \Lambda\}$.

We formulate theoretical guarantees for recovery of cosparse signals (P_1) via tangent cones that are similar to the conditions stated in [6], [12]. Based on this, we are able to provide the following bound on the number of Gaussian measurements required for nonuniform recovery.

Theorem 1: Let x be l-cosparse with $l = p - s$, that is, Ωx is s-sparse. Let $M \in \mathbb{R}^{m\times d}$ be a Gaussian random matrix and $0 < \varepsilon < 1$. If

$$\frac{m^2}{m+1} \ge \frac{2Bs}{A}\left(\sqrt{\ln\frac{ep}{s}} + \sqrt{\frac{A\ln(\varepsilon^{-1})}{Bs}}\right)^2, \tag{3}$$

then with probability at least $1 - \varepsilon$, vector x is the unique minimizer of $\|\Omega z\|_1$ subject to $Mz = Mx$.

Roughly speaking, a fixed l-cosparse vector is recovered with high probability from $m > 2(B/A)s\ln(ep/s)$ Gaussian measurements. For $\Omega = \mathrm{Id}$, this bound strengthens a result in [6]. We can also incorporate the case of noisy measurements (1). But for the ease of presentation, we omit it here.

Usually, the signals to be recovered are only approximately cosparse. The quantity

$$\sigma_s(\Omega x)_1 := \inf\{\|\Omega x - z\|_1 : z \text{ is } s\text{-sparse}\}$$

describes the ℓ_1-best approximation error to Ωx by s-sparse vectors. The Ω-null space property of M to be defined below ensures stability of reconstruction. Analyzing it for Gaussian random matrices leads to the following stable and uniform recovery result.

Theorem 2: Let $M \in \mathbb{R}^{m\times d}$ be a Gaussian random matrix, $0 < \rho < 1$ and $0 < \varepsilon < 1$. If

$$\frac{m^2}{m+1} \geq \frac{2Bs\left(1+\rho^{-1}\right)^2}{A}\left(\sqrt{\ln\frac{ep}{s}} + \frac{1}{\sqrt{2}} + \sqrt{\frac{A\ln(\varepsilon^{-1})}{Bs}}\right)^2, \tag{4}$$

then with probability at least $1-\varepsilon$ for every vector $x \in \mathbb{R}^d$ a minimizer $\hat{x}$ of $\|\Omega z\|_1$ subject to $Mz = Mx$ approximates x with ℓ_2-error

$$\|x - \hat{x}\|_2 \leq \frac{2(1+\rho)^2}{\sqrt{A}(1-\rho)}\frac{\sigma_s(\Omega x)_1}{\sqrt{s}}.$$

For the standard case $\Omega = \mathrm{Id}$, this theorem improves the main result in [18] with respect to the constant and adds stability in ℓ_2.

We will give proof sketches here. Detailed arguments will be contained in [15].

We use the notation Ω_Λ to refer to a submatrix of Ω with the rows indexed by Λ. $(\Omega x)_S$ stands for the vector in $\mathbb{R}^p$ whose entries indexed by S coincide with the entries of Ωx and the rest are filled by 0. Let B_2^p denote a unit ball in $\mathbb{R}^p$ with respect to the ℓ_2-norm.

II. Nonuniform recovery from Gaussian measurements

In the present section we provide bounds on the number of measurements required for exact recovery of x by (P_1), where $M \in \mathbb{R}^{m\times d}$ is a Gaussian random matrix. We use the idea presented in [6], that requires to calculate the Gaussian widths of tangent cones.

For fixed $x \in \mathbb{R}^d$, we define the convex cone

$$T(x) = \mathrm{cone}\{z - x : z \in \mathbb{R}^d,\ \|\Omega z\|_1 \leq \|\Omega x\|_1\}.$$

Theorem 3: Let $M \in \mathbb{R}^{m\times d}$. A vector $x \in \mathbb{R}^d$ is the unique minimizer of $\|\Omega z\|_1$ subject to $Mz = Mx$ if and only if $\ker M \cap T(x) = \{0\}$.

Proof: First assume that $\ker M \cap T(x) = \{0\}$. Let $z \in \mathbb{R}^d$ be a vector that satisfies

$$\|\Omega z\|_1 \leq \|\Omega x\|_1 \ \text{ subject to } \ Mz = Mx.$$

This means that $z - x \in T(x)$ and $z - x \in \ker M$. According to our assumption we conclude that $z - x = 0$, so that x is the unique minimizer.

On the other hand, if x is the unique minimizer of (P_1), then $\|\Omega(x+v)\|_1 > \|\Omega x\|_1$ for all $v \in \ker M \setminus \{0\}$, which implies that $v \notin T(x)$. This means that

$$(\ker M \setminus \{0\}) \cap T(x) = \emptyset$$

or equivalently $\ker M \cap T(x) = \{0\}$. ∎

To prove Theorem 1 we rely on Theorem 3, which requires that the null space of the measurement matrix M misses the set $T(x)$. The next ingredient of the proof is a variation of Gordon's escape through the mesh theorem [9], which was first used in the context of compressed sensing in [18]. To formulate this theorem whose proof will be present in a journal paper in preparation, we introduce some notation.

Let $g \in \mathbb{R}^m$ be a standard Gaussian random vector. Then for

$$E_m := \mathbb{E}\,\|g\|_2 = \sqrt{2}\,\frac{\Gamma\left((m+1)/2\right)}{\Gamma\left(m/2\right)}$$

we have

$$\frac{m}{\sqrt{m+1}} \leq E_m \leq \sqrt{m}.$$

For a set $T \subset \mathbb{R}^d$ we define its Gaussian width by

$$\ell(T) := \mathbb{E}\sup_{x\in T}\langle x, g\rangle,$$

where $g \in \mathbb{R}^d$ is a standard Gaussian random vector.

Theorem 4: Let $\Omega \in \mathbb{R}^{p\times d}$ be a frame with constants A, $B > 0$. Let $M \in \mathbb{R}^{m\times d}$ be a Gaussian random matrix and T be a subset of the unit sphere $S^{d-1} = \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$. Then, for $t > 0$, it holds

$$\mathbb{P}\left(\inf_{x\in T}\|Mx\|_2 > E_m - \frac{1}{\sqrt{A}}\ell\left(\Omega(T)\right) - t\right) \geq 1 - e^{-\frac{t^2}{2}}, \tag{5}$$

where $\Omega(T)$ corresponds to the set of elements produced by applying Ω on elements from T.

With $T := T(x)\cap S^{d-1}$ the number of Gaussian measurements required to guarantee the exact reconstruction of x with probability $1 - e^{-t^2/2}$ is determined by

$$E_m \geq \frac{1}{\sqrt{A}}\ell(\Omega(T)) + t.$$

If Ω is a frame, then

$$\Omega(T) \subset \Omega(T(x)) \cap \Omega(S^{d-1}) \subset K(\Omega x) \cap \left(\sqrt{B}B_2^p\right),$$

where

$$K(\Omega x) = \mathrm{cone}\left\{y - \Omega x : y \in \mathbb{R}^p,\ \|y\|_1 \leq \|\Omega x\|_1\right\}.$$

The supremum over a larger set can only increase, hence

$$\ell(\Omega(T)) \leq \sqrt{B}\ell\left(K(\Omega x) \cap B_2^p\right). \tag{6}$$

We next give an upper bound for $\ell\left(K(\Omega x)\cap B_2^p\right)$ involving the polar cone $\mathcal{N}(\Omega x) = K(\Omega x)^\circ$ defined by

$$\mathcal{N}(\Omega x) = \{z \in \mathbb{R}^p : \langle z, y - \Omega x\rangle \leq 0 \text{ for all } y \in \mathbb{R}^p \text{ such that } \|y\|_1 \leq \|\Omega x\|_1\}.$$

Proposition 1: Let $g \in \mathbb{R}^p$ be a standard Gaussian random vector. Then

$$\ell\left(K(\Omega x)\cap B_2^p\right) \leq \mathbb{E}\min_{z\in\mathcal{N}(\Omega x)}\|g - z\|_2. \tag{7}$$

The proof is an application of convex analysis, see [1], [6]. Now the problem of estimating $\ell(\Omega(T))$ is reduced to

bounding $\mathbb{E} \min_{z\in N(\Omega(x))} \|g-z\|_2$, where Ωx is an s-sparse vector. By Hölder's inequality

$$\left(\mathbb{E} \min_{z\in\mathcal{N}(\Omega x)} \|g-z\|_2\right)^2 \le \mathbb{E} \min_{z\in\mathcal{N}(\Omega x)} \|g-z\|_2^2 \qquad (8)$$

and with some extra calculation (improving slightly over a bound in [6]) we can show that

$$\mathbb{E} \min_{z\in\mathcal{N}(\Omega x)} \|g-z\|_2^2 \le 2s \ln \frac{ep}{s}.$$

Together with inequalities (6) and (7) this gives

$$\ell(\Omega(T))^2 \le 2Bs \ln \frac{ep}{s}.$$

Proof of Theorem 1: Set $t = \sqrt{2\ln(\varepsilon^{-1})}$. The fact that $E_m \ge m/\sqrt{m+1}$ along with condition (3) yields

$$E_m \ge \frac{1}{\sqrt{A}} \ell(\Omega(T)) + t.$$

Theorem 4 implies

$$\mathbb{P}\left(\inf_{x\in T} \|Mx\|_2 > 0\right) \ge 1 - e^{-\frac{t^2}{2}} = 1-\varepsilon,$$

which guarantees that $\ker M \cap T(x) = \{0\}$ with probability at least $1-\varepsilon$. As a final step we apply Theorem 3. ∎

III. Ω-null space property

The proof of Theorem 2 is based on the following concept.

Definition 2: A matrix $M \in \mathbb{R}^{m\times d}$ is said to satisfy the ℓ_2-stable Ω-null space property of order s with $0<\rho<1$, if for any set $\Lambda \subset [p]$ with $\#\Lambda \ge p-s$ it holds

$$\|\Omega_{\Lambda^c} v\|_2 < \frac{\rho}{\sqrt{s}} \|\Omega_\Lambda v\|_1 \quad \text{for all } \ v \in \ker M \setminus \{0\}. \qquad (9)$$

This is the strengthened version of the recovery condition stated in [13]. If Ω is the identity map $\mathrm{Id} : \mathbb{R}^d \to \mathbb{R}^d$, then condition (9) becomes the standard ℓ_2-stable null space property [8].

Theorem 5: Let $\Omega \in \mathbb{R}^{p\times d}$ be a frame and $M \in \mathbb{R}^{m\times d}$ satisfy the ℓ_2-stable Ω-null space property of order s with constant $0<\rho<1$. Then for any $x \in \mathbb{R}^d$ the solution $\hat{x}$ of (P_1) with $y = Mx$ approximates the vector x with ℓ_2-error

$$\|x-\hat{x}\|_2 \le \frac{2(1+\rho)^2}{\sqrt{A}(1-\rho)} \frac{\sigma_s(\Omega x)_1}{\sqrt{s}}. \qquad (10)$$

Inequality (10) means that l-cosparse vectors are exactly recovered by (P_1) and vectors $x \in \mathbb{R}^d$, such that Ωx is close to an s-sparse vector in ℓ_1, can be well approximated in ℓ_2 by a solution of (P_1). The proof goes along the same lines as in the standard case. For the sake of brevity we omit it here.

IV. Uniform recovery from Gaussian measurements

The ℓ_2-stable Ω-null space property of order s of the measurement matrix $M \in \mathbb{R}^{m\times d}$ ensures the exact recovery of any l-cosparse vector by solving (P_1). The same strategy as in the Section II allows us to give the bound on the number of Gaussian measurements required for the ℓ_2-stable Ω-null space property to hold.

To prove Theorem 2 let us introduce the set

$$W_{\rho,s} := \left\{ w \in \mathbb{R}^d : \|\Omega_{\Lambda^c} w\|_2 \ge \rho/\sqrt{s} \|\Omega_\Lambda w\|_1 \right. \\ \left. \text{for some } \Lambda \subset [p],\ \#\Lambda = p-s \right\}.$$

If

$$\min\left\{ \|Mw\|_2 : w \in W_{\rho,s} \cap S^{d-1} \right\} > 0, \qquad (11)$$

then for all $w \in \ker M \setminus \{0\}$ and any $\Lambda \subset [p]$ with $\#\Lambda = p-s$ we have

$$\|\Omega_{\Lambda^c} w\|_2 < \frac{\rho}{\sqrt{s}} \|\Omega_\Lambda w\|_1,$$

which implies that M satisfies the ℓ_2-stable Ω-null space property of order s. To show (11) we apply Theorem 4, according to which we have to study the Gaussian width of the set $\Omega\left(W_{\rho,s} \cap S^{d-1}\right)$. Since Ω is a frame, we have

$$\Omega\left(W_{\rho,s} \cap S^{d-1}\right) \subset \Omega\left(W_{\rho,s}\right) \cap \left(\sqrt{B} B_2^p\right) \subset T_{\rho,s} \cap \left(\sqrt{B} B_2^p\right),$$

with

$$T_{\rho,s} = \left\{ w \in \mathbb{R}^p : \|w_S\|_2 \ge \rho/\sqrt{s} \|w_{S^c}\|_1 \right. \\ \left. \text{for some } S \subset [p],\ \#S = s \right\}.$$

Then

$$T_{\rho,s} \cap \left(\sqrt{B} B_2^p\right) = \bigcup_{\#S=s} \left\{ w \in \mathbb{R}^p : \|w\|_2 \le \sqrt{B}, \right. \\ \left. \|w_S\|_2 \ge \frac{\rho}{\sqrt{s}} \|w_{S^c}\|_1 \right\}.$$

Lemma 1: Let the set D be defined by

$$D := \operatorname{conv}\left\{ x \in S^{p-1} : \#\operatorname{supp} x \le s \right\}.$$

Then

$$T_{\rho,s} \cap \left(\sqrt{B} B_2^p\right) \subset \left(1+\rho^{-1}\right) \left(\sqrt{B} D\right). \qquad (12)$$

A similar result was stated as Lemma 4.5 in [18], so we omit the proof.

Lemma 1 implies that

$$\ell\left(T_{\rho,s} \cap \left(\sqrt{B} B_2^p\right)\right) \le \sqrt{B}\left(1+\rho^{-1}\right) \ell(D).$$

Lemma 2: The Gaussian width of D satisfies

$$\ell(D) \le \sqrt{2s \ln \frac{ep}{s}} + \sqrt{s}.$$

Proof: Due to the definition of the Gaussian width

$$\ell(D) = \mathbb{E} \sup_{x\in D} \langle g, x\rangle = \mathbb{E} \sup_{\substack{\|x\|_2=1,\\ \#\operatorname{supp} x \le s}} \langle g, x\rangle, \qquad (13)$$

where $g \in \mathbb{R}^p$ is a standard Gaussian random vector. Hölder's inequality applied to (13) and an estimate on the maximum

squared ℓ_2-norm of a sequence of standard Gaussian random vectors (see e.g. [16, Lemma 3.2]) give

$$\ell(D) \leq \mathbb{E} \max_{S\subset[p],\#S=s} \|g_S\|_2 \leq \sqrt{\mathbb{E} \max_{S\subset[p],\#S=s} \|g_S\|_2^2}$$
$$\leq \sqrt{2\ln\binom{p}{s}} + \sqrt{s} \leq \sqrt{2s\ln\frac{ep}{s}} + \sqrt{s}.$$

The last inequality follows from the fact that

$$\binom{p}{s} \leq \left(\frac{ep}{s}\right)^s.$$

■

Proof of Theorem 2: The reasoning above shows that

$$\ell\left(\Omega\left(W_{\rho,s}\cap S^{d-1}\right)\right) \leq \sqrt{B}\left(1+\rho^{-1}\right)\ell(D)$$
$$\leq \sqrt{B}\left(1+\rho^{-1}\right)\left(\sqrt{2s\ln\frac{ep}{s}}+\sqrt{s}\right).$$

Set $t = \sqrt{2\ln(\varepsilon^{-1})}$. The fact that $E_m \geq m/\sqrt{m+1}$ along with condition (4) yields

$$E_m \geq \frac{1}{\sqrt{A}} l\left(\Omega\left(W_{\rho,s}\cap S^{d-1}\right)\right) + t.$$

Theorem 4 implies

$$\mathbb{P}\left(\inf\|Mw\|_2 > 0 : w \in W_{\rho,s}\cap S^{d-1}\right)$$
$$\geq 1 - e^{-\frac{t^2}{2}} = 1-\varepsilon,$$

which guarantees

$$\|\Omega_{\Lambda^c} w\|_2 < \frac{\rho}{\sqrt{s}}\|\Omega_\Lambda w\|_1$$

for all $w \in \ker M\backslash\{0\}$ and any $\Lambda\subset[p]$ with $\#\Lambda = p-s$. This means that M satisfies the ℓ_2-stable Ω-null space property of order s. Finally, apply Theorem 5. ■

V. Uniform recovery from Gaussian measurements

In this work we provided conditions that guarantee the uniqueness of the solution of the optimization problem (P_1), when the analysis operator is given by a frame. The modification of the Gordon's escape through the mesh theorem allowed to derive a bound on the number of Gaussian random measurements needed to satisfy these conditions.

References

[1] S. Boyd, L. Vandenberghe. Convex optimization. Cambridge University Press, Cambridge, 2004.

[2] J.-F. Cai, S. Osher, Z. Shen. Split Bregman methods and frame based image restoration. *Multiscale Model. Simul.*, 8(2):337–369, 2009/10.

[3] E. J. Candès, D. L. Donoho. New tight frames of curvelets and optimal representations of objects with piecewise C^2 singularities. *Comm. Pure Appl. Math.*, 57(2):219–266, 2004.

[4] E. J. Candès, Y. C. Eldar, D. Needell, P. Randall. Compressed sensing with coherent and redundant dictionaries. *Appl. Comput. Harmon. Anal.*, 31(1):59–73, 2011.

[5] T. Chan, J. Shen. Image Processing and Analysis: Variational, PDE, Wavelet, and Stochastic Methods. SIAM, 2005.

[6] V. Chandrasekaran, B. Recht, P. A. Parrilo, A. S. Willsky. The Convex Geometry of Linear Inverse Problems. *Found. Comput. Math.*, 12(6):805–849, 2012.

[7] M. Elad, P. Milanfar, R. Rubinstein. Analysis versus synthesis in signal priors. *Inverse Problems*, 23(3):947–968, 2007.

[8] S. Foucart, H. Rauhut. A Mathematical Introduction to Compressive Sensing. Birkhäuser, to appear.

[9] Y. Gordon. On Milman's inequality and random subspaces which escape through a mesh in $\mathbf{R}^n$. In *Geometric aspects of functional analysis (1986/87)*, volume 1317 of *Lecture Notes in Math.*, pages 84–106. Springer, Berlin, 1988.

[10] Y. Liu, T. Mi, Sh. Li. Compressed sensing with general frames via optimal-dual-based ℓ_1-analysis. *IEEE Transactions on information theory*, 58(7):4201–4214, 2012.

[11] S. Mallat. A Wavelet Tour of Signal Processing: The Sparse Way. Academic Press, 2008.

[12] S. Mendelson, A. Pajor, N. Tomczak-Jaegermann. Reconstruction and subgaussian operators in asymptotic geometric analysis. *Geom. Funct. Anal.*, 17(4):1248–1282, 2007.

[13] S. Nam, M.E. Davies, M. Elad, R. Gribonval. The cosparse analysis model and algorithms. *Appl. Comput. Harmon. Anal.*, 34(1):30–56, 2013.

[14] D. Needell, R. Ward. Stable image reconstruction using total variation minimization. http://arxiv.org/abs/1202.6429

[15] H.Rauhut, M. Kabanava. Analysis ℓ_1–recovery with frames and Gaussian measurements. In preparation.

[16] N. Rao, B. Recht, R. Nowak. Tight measurement bounds for exact recovery of structured sparse signals. In Proceedings of AISTATS, 2012.

[17] A. Ron, Z. Shen. Affine systems in $L_2(R^d)$: the analysis of the analysis operator. *J. Funct. Anal.*, 148(2):408–447, 1997.

[18] M. Rudelson, R. Vershynin. On sparse reconstruction from Fourier and Gaussian measurements. *Comm. Pure Appl. Math.*, 61(8):1025–1045, 2008.

[19] I. Selesnick, M. Figueiredo. Signal restoration with overcomplete wavelet transforms: comparison of analysis and synthesis priors. Proceedings of SPIE, vol. 7446, 2009, p. 74460D.

[20] S. Vaiter, G. Peyré, Ch. Dossal, J. Fadili. Robust sparse analysis regularization. *IEEE Transactions on information theory*, 59(4):2001–2016, 2013.

q-ary Compressive Sensing

Youssef Mroueh
LCSL/CBCL - CSAIL, MIT, IIT
Email: ymroueh@mit.edu

Lorenzo Rosasco
DIBRIS, Unige and LCSL - MIT, IIT
Email: lrosasco@mit.edu

***Abstract*—We introduce q-ary compressive sensing, an extension of 1-bit compressive sensing. We propose a novel sensing mechanism and a corresponding recovery procedure. The recovery properties of the proposed approach are analyzed both theoretically and empirically. Results in 1-bit compressive sensing are recovered as a special case. Our theoretical results suggest a tradeoff between the quantization parameter q and the number of measurements m, in controlling the error and robustness to noise of the resulting recovery algorithm.**

I. Introduction

Reconstructing signals from discrete measurements is a classic problem in signal processing. Properties of the signal allow the reconstruction from a minimal set of measurements. The classical Shannon sampling result ensures that band limited signals can be reconstructed by a linear procedure, as long as a number of linear measurements, at least twice the maximum frequency, is available. Modern data analysis typically requires recovering high dimensional signals from few inaccurate measurements. Indeed, the development of Compressed Sensing (CS) and Sparse Approximation [1] shows that this is possible for signals with further structure. For example, d-dimensional, s-sparse signals[1] can be reconstructed with high probability through convex programming, given $m \sim s \log(d/s)$ random linear measurements.

Non-linear measurements have been recently considered in the context of 1-bit compressive sensing [2], [3], [4], [5], [6] (`http://dsp.rice.edu/1bitCS/`). Here, binary (one-bit) measurements are obtained by applying, for example, the "sign" function[2] to linear measurements. More precisely, given $x \in \mathbb{R}^d$, a measurement vector is given by $y = (y_1, \ldots, y_m)$, where $y_i = sign(\langle w_i, x \rangle)$ with $w_i \sim \mathcal{N}(0, I_d)$ independent Gaussian random vectors, for $i = 1, \ldots, m$. It is possible to prove [4] that, for a signal $x \in K \cap \mathbb{B}^d$ ($\mathbb{B}^d$ is the unit ball in $\mathbb{R}^d$), the solution $\hat{x}_m$ to the problem

$$\max_{x \in K} \sum_{i=1}^{m} y_i \langle w_i, x \rangle , \tag{1}$$

satisfies $\|\hat{x}_m - x\|^2 \leq \frac{\delta}{\sqrt{\frac{2}{\pi}}}$, with probability $1 - 8\exp(-c\delta^2 m)$, $\delta > 0$, as long as $m \geq C\delta^{-2}\omega(K)^2$ [4]. Here, C denotes a universal constant and $\omega(K) = \mathbb{E}\sup_{x \in K-K} \langle w, x \rangle$ the Gaussian mean width of K, which can be interpreted as a complexity measure. If K is a convex set, problem (1) can be solved efficiently.

In this paper, borrowing ideas from signal classification and machine learning, we discuss a novel sensing strategy, based on *q-ary nonlinear measurements*, and a corresponding recovery procedure.

II. q-ary Compressive Sensing

A. Sensing and Recovery

The sensing procedure we consider is given by a map C from $K \cap \mathbb{B}^d$ to $\mathcal{F} = \{0, \ldots, q-1\}^m$, where $K \subset \mathbb{R}^d$. To define C we need the following definitions.

Definition 1 (Simplex Coding [7])**.** *The simplex coding map is $S : \{0, \ldots, q-1\} \to \mathbb{R}^{q-1}$, $S(j) = s_j$, where*
1) $\|s_j\|^2 = 1$,
2)$\langle s_j, s_i \rangle = -\frac{1}{q-1}$, for $i \neq j$,
3) $\sum_{j=0}^{q-1} s_j = 0$.

Definition 2 (q-ary Quantized Measurements)**.** *Let $W \in \mathbb{R}^{q-1,d}$ be a Gaussian random matrix, i.e. $W_{ij} \sim \mathcal{N}(0,1)$ for all i, j. Then, $Q : K \cap \mathbb{B}^d \to \{0, \ldots, q-1\}$,*

$$Q(x) = Q_W(x) = \arg\max_{j=0\ldots q-1} \langle s_j, Wx \rangle ,$$

is called a q-ary quantized measurement.

Then, we can define the q-ary sensing strategy induced by non-linear quantized measurements.

Definition 3 (q-ary Sensing)**.** *Let $W_1, \ldots, W_m$, be independent Gaussian random matrices in*

[1] A d-dimensional signal, that is a vector in $\mathbb{R}^d$, is s-sparse if only s of its components are different from zero.

[2] More generally, any function $\theta : \mathbb{R} \to [-1, 1]$, such that $\mathbb{E}(g\theta(g) > 0)$ can be used.

$\mathbb{R}^{q-1,d}$ *and* $Q_{W_i}(x), i = 1, \dots, m$ *as in Def. 2. The* q*-ary sensing is* $C : K \cap \mathbb{B}^d \to \{0, \dots, q-1\}^m$,

$$C(x) = (Q_{W_1}(x), \dots Q_{W_m}(x)),$$

$\forall x \in K \cap \mathbb{B}^d$.

Before describing the recovery strategy we consider, we discuss the connection to 1-bit CS and binary embeddings [8] [6] .

Remark 1 (Connection to 1-bit CS)**.** *If* $q = 2$*,* W *reduces to a Gaussian random vector, and* $2Q(x) - 1 = sign(Wx)$*, so that the* q*-ary quantized measurements become equivalent to those considered in* 1*-bit CS.*

Remark 2 (Sensing and Embeddings)**.** *It can be shown that* C *defines an* ϵ*-isometric embedding of* $(K, \|\cdot\|)$ *into* $(\mathcal{F}, d_H)$ *– up-to a bias term. Here* d_H *is the (normalized) Hamming distance,* $d_H(u, v) = \frac{1}{m}\sum_{i=1}^m \mathbb{1}_{u_i \neq v_i}, u, v \in \mathcal{F}$*. This analysis is deferred to the long version of this paper.*

In this paper, we are interested in provably (and efficiently) recovering a signal x from its q-ary measurements $y = (y_1, \dots, y_m) = C(x)$. Following [4], we consider the recovery strategy $D : \{0, \dots, q-1\}^m \to K \cap \mathbb{B}^d$ defined by,

$$D(y) = \arg\max_{u \in K \cap \mathbb{B}^d} \frac{1}{m} \sum_{i=1}^m \langle s_{y_i}, W_i u \rangle . \qquad (2)$$

The above problem is convex if K is convex and can be solved efficiently, see Section III-A. In the next section, we prove that the solution to Problem (2) has good recovery guarantees both in noiseless and noisy settings.

Remark 3 (Connection to Classification)**.** *The inspiration for considering* q*-ary CS stems from an analogy between* 1*-bit compressed sensing and binary classification in machine learning. In this view, Definition* (3) *is related to the approach proposed for multi category classification in [7]. Following these ideas, we can extend the recovery strategy* (2) *by considering*

$$D_V(y) = \arg\min_{u \in K \cap \mathbb{B}^d} \frac{1}{m} \sum_{i=1}^m V(-\langle s_{y_i}, W_i u \rangle), \qquad (3)$$

where V *is a convex, Lipschitz, non-decreasing loss function* $V : \mathbb{R} \to \mathbb{R}^+$*. Problem* (2) *corresponds to the choice* $V(x) = x$*. Other possible choices include* $V(x) = \max(1+x, 0)$*,* $V(x) = \log(1 + e^x)$*, and* $V(x) = e^x$*. Each of these loss functions can be seen as convex relaxations of the* 0-1 *loss function, defined as* $V(x) = 0$ *if* $x \leq 0$*, and* 1 *otherwise. The* $0 - 1$ *loss defines the misclassification risk, which corresponds to Hamming distance in CS, and is a natural measure of performance while learning classification rules.*

Remark 4 (Recovery of Distorted Signals)**.** *We note that the* q*-ary approach could be of particular interest in situations where the signals can undergo unknown non-linear distortions, because of the robustness of the maximum in the definition of the* q*-ary measurements.*

B. Recovery guarantees: Noiseless Case

The following theorem describes the recovery guarantees for the proposed procedure for signals in a set K of Gaussian mean width $w(K)$. We first consider a noiseless scenario.

Theorem 1. *Let* $\delta > 0$*,* $m \geq C\delta^{-2}w(K)^2$*. Then with probability at least* $1 - 8\exp(-c\delta^2 m)$*, the solution* $\hat{x}_m = D(y)$ *of problem* (2) *satisfies,*

$$\|\hat{x}_m - x\|^2 \leq \frac{\delta}{\sqrt{\log(q)}}. \qquad (4)$$

A proof sketch of the above result is given in Section II-D, while the complete proof is deferred to the long version of the paper. Here, we add four comments. First, it can be shown the the above result bound is derived from an error bound,

$$||\hat{x}_m - x||^2 \leq C(\frac{w(K)}{\sqrt{\log(q)m}} + t), \qquad (5)$$

with probability at least, $1 - 4\exp(-2t^2), t > 0$.

Second, Inequalities (4), (5) can be compared to results in 1-bit CS. For the same number of measurements, $m \geq C\delta^{-2}w(K)^2$, the error for q-ary CS is $\frac{\delta}{\sqrt{\log(q)}}$, in contrast with $\frac{\delta}{\sqrt{\frac{2}{\pi}}}$ in the 1-bit CS [4], at the expense of a more demanding sensing procedure. Also note that, for $q = 2$, we recover the result in 1-bit CS as a special case. Third, we see that for a given accuracy our results highlight a trade-off between the number of q-ary measurements m and the quantization parameter q. To achieve an error ϵ with a memory budget of ℓ bits, one can choose m and q so that $\epsilon = O(\frac{1}{\sqrt{m\log(q)}})$, and $m\log_2(q) = \ell$ (see also section III-B). Finally, in the following we will be interested in K being the set of s-sparse signals. Following again [4], it is interesting to consider in Problem (2) the relaxation

$$K_1 = \{u \in \mathbb{R}^d : \|u\|_1 \leq \sqrt{s}, \|u\|_2 \leq 1\}.$$

With this choice, it it possible to prove that $w(K_1) \leq C\sqrt{s\log(\frac{2d}{s})}$, and that for $m \geq C\delta^{-2}s\log(\frac{2d}{s})$, the solution of the convex program (2) on K_1 satisfies, $\|\hat{x}_m - x\|^2 \leq \frac{\delta}{\sqrt{\log(q)}}$. We end noting that other choices of K are possible, for example in [9] the set of group sparse signals (and its Gaussian width) is studied.

C. Recovery Guarantees: Noisy Case

Next we discuss the q-ary approach in two noisy settings, related to those considered in [4].

Noise before quantization. For $i = 1, \dots, m$, let

$$y_i = \arg\max_{j=0\dots q-1} \{\langle s_j, W_i x\rangle + g_j\}, \quad (6)$$

with g_j independent Gaussian random variables of variance σ^2. In this case, it is possible to prove that, for $m \geq C\delta^{-2}w(K)^2$,

$$\|\hat{x}_m - x\|^2 \leq \frac{\delta\sqrt{1+\sigma^2}}{\sqrt{\log(q)}},$$

with probability at least $1-8\exp(-c\delta^2 m)$. The quantization level q can be chosen to adjust to the noise level σ for a more robust recovery of x. This result can be viewed in the perspective of the *bit-depth versus measurement-rates* perspective studied in [10], where it is shown that 1-bit CS outperforms conventional scalar quantization. In this view, $q-$ary CS provides a new way to adjust the quantization parameter to the noise level.

Inexact maximum. For $i = 1, \dots, m$, let $y_i = Q_{W_i}(x)$, with probability p, and $y_i = r$ with probability $1-p$, with r drawn uniformly at random from $\{0, \dots, q-1\}$. In this case, it is possible to prove that, for $m \geq C\delta^{-2}w(K)^2$,

$$\|\hat{x}_m - x\|^2 \leq \frac{\delta}{\sqrt{\log(q)}(2p-1)}.$$

with probability at least $1-8\exp(-c\delta^2 m)$. The signal x can be recovered even if *nearly half* of the q-ary bits are flipped.

D. Elements of the proofs

We sketch the main steps in proving our results. The proof of Theorem 1 is based on: 1) deriving a bound in expectation, and 2) deriving a concentration result. The proof of the last step uses Gaussian concentration inequality extending the proof strategy in [4]. Step 1) gives the bound

$$\mathbb{E}\left(||\hat{x}_m - x||^2\right) \leq \frac{w(K)}{C\sqrt{\log(q)m}},$$

the proof of which is based on the following proposition.

Proposition 1. *Let $\mathcal{E}_x(u) = \mathbb{E}_W(\langle s_\gamma, Wu\rangle)$, where $\gamma = Q_W(x)$. Then, $\forall u \in \mathbb{B}^d$, we have,*

$$\frac{1}{2}\|u - x\|^2 \leq \frac{1}{\lambda(q)}\left(\mathcal{E}_x(x) - \mathcal{E}_x(u)\right),$$

where $\lambda(q) = \mathbb{E}_{\bar{\gamma},g}(\langle s_{\bar{\gamma}}, g\rangle)$, and $g \sim \mathcal{N}(0, I_{q-1})$, and $\bar{\gamma} = \arg\max_{j=0\dots q-1}\langle s_j, g\rangle$.

Using results in empirical process theory it possible to show that

$$|\mathcal{E}_x(x) - \mathcal{E}_x(\hat{x}_m)| \leq C\frac{w(K)}{\sqrt{m}}.$$

The bound on the expected recovery follows combining the above inequality and Proposition 1 with the inequality,

$$\lambda(q) \geq C\sqrt{\log(q)},$$

which is proved using Slepian's inequality and Sudakov minoration.

Results in the noisy settings follow from suitable estimates of $\lambda(q)$. Indeed, for the *noise before quantization* case it can be proved that $\lambda(q) \geq C\sqrt{\frac{\log(q)}{1+\sigma^2}}$. For the *inexact maximum* case one has

$$\begin{aligned}\lambda(q) \quad &= \mathbb{E}_{y,g}(\langle s_y, g\rangle) = \\ &p\mathbb{E}(\max_{j=1\dots q}\langle s_j, g\rangle) + (1-p)\mathbb{E}(\langle s_r, g\rangle) \geq \\ &Cp\sqrt{\log(q)} + (1-p)\mathbb{E}(\min_{j=1\dots q}\langle s_j, g\rangle) \geq \\ &(2p-1)C\sqrt{\log(q)}.\end{aligned}$$

III. Experimental Validation

A. An Algorithm for Sparse recovery

In our experiments, we considered the following variation of problem (2), Let $\xi_i = s_{y_i}^\top W_i \in \mathbb{R}^d, i = 1 \dots m$.

$$\max_{u, \|u\|_2 \leq 1} \frac{1}{m}\sum_{i=1}^m \langle \xi_i, u\rangle - \eta\|u\|_1, \quad (7)$$

where $\eta > 0$. The above problem can be solved efficiently using Proximal Methods [11]. Indeed, a solution can be computed via the iteration,

$$\begin{aligned}u_{t+1} &= u_t + \frac{\nu_t}{m}\sum_{i=1}^m \xi_i, \\ u_{t+1} &= Prox_\eta(u_{t+1}), \\ u_{t+1} &= u_{t+1}\min(\frac{1}{\|u_{t+1}\|_2}, 1).\end{aligned}$$

Where ν_t is the gradient step size, and $Prox_\eta$ acts component-wise as $\max(1-\frac{\eta}{|u_i|}, 0)u_i$. The iteration is initialized randomly to a unit vector.

Remark 5. *The computational complexity of the sensing process depends on both m and q, while, once computed ξ_i, that of the recovery algorithm depends only on m, and is the same as in 1-bit CS. In this sense, given a bit rate, the same precision can achieved by 1-bit CS and q-ary CS, with a better computational complexity for the decoding in the q-ary case.*

B. Sparse Recovery

We tested our approach for recovering a signal from from its q-ary measurements. We considered sparse signals of dimension d generated via a Gauss-Bernoulli model. In Figure 1.(a), we see that the reconstruction error $\hat{x}_m$ (in blue), for varying m and q fixed, follows the theoretical bound $\frac{1}{\sqrt{m}}$ (in red). In Figure 1.(b), we see that the reconstruction error of $\hat{x}_m$ (in blue), for varying q and m fixed, follows the theoretical bound $\frac{1}{\sqrt{\log(q)}}$ (in red). Figures 1.(c), and 1.(d) highlight the tradeoff between the number of measurements and the quantization parameter. For a precision ϵ, and a memory budget 2^B, one can choose an operating point (m, q), according to the theoretical bound $\frac{1}{\sqrt{m \log(q)}}$.

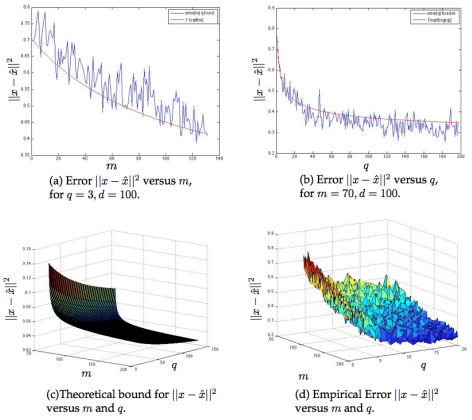


Fig. 1. q-ary Compressive Sensing: Quantization/Number of measurements tradeoff.

C. Image Reconstruction

We considered the problem of recovering an image from q-ary measurements. We used the $8-$bit grayscale boat image of size 64×64 pixels shown in Figure 2(a). We extracted and thresholded the wavelet coefficients to get a sparse signal. We normalized the resulting vector of wavelets coefficients of dimension $d = 3840$ to obtain a unit vector. Then, we performed sensing and recovery with $q = 2^5$ (5-bit compressive sensing) and $q = 2$ (1-bit compressive sensing) for the same $m = 2048 < d$. We compared the SNR of the corresponding reconstructed images in noiseless (Figures 2(b)-(c)), and noisy settings (noise before quantization model (6), with $\sigma = 0.8$), Figures 2(d)-(e). Note that in this setting we are comparing 1-bit CS and q-ary CS, for the same decoding time (same m). The results confirm our theoretical findings: higher quantization improves the SNR, as well as robustness to noise of q-ary CS.

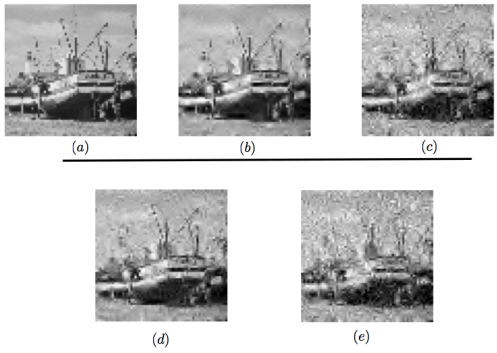


Fig. 2. Image recovery with q-ary CS. (a) Original image. (b) Reconstruction with no-noise: $q = 2^5$, SNR $= 20.2$ dB. (c) Reconstruction with no-noise: $q = 2$, SNR $= 16.2$ dB. (d) Reconstruction with noise: $q = 2^5$, SNR $= 18.3$ dB. (e) Reconstruction with noise: $q = 2$, SNR $= 15$ dB.

REFERENCES

[1] Y.C. Eldar, and G. Kutyniok Eds, Compressed Sensing. Cambridge University Press, 2012.

[2] P. T. Boufounos and R. G. Baraniuk. 1-bit Compressive Sensing, in Proceedings of Conference on Information Science and Systems (CISS), Princeton, NJ, March 2008.

[3] J. Z. Sun and V. K. Goyal, Optimal Quantization of Random Measurements in Compressed Sensing, Proc. IEEE Int. Symp. Information Theory (Seoul, Korea), June-July 2009.

[4] Y. Plan and R. Vershynin, Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach, IEEE Transactions on Information Theory, to appear.

[5] Plan, Y. and Vershynin, R. (2013), One-bit compressed sensing by linear programming. Comm. Pure Appl. Math.

[6] L. Jacques, J. N. Laska, P. T. Boufounos, and R. G. Baraniuk, Robust 1-Bit Compressive Sensing via Binary Stable Embeddings of Sparse Vectors, IEEE Trans. Info. Theory, v. 59, no. 4, April, 2013.

[7] Y. Mroueh, T. Poggio, L. Rosasco, and J.J Slotine, Multiclass learning with simplex coding, NIPS 2012.

[8] Y. Plan and R. Vershynin, Dimension reduction by random hyperplane tessellations, submitted 2011.

[9] N. Rao, R. Nowak and B. Recht, Tight Measurement Bounds for Exact Recovery of Structured Sparse Signals, Arxiv preprint,2011.

[10] J. Laska, and R G. Baraniuk, Regime Change: Bit-Depth versus Measurement-Rate in Compressive Sensing, Arxiv preprint, available online at http://arxiv.org/abs/1110.3450, 2011.

[11] S. Boyd and L. Vandenberghe. Convex Optimization. Cambridge University Press, March 2004. ISBN 0521833787.

Low-rank Tensor Recovery via Iterative Hard Thresholding

Holger Rauhut
RWTH Aachen University
Templergraben 55,
52056 Aachen, Germany
Email: rauhut@mathc.rwth-aachen.de

Reinhold Schneider
Technische Universität Berlin
Straße des 17. Juni 136,
10623 Berlin, Germany
Email: schneidr@math.tu-berlin.de

Željka Stojanac
Hausdorff Center for Mathematics
Institut für Numerische Simulation
Endenicher Allee 62,
53115 Bonn, Germany
Email: zeljka.stojanac@hcm.uni-bonn.de

***Abstract*—We study recovery of low-rank tensors from a small number of measurements. A version of the iterative hard thresholding algorithm (TIHT) for the higher order singular value decomposition (HOSVD) is introduced. As a first step towards the analysis of the algorithm, we define a corresponding tensor restricted isometry property (HOSVD-TRIP) and show that Gaussian and Bernoulli random measurement ensembles satisfy it with high probability.**

I. Introduction and Motivation

Low-rank recovery builds on ideas from the theory of compressive sensing which predicts that sparse vectors can be recovered efficiently from incomplete measurements via efficient algorithms including ℓ_1-minimization. Given a matrix $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2}$ of rank at most $r \ll \min\{n_1, n_2\}$, the goal of the low-rank matrix recovery is to reconstruct $\mathbf{X}$ from linear measurements $\mathbf{y} = \mathcal{A}(\mathbf{X})$, where $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$ with $m \ll n_1 n_2$. Unfortunately, the natural approach of finding the solution of the optimization problem

$$\min_{\mathbf{Z} \in \mathbb{R}^{n_1 \times n_2}} \operatorname{rank}(\mathbf{Z}) \quad \text{s.t.} \quad \mathcal{A}(\mathbf{Z}) = \mathbf{y}, \tag{1}$$

is NP-hard. Nevertheless, it has been shown that solving the convex optimization problem

$$\min_{\mathbf{Z} \in \mathbb{R}^{n_1 \times n_2}} \|\mathbf{Z}\|_* \quad \text{s.t.} \quad \mathcal{A}(\mathbf{Z}) = \mathbf{y}, \tag{2}$$

reconstructs $\mathbf{X}$ exactly under suitable conditions on $\mathcal{A}$. The required number of measurements scales as $m \geq Cr \max\{n_1, n_2\}$ for Gaussian measurement ensembles [11], [2].

In this note, we go one step further and consider the recovery of low-rank tensors $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ from a small number of linear measurements $\mathbf{y} = \mathcal{A}(\mathbf{X})$, where $\mathcal{A} : \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d} \to \mathbb{R}^m$ and $m \ll n_1 n_2 \cdots n_d$. Again, we are led to consider the rank-minimization problem

$$\min_{\mathbf{Z} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}} \operatorname{rank}(\mathbf{Z}) \quad \text{s.t.} \quad \mathbf{y} = \mathcal{A}(\mathbf{Z}). \tag{3}$$

Different notions of the tensor rank have been introduced, which correspond to different decompositions. One possibility is to define the rank of an arbitrary tensor $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$, analogously to the matrix rank, as the smallest number of rank one tensors that sum up to $\mathbf{X}$, where a rank one tensor is of the form $\mathbf{A} = \mathbf{u}_1 \otimes \mathbf{u}_2 \otimes \cdots \otimes \mathbf{u}_d$ or elementwise $\mathbf{A}(i_1, i_2, \ldots, i_d) = \mathbf{u}_1(i_1)\, \mathbf{u}_2(i_2) \cdots \mathbf{u}_d(i_d)$. Expectedly, the problem (3) is NP hard [8]. Although it is possible to define an analog of the nuclear norm $\|\cdot\|_*$ for tensors and consider the minimization problem

$$\min_{\mathbf{Z} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}} \|\mathbf{Z}\|_* \quad \text{s.t.} \quad \mathbf{y} = \mathcal{A}(\mathbf{Z}),$$

the computation of $\|\cdot\|_*$ and thereby this problem is NP hard [8] as well for tensors of order $d \geq 3$.

The previous approaches to low-rank tensor recovery and tensor completion [3] and [9] are based on the sum of nuclear norms of matrices obtained as unfoldings of the tensor (see below for the notion of unfolding). Only numerical experiments have been performed in these papers and at least from a theoretical point of view, we do not believe this to be the right approach since the tensor structure is lost.

We consider a generalization of the singular value decomposition, called HOSVD (higher order singular value decomposition). This decomposition is used in e.g. data mining for handwritten digit classification [12], in signal processing to extend Wiener filters [10], in computer vision [13] and in other applications.

As a proxy for (3) we propose an extension of the IHT algorithm already used for recovery of sparse signals [1] and low-rank matrices [5]. The tensor iterative hard thresholding algorithm (TIHT algorithm) is presented in Section IV. In the last section, we introduce the tensor restricted isometry property (HOSVD-TRIP) and also show that random linear mappings satisfy the HOSVD-TRIP with high probability, under suitable conditions.

The version for the tensor train decomposition (TT decomposition) and hierarchical tucker format (HT decomposition) will be treated in a journal paper in preparation.

II. Notation

We work with tensors $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ of order d. With $\mathbf{X}_{i_k = p}$, for all $p \in [n_k]$, where $[n_k] = \{1, 2, \ldots, n_k\}$, we denote the $(d-1)$-dimensional tensor (called subtensor) that is obtained by fixing the k-th component of a tensor $\mathbf{X}$ to p i.e., $\mathbf{X}_{i_k = p}(i_1, \ldots, i_{k-1}, i_{k+1}, \ldots, i_d) = \mathbf{X}(i_1, \ldots, i_{k-1}, p, i_{k+1}, \ldots, i_d)$, for all $i_l \in [n_l]$ and for all $l \in [d] \setminus \{k\}$. A matrix obtained by taking the first r_k

columns of the matrix $\mathbf{U}$ is denoted by $\mathbf{U}(:,[r_k])$. Similary, $\mathbf{S}([r_1],[r_2],\ldots,[r_d]) \in \mathbb{R}^{r_1\times r_2\times\cdots\times r_d}$ is defined elementwise as $\mathbf{S}([r_1],[r_2],\ldots,[r_d])(i_1,i_2,\ldots,i_d) = \mathbf{S}(i_1,i_2,\ldots,i_d)$, for all $i_k \in [r_k]$ and for all $k \in [d]$.

Matrices will be denoted with capital bold letters, linear mappings with caligraphic capital letters and vectors with small bold letters.

The inner product of two tensors $\mathbf{X},\mathbf{Y} \in \mathbb{R}^{n_1\times n_2\times\cdots\times n_d}$ is defined as

$$\langle \mathbf{X},\mathbf{Y}\rangle = \sum_{i_1=1}^{n_1}\sum_{i_2=1}^{n_2}\cdots\sum_{i_d=1}^{n_d}\mathbf{X}(i_1,i_2,\ldots,i_d)\,\mathbf{Y}(i_1,i_2,\ldots,i_d).$$

The (Frobenius) norm of a tensor $\mathbf{X} \in \mathbb{R}^{n_1\times n_2\times\cdots\times n_d}$, $\|\mathbf{X}\|_F$, induced by this inner product, is

$$\|\mathbf{X}\|_F = \sqrt{\sum_{i_1=1}^{n_1}\sum_{i_2=1}^{n_2}\cdots\sum_{i_d=1}^{n_d}\mathbf{X}^2(i_1,i_2,\ldots,i_d)}.$$

Matricization (unfolding) is the operation that transforms a tensor into a matrix. The mode-k matricization of a tensor $\mathbf{X} \in \mathbb{R}^{n_1\times n_2\times\cdots\times n_d}$ is denoted by $\mathbf{X}_{(k)}$, $\mathbf{X}_{(k)} \in \mathbb{R}^{n_k\times n_1\cdots n_{k-1}n_{k+1}\cdots n_d}$. The rows of the matrix $\mathbf{X}_{(k)}$ are determined by the k-th component of the tensor $\mathbf{X}$, whereas all the remaining components determine its column, i.e.,

$$\mathbf{X}_{(k)}(i_k;(i_1,\ldots,i_{k-1},i_{k+1},\ldots,i_d)) = \mathbf{X}(i_1,\ldots,i_d).$$

For $\mathbf{X} \in \mathbb{R}^{n_1\times n_2\times\cdots\times n_d}$, $\mathbf{A} \in \mathbb{R}^{J\times n_k}$ and $k \in [d]$, the k-mode multiplication, $\mathbf{X}\times_k\mathbf{A} \in \mathbb{R}^{n_1\times\cdots\times n_{k-1}\times J\times n_{k+1}\times\cdots\times n_d}$ is defined elementwise as

$$(\mathbf{X}\times_k\mathbf{A})(i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_d) = \\ = \sum_{i_k=1}^{n_k}\mathbf{X}(i_1,\ldots,i_d)\,\mathbf{A}(j,i_k).$$

Remark 1: Notice that the SVD decomposition of a matrix $\mathbf{X} \in \mathbb{R}^{n_1\times n_2}$ can be written using the above notation as $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \mathbf{\Sigma}\times_1\mathbf{U}\times_2\mathbf{V}$.

III. HOSVD DECOMPOSITION

The Tucker decomposition, and in particular the HOSVD decomposition [7], decomposes a tensor into a set of matrices and one tensor.

Definition 1 (Tucker decomposition): Given a tensor $\mathbf{X} \in \mathbb{R}^{n_1\times n_2\times\cdots\times n_d}$ the decomposition

$$\mathbf{X} = \mathbf{S}\times_1\mathbf{U}_1\times_2\mathbf{U}_2\times\cdots\times_d\mathbf{U}_d,$$

or elementwise

$$\mathbf{X}(i_1,i_2,\ldots,i_d) = \sum_{j_1=1}^{n_1}\sum_{j_2=1}^{n_2}\cdots\sum_{j_d=1}^{n_d}\mathbf{S}(j_1,\ldots,j_d)\cdot \\ \cdot\mathbf{U}_1(i_1,j_1)\,\mathbf{U}_2(i_2,j_2)\cdots\mathbf{U}_d(i_d,j_d)$$

is called Tucker decomposition. The tensor $\mathbf{S} \in \mathbb{R}^{r_1\times r_2\times\cdots\times r_d}$ is called the core tensor and $\mathbf{U}_i \in \mathbb{R}^{n_i\times n_i}$, for all $i \in [d]$, are matrices.

Remark 2: Given invertible matrices $\mathbf{U}_i \in \mathbb{R}^{n_i\times n_i}$, the Tucker decomposition of a tensor $\mathbf{X}$ always exists since

$$\mathbf{S} = \mathbf{X}\times_1\mathbf{U}_1^{-1}\times_2\mathbf{U}_2^{-1}\times\cdots\times\mathbf{U}_d^{-1}$$

defines the core tensor.

Definition 2 (HOSVD decomposition): The HOSVD is a special case of the Tucker decomposition where

- the $\mathbf{U}_k$ are unitary $n_k\times n_k$-matrices, for all $k \in [d]$,
- any two subtensors of the core tensor $\mathbf{S}$ are orthogonal, i.e., $\langle S_{i_k=p}, S_{i_k=q}\rangle = 0$, for all $k \in [d]$ and for all $p \neq q$,
- the subtensors of the core tensor $\mathbf{S}$ are ordered according to their Frobenius norm, i.e., $\|S_{i_k=1}\|_F \geq \|S_{i_k=2}\|_F \geq \ldots \geq \|S_{i_k=n_k}\|_F \geq 0$, for all $k \in [d]$.

Definition 3 (HOSVD-rank): Let $\mathbf{X} \in \mathbb{R}^{n_1\times n_2\times\cdots\times n_d}$. The k-rank of $\mathbf{X}$, denoted by $R_k = \operatorname{rank}_k(\mathbf{X})$, is the rank of the k-th unfolding, i.e.,

$$\operatorname{rank}_k(\mathbf{X}) = \operatorname{rank}\left(\mathbf{X}_{(k)}\right).$$

The HOSVD-rank of a tensor $\mathbf{X}$ is the vector $\mathbf{r}_{\text{HOSVD}}(\mathbf{X}) = (R_1,R_2,\ldots,R_d)$.

Remark 3 ([7]): Let the HOSVD of a tensor $\mathbf{X} \in \mathbb{R}^{n_1\times n_2\times\cdots\times n_d}$ be given as in Definition 1 and let r_k be equal to the highest index for which $\|S_{i_k=r_k}\|_F > 0$. Then

$$r_k = \operatorname{rank}_k(\mathbf{X}) = R_k.$$

Remark 4: Let $\mathbf{X} \in \mathbb{R}^{n_1\times n_2\times\cdots\times n_d}$ be a tensor of HOSVD-rank $(r_1,r_2,\ldots,r_d)$ and let $\mathbf{X} = \mathbf{S}\times_1\mathbf{U}_1\times_2\mathbf{U}_2\times\cdots\times_d\mathbf{U}_d$ be its HOSVD decomposition. Then $\mathbf{X}$ can be written as

$$\mathbf{X} = \overline{\mathbf{S}}\times_1\overline{\mathbf{U}}_1\times_2\overline{\mathbf{U}}_2\times\cdots\times_d\overline{\mathbf{U}}_d,$$

where $\overline{\mathbf{S}} = \mathbf{S}([r_1],[r_2],\ldots,[r_d]) \in \mathbb{R}^{r_1\times r_2\times\cdots\times r_d}$, $\overline{\mathbf{U}}_k = \mathbf{U}_k(:,[r_k]) \in \mathbb{R}^{n_k\times r_k}$, for all $k \in [d]$. From now on, we will assume that the HOSVD decomposition of an arbitrary tensor is of this form.

IV. TENSOR ITERATIVE HARD THRESHOLDING ALGORITHM

In this section we present the tensor iterative hard thresholding algorithm (TIHT) and the corresponding numerical results.

In the TIHT algorithm, $\mathcal{H}_{\mathbf{r}}(\mathbf{X})$ denotes the rank-$\mathbf{r}$ approximation of the tensor $\mathbf{X}$ obtained by restricting the components of its HOSVD decomposition. To be more precise, if $\mathbf{X} = \mathbf{S}\times_1\mathbf{U}_1\times\cdots\times_d\mathbf{U}_d$ is its HOSVD decomposition, then $\mathcal{H}_{\mathbf{r}}(\mathbf{X}) = \overline{\mathbf{S}}\times_1\overline{\mathbf{U}}_1\times\cdots\times_d\overline{\mathbf{U}}_d$ where $\overline{\mathbf{S}} = \mathbf{S}([r_1],\ldots,[r_d]) \in \mathbb{R}^{r_1\times r_2\times\cdots\times r_d}$ and $\overline{\mathbf{U}}_k = \mathbf{U}_k(:,[r_k]) \in \mathbb{R}^{n_k\times r_k}$ for all $k \in [d]$.

Remark 5: In the case of sparse vector recovery and of low-rank matrix recovery, the operator $\mathcal{H}_r$ returns the best r-sparse approximation [1] and best rank-r approximation [5], respectively. This fact is often used in the analysis of the algorithm. However, the rank-$\mathbf{r}$ approximation $\mathcal{H}_{\mathbf{r}}(\mathbf{X})$ of an arbitrary d-th order tensor $\mathbf{X}$ is not necessarily its best rank-$\mathbf{r}$ approximation $\mathbf{X}_{\text{BEST}}$ [4]. To be more precise,

$$\|\mathbf{X}-\mathcal{H}_{\mathbf{r}}(\mathbf{X})\|_F \leq \sqrt{d}\,\|\mathbf{X}-\mathbf{X}_{\text{BEST}}\|_F.$$

Tensor iterative hard thresholding algorithm (TIHT algorithm)

Input: measurement ensemble $\mathcal{A}$, measurement vector $\mathbf{y} = \mathcal{A}(\mathbf{X})$, rank level $\mathbf{r}$
$\mathbf{X}^0 = \mathcal{H}_{\mathbf{r}}(\mathcal{A}^*(\mathbf{y}))$, $j = 0$.
repeat
Compute $\mu_j = \frac{\|\mathcal{A}^*(\mathbf{y}-\mathcal{A}(\mathbf{X}^j))\|_F^2}{\|\mathcal{A}(\mathcal{A}^*(\mathbf{y}-\mathcal{A}(\mathbf{X}^j)))\|_2^2}$.
Set $\mathbf{X}^{j+1} = \mathcal{H}_{\mathbf{r}}\left(\mathbf{X}^j + \mu_j \mathcal{A}^*\left(\mathbf{y} - \mathcal{A}\left(\mathbf{X}^j\right)\right)\right)$.
j=j+1
until a stopping criterion is met at $\bar{j} = j$.
Output: the $\mathbf{r}$-rank tensor $\mathbf{X}^{\#} = \mathbf{X}^j$

This fact causes significant obstacles in the theoretical anaysis of the TIHT. Nevertheless, as we present in the following, the algorithm still works quite well in practice.

We present the numerical results only for 3rd order tensors $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$. In the first two experiments we consider a cubic tensor, i.e., $n_1 = n_2 = n_3 = 10$, with equal and unequal ranks of its unfoldings, respectively. In the last case we consider a non-cubic tensor $\mathbf{X} \in \mathbb{R}^{6 \times 10 \times 15}$ with equal ranks of the unfoldings, i.e., $r_1 = r_2 = r_3 = r$.

For fixed tensor dimensions $n_1 \times n_2 \times n_3$, fixed HOSVD-rank $\mathbf{r} = (r_1, r_2, r_3)$ and a fixed number of measurements m we performed 200 simulations.

We consider an algorithm to successfully recover the sensed tensor $\mathbf{X}_0$ if it returns a tensor $\mathbf{X}^{\#}$ s.t. $\left\|\mathbf{X}_0 - \mathbf{X}^{\#}\right\|_F < 10^{-3}$.

The algorithm stops if $\left\|\mathbf{X}^j - \mathbf{X}^{j-1}\right\|_F < 10^{-4}$ in which case we say that the algorithm converged, or it stops if it reached 5000 iterations.

A linear mapping $\mathcal{A} : \mathbb{R}^{n_1 \times n_2 \times n_3} \to \mathbb{R}^m$ is defined by tensors $\mathbf{A}_k \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ via $[\mathcal{A}(\mathbf{X})](k) = \langle \mathbf{X}, \mathbf{A}_k \rangle$, for all $k \in [m]$. The entries of the tensors $\mathbf{A}_k$ are i.i.d. Gaussian $\mathcal{N}\left(0, \frac{1}{m}\right)$. We generate tensors $\mathbf{X}^0 \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ of rank $\mathbf{r} = (r_1, r_2, r_3)$ via its Tucker decomposition. If $\mathbf{X}^0 = \mathbf{S} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times_3 \mathbf{U}_3$ is its Tucker decomposition, each of the elements of the tensor $\mathbf{S}$ is taken independently from the normal distribution, $\mathcal{N}(0,1)$, and the components $\mathbf{U}_k \in \mathbb{R}^{n_k \times r_k}$ are the first r_k left singular vectors of a matrix $\mathbf{M}_k \in \mathbb{R}^{n_k \times n_k}$ whose elements are also drawn independently from the normal distribution $\mathcal{N}(0,1)$.

In Figure 1 and Figure 2 we present the recovery results for low-rank tensors of size $10 \times 10 \times 10$. The horizontal axis represents the number of measurements taken with respect to the number of degrees of freedom of an arbitrary tensor of this size. To be more precise, for a tensor of size $n_1 \times n_2 \times n_3$, the number $\bar{n}$ on the horizontal axis represents $m = \left\lceil n_1 n_2 n_2 \frac{\bar{n}}{100} \right\rceil$ measurements. The vertical axis represents the percentage of the successful recovery. The numerical results for tensors of rank $\mathbf{r} = (1,1,1)$, $\mathbf{r} = (2,2,2)$, $\mathbf{r} = (5,5,5)$ and $\mathbf{r} = (7,7,7)$ are presented in Figure 1. Notice that only for the rank $\mathbf{r} = (7,7,7)$, 33% of measurements were not enough for a full recovery. In this case 54% of the measurements and on average 1107 iterations were needed. For tensors of rank $\mathbf{r} = (1,1,1)$ already with 9% of measurements we obtain a full recovery in 321 iterations on average. The algorithm ended on average in 185, 337 and 547 iterations for 20%, 21% and 33% of measurements for ranks $\mathbf{r} = (2,2,2)$, $\mathbf{r} = (3,3,3)$ and $\mathbf{r} = (5,5,5)$, respectively.

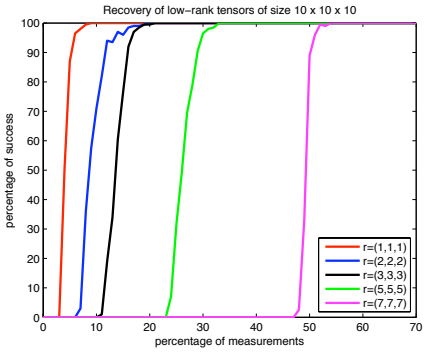


Fig. 1. Recovery of low rank 10 x 10 x 10 tensors of the same rank

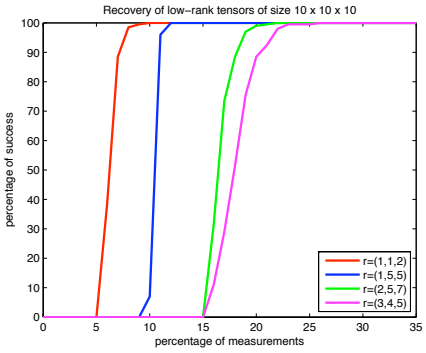


Fig. 2. Recovery of low rank $10 \times 10 \times 10$ tensors of a different rank

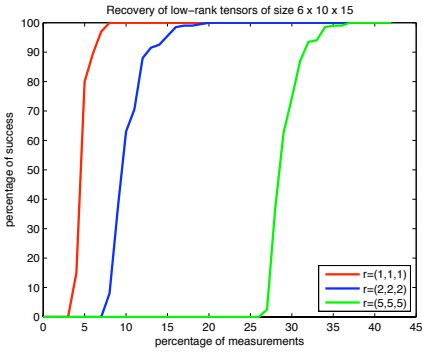


Fig. 3. Recovery of low rank $6 \times 10 \times 15$ tensors of a different rank

In Figure 2 we present the results for tensors of rank $\mathbf{r} = (1,2,2)$, $\mathbf{r} = (1,5,5)$, $\mathbf{r} = (2,5,7)$ and $\mathbf{r} = (3,4,5)$. Only 26% of measurements were enough for a full recovery. For 10%, 12%, 22% and 26% of measurements, the algorithm converged on average in 588, 1912, 696, 384 iterations, for the various ranks respectively.

We obtained similar results for recovery of low-rank tensors of size $6 \times 10 \times 15$ and ranks $\mathbf{r} = (1,1,1)$, $\mathbf{r} = (2,2,2)$ and $\mathbf{r} = (5,5,5)$ - see Figure 3. We managed to get a full recovery from 8% of measurements for the rank $\mathbf{r} = (1,1,1)$, and 20% and 37% of measurements for the remaining ranks. The algorithm ended on average in 511, 214 and 501 iterations, for

the various ranks and number of measurements, respectively.

V. HOSVD TENSOR RIP

The analysis of the IHT algorithm for recovery of sparse vectors [1] and low-rank matrices [5] is based on the corresponding notion of restricted isometry property (RIP). Therefore, we start by introducing an analog for tensors, which we call the tensor restricted isometry property (HOSVD-TRIP).

Definition 4 (HOSVD-TRIP): Let $\mathcal{A}\colon \mathbb{R}^{n_1\times n_2\times\cdots\times n_d} \to \mathbb{R}^m$ be a measurement ensemble. Then for each d-tuple of the integers $\mathbf{r}$, $\mathbf{r} = (r_1, r_2, \ldots, r_d)$, where $r_i \in [n_i]$, for all $i \in [d]$, the tensor restricted isometry constant $\delta_{\mathbf{r}}$ of $\mathcal{A}$ is the smallest quantity such that

$$(1-\delta_{\mathbf{r}})\left\|\mathbf{X}\right\|_F^2 \leq \left\|\mathcal{A}\left(\mathbf{X}\right)\right\|_2^2 \leq (1+\delta_{\mathbf{r}})\left\|\mathbf{X}\right\|_F^2 \qquad (4)$$

holds for all tensors of HOSVD-rank at most $\mathbf{r}$.

We say that $\mathcal{A}$ satisfies the HOSVD-TRIP at rank $\mathbf{r}$ if $\delta_{\mathbf{r}}$ is bounded by a sufficiently small constant between 0 and 1.

A random variable X is called L-subgaussian if there exists a constant $L > 0$ s.t. $\mathbb{E}\left[\exp\left(tX\right)\right] \leq \exp\left(L^2t^2/2\right)$ holds for all $t \in \mathbb{R}$. We call $\mathcal{A} : \mathbb{R}^{n_1\times n_2\times\cdots\times n_d} \to \mathbb{R}^m$ an L-subgaussian measurement ensemble if all elements of $\mathcal{A}$, interpreted as a tensor in $\mathbb{R}^{n_1\times n_2\times\cdots\times n_d\times m}$, are independent mean-zero, variance one, L-subgaussian variables. Gaussian and Bernoulli random measurement ensembles where the entries are standard normal distributed random variables and Rademacher ± 1 variables, respectedly, are special cases of 1-subgaussian measurement ensembles.

Theorem 1: Let $\delta, \varepsilon \in (0,1)$. A random draw of an L-subgaussian measurement ensemble $\mathcal{A} : \mathbb{R}^{n_1\times n_2\times\cdots\times n_d} \to \mathbb{R}^m$ satisfies $\delta_{\mathbf{r}} \leq \delta$ with probability at least $1-\varepsilon$ provided

$$m \geq C\delta^{-2}\max\left\{\left(r^d + dnr\right)\log\left(d\right), \log\left(\varepsilon^{-1}\right)\right\},$$

where $n = \max\left\{n_i : i \in [d]\right\}$ and $r = \max\left\{r_i : i \in [d]\right\}$. The constant $C > 0$ depends only on subgaussian parameter L.

The proof of Theorem 1 uses ε-nets.

Definition 5: A set $\mathcal{N}_\varepsilon \subset X$ is called an ε-net of X with respect to the norm $\left\|\cdot\right\|$ if for each $v \in X$, there exists $v_0 \in \mathcal{N}_\varepsilon$ with $\left\|v_0 - v\right\| \leq \varepsilon$. The minimal cardinality of an ε-net of X with respect to the norm $\left\|\cdot\right\|$, if finite, is denoted $\mathcal{N}\left(X, \left\|\cdot\right\|, \varepsilon\right)$ and is called the covering number of X (at scale ε).

Lemma 1 (Covering number of low-HOSVD-rank tensors): Let

$$S_{\mathbf{r}} = \left\{\mathbf{X} \in \mathbb{R}^{n_1\times n_2\times\cdots\times n_d} : \mathrm{rank}_{\mathrm{HOSVD}}\left(\mathbf{X}\right) \leq \mathbf{r}, \left\|\mathbf{X}\right\|_F = 1\right\}.$$

Then there exists an ε-net $\mathcal{N}_\varepsilon$ of $S_{\mathbf{r}}$ with respect to the Frobenius norm obeying

$$\mathcal{N}\left(S_{\mathbf{r}}, \left\|\cdot\right\|_F, \varepsilon\right) \leq \left(3\left(d+1\right)/\varepsilon\right)^{r_1r_2\cdots r_d + \sum_{i=1}^d n_ir_i}. \qquad (5)$$

The proof of the above lemma follows a similar strategy as in [2] and will be presented in a forthcoming journal paper.

Sketch of the proof of the Theorem 1: We use a tool developed in [6]. We write

$$\mathcal{A}\left(\mathbf{X}\right) = \mathbf{V}_{\mathbf{X}}\boldsymbol{\xi},$$

where $\boldsymbol{\xi}$ is an L-subgaussian random vector of length $n_1n_2\cdots n_dm$ and $\mathbf{V}_{\mathbf{X}}$ is the $m \times n_1n_2\cdots n_dm$ block-diagonal matrix

$$\mathbf{V}_{\mathbf{X}} = \frac{1}{\sqrt{m}}\begin{bmatrix} \mathbf{x}^T & 0 & \cdots & 0 \\ 0 & \mathbf{x}^T & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \mathbf{x}^T \end{bmatrix},$$

where $\mathbf{x}$ is the vectorized version of the tensor $\mathbf{X}$. With this notation the restricted isometry constant is given by

$$\delta_{\mathbf{r}} = \sup_{\mathbf{X}\in S_{\mathbf{r}}}\left|\left\|\mathbf{V}_{\mathbf{X}}\boldsymbol{\xi}\right\|_2^2 - \mathbb{E}\left\|\mathbf{V}_{\mathbf{X}}\boldsymbol{\xi}\right\|_2^2\right|.$$

Theorem 3.1 in [6] provides a general probabilistic bound for expressions in the form of the right hand side above in terms of the diameters $d_F(\mathcal{B})$ and $d_{2\to 2}(\mathcal{B})$ of the set $\mathcal{B} := \{\mathbf{V}_{\mathbf{X}} : \mathbf{X} \in S_{\mathbf{r}}\}$ with respect to the Frobenius norm and the operator norm, as well as in terms of Talagrand's functional $\gamma_2(\mathcal{B}, \left\|\cdot\right\|_{2\to 2})$. It is straightforward to see that $d_F(\mathcal{B}) = 1$ and $d_{2\to 2}(\mathcal{B}) = \frac{1}{\sqrt{m}}$. The bound of the γ_2-functional via a Dudley type integral [6] yields

$$\gamma_2\left(\mathcal{B}, \left\|\cdot\right\|_{2\to 2}\right) \leq C\frac{1}{\sqrt{m}}\int_0^1 \sqrt{\log\left(\mathcal{N}\left(S_{\mathbf{r}}, \left\|\cdot\right\|_2, u\right)\right)}du.$$

Using (5) for $\mathcal{N}\left(S_{\mathbf{r}}, \left\|\cdot\right\|_F, u\right)$ we reach

$$\gamma_2\left(\mathcal{B}, \left\|\cdot\right\|_{2\to 2}\right) \leq \tilde{C}\sqrt{\frac{\left(r_1r_2\cdots r_d + \sum_{i=1}^d n_ir_i\right)\log\left(d\right)}{m}}.$$

The claim follows then from Theorem 3.1 in [6]. ∎

REFERENCES

[1] T. Blumensath, M. E. Davis, *Iterative hard thresholding for compressed sensing*. Appl. Comput. Harmon. Anal., 27(3):265–274, 2009.

[2] E. J. Candès , Y. Plan, *Tight oracle bounds for low-rank matrix recovery from a minimal number of random measurements*. IEEE Transactions on Information Theory, 57(4):2342–2359, 2009.

[3] S. Gandy, B. Recht, I. Yamada, *Tensor completion and low-n-rank tensor recovery via convex optimization*. Inverse Problems, 27(2):19pp, 2011.

[4] L. Grasedyck, *Hierarchical Singular Value Decomposition of Tensors*. SIAM J. Matrix Anal. Appl., 31(4):2029–2054, 2010.

[5] P. Jain, R. Meka, I. S. Dhillon, *Guaranteed Rank Minimization via Singular Value Projection*. Neural Information Processing Systems, pp. 937–945, 2010.

[6] F. Krahmer, S. Mendelson, H. Rauhut *Suprema of Chaos Processes and the Restricted Isometry Property*. Comm. Pure Appl. Math., to appear

[7] L. de Lathauwer, B. de Moor, J. Vandewalle, *A multilinear singular value decomposition*. SIAM J. Matrix Anal. Appl., 21(4):1253–1278, 2000.

[8] L.-H. Lim, C. J. Hillar, *Most Tensor Problems are NP-Hard*. http://arxiv.org/pdf/0911.1393v3.pdf, 2012.

[9] J. Liu, P. Musiaski, P. Wonka, J. Ye, *Tensor Completion for Estimating Missing Values in Visual Data*. IEEE Transactions on Pattern Analyis and Machine Intelligence, 35(1):208–220, 2013.

[10] D. Muti, S. Bourennane, *Multidimensional filtering based on a tensor approach*. Signal Process., 85(12):2338–2353, 2005.

[11] B. Recht, M. Fazel, P. A. Parillo, *Guaranteed minimum-rank solution of linear matrix equations via nuclear norm minimization*. SIAM Rev., 52(3):471–501, 2010.

[12] B. Savas, L. Eldén, *Handwritten digit classification using higher order singular value decomposition*. Pattern Recog., 40(3):993–1008, 2003.

[13] M. A. O. Vasilescu, D. Terzopoulos, *Multilinear analysis of image ensembles: TensorFaces*. Lecture Notes in Computer Science, 2350:447–460, Springer-Verlag, 2002.

(Non-)Density Properties of Discrete Gabor Multipliers

Dominik Bayer
Acoustics Research Institute
Austrian Academy of Sciences
Wien, Austria
Email: bayerd@kfs.oeaw.ac.at

Peter Balazs
Acoustics Research Institute
Austrian Academy of Sciences
Wien, Austria
Email: peter.balazs@oeaw.ac.at

***Abstract*—This paper is concerned with the possibility of approximating arbitrary operators by multipliers for Gabor frames or more general Bessel sequences. It addresses the question of whether sets of multipliers (whose symbols come from prescribed function classes such as ℓ^2) constitute dense subsets of various spaces of operators (such as Hilbert-Schmidt class). We prove a number of negative results that show that in the discrete setting subspaces of multipliers are usually not dense and thus too small to guarantee arbitrary good approximation. This is in contrast to the continuous case.**

I. Preliminaries

All Hilbert spaces are assumed to be separable and infinite-dimensional.

A. Bessel sequences

Let H be a Hilbert space with inner product, linear in the first argument, denoted by $\langle \cdot, \cdot \rangle$. A sequence (f_n), $n \in \mathbb{N}$, of elements of H is called a *Bessel sequence* if there exists a constant $B > 0$ such that $\sum_{n\in\mathbb{N}} |\langle h, f_n\rangle|^2 \leq B\|h\|^2$ for all $h \in H$. Any such number B is called a *Bessel bound* of the Bessel sequence, the smallest such constant the *optimal Bessel bound*. If a Bessel sequence satisfies additionally the analogous inequality from below, i.e. there exists a constant $A > 0$ such that $\sum_{n\in\mathbb{N}} |\langle h, f_n\rangle|^2 \geq A\|h\|^2$ for all $h \in H$, then the sequence is called a *frame* for H. Prominent examples of Bessel sequences are orthonormal systems, which are Bessel sequences with Bessel bound 1. For a Bessel sequence (f_n), the *analysis operator* $C : H \to \ell^2$, $h \mapsto Ch := (\langle h, f_n\rangle)_{n\in\mathbb{N}}$, and the *synthesis operator* $D : \ell^2 \to H$, $c = (c_n) \mapsto Dc := \sum_{n\in\mathbb{N}} c_n f_n$ (the series converges in the norm topology of H), are well-defined and adjoint to each other: $C = D^*$.

A useful characterization of Bessel sequences is the following (cf. [4]):

Lemma I.1. *Let (f_n) be a sequence in H and (e_n) be an arbitrary orthonormal basis. Then (f_n) is a Bessel sequence if and only if there exists a bounded operator $T \in B(H)$ with $f_n = Te_n$ for all $n \in \mathbb{N}$. The optimal Bessel bound B is given by $B = \|T\|^2_{B(H)}$.*

We will often use the following basic fact about Bessel sequences (see e.g. [4]):

Lemma I.2. *Let (f_n) be a Bessel sequence with Bessel bound B. Then, for all $n \in \mathbb{N}$,*

$$\|f_n\| \leq \sqrt{B}.$$

B. Time-frequency analysis

In the Hilbert space $L^2(\mathbb{R})$, define the *translation operator* $T_x f(t) = f(t - x)$ and the *modulation operator* $M_\omega f(t) = e^{-2\pi i\omega t} f(t)$ (for $f \in L^2$ and $x, \omega \in \mathbb{R}$). These are unitary operators on L^2. They combine to form the *time-frequency shift* $\pi(x, \omega) = M_\omega T_x$. The *short-time Fourier transform (STFT)* of f with window g is defined as the bilinear time-frequency distribution

$$V_g f(x, \omega) = \int_{\mathbb{R}} f(t)\overline{g(t - x)}e^{-2\pi i\omega t}\, dt = \langle f, \pi(x, \omega)g\rangle.$$

If (x_n, ω_n), $n \in \mathbb{N}$, is a discrete subset of $\mathbb{R}^2$ and $h \in L^2$, then the family of functions $(\pi(x_n, \omega_n)h)$ is called a *Gabor system*. If a Gabor system constitutes a Bessel sequence or a frame for L^2, we speak of a *Bessel Gabor system* or *Gabor frame*, respectively. Another important time-frequency distribution is the *(cross) Wigner distribution* of f and g:

$$W(f, g)(x, \omega) = \int_{\mathbb{R}} f(x + \frac{t}{2})\overline{g(x - \frac{t}{2})}e^{-2\pi i\omega t}\, dt.$$

It is related to the STFT via the formula $W(f, g)(x, \omega) = 2e^{4\pi i\omega x} V_{\tilde{g}} f(2x, 2\omega)$ with $\tilde{g}(t) = g(-t)$. Both STFT and Wigner distribution are in $L^2(\mathbb{R}^2)$ if f and g are in $L^2(\mathbb{R})$. Both can be defined for larger classes of functions or even distributions for f and g. The Wigner distribution is associated to the *Weyl calculus*: every continuous operator $T : \mathcal{S} \to \mathcal{S}'$ from *Schwartz class* $\mathcal{S}$ to the *tempered distributions* $\mathcal{S}'$ can be described in the form $\langle Tf, g\rangle = \langle \sigma, W(g, f)\rangle$ for $f, g \in \mathcal{S}$, with a suitable unique (distributional) *Weyl symbol* $\sigma \in \mathcal{S}'(\mathbb{R}^2)$. If T is a Hilbert-Schmidt operator, then one has $\sigma \in L^2(\mathbb{R}^2)$. For all of these facts and many more we refer to [6].

C. Compact and Schatten class operators

A bounded operator $T : H \to H$ is *compact* if the image of any bounded sequence under T contains a convergent subsequence. A compact operator always has a *spectral*

representation $T(\cdot) = \sum_k s_k(T)\,\langle \cdot, \phi_k\rangle\,\psi_k$ with suitably chosen orthonormal systems (ϕ_k), (ψ_k) and a unique sequence $(s_k(T))$ with $s_1(T) \geq s_2(T) \geq \ldots \geq 0$, $k \in \mathbb{N}$. The sequence $(s_k(T))$ is the sequence of *singular values* of T. The operator belongs to *Schatten p-class* $\mathcal{S}^p(H)$, $1 \leq p < \infty$, if $\sum_k |s_k(T)|^p < \infty$. These are Banach spaces with norm $\|T\|_{\mathcal{S}^p} = \|(s_k(T))\|_p = \left(\sum_k |s_k(T)|^p\right)^{1/p}$. $\mathcal{S}^2(H)$ is also called *Hilbert-Schmidt class*, $\mathcal{S}^1(H)$ *trace class*. We use the notation $\mathcal{S}^\infty(H)$ to denote the set $B(H)$ of all bounded operators on H, and $\mathcal{S}^0(H) = K(H)$ to denote the set of all compact operators on H. For more information, refer to e.g. [5] or [7].

II. Bessel Multipliers

Definition II.1. *Let (f_n) and (g_n) be Bessel sequences in H and $m = (m_n) \in \ell^\infty$. The* Bessel multiplier *with* symbol m *(associated to the sequences (f_n) and (g_n)) is defined as the linear operator on H given by*

$$\mathcal{A}^{(f_n),(g_n)}(m)(h) := \sum_n m_n \langle h, f_n\rangle g_n, \qquad h \in H.$$

In order to simplify notation, we will usually suppress the dependence on the Bessel sequences (f_n) and (g_n) and simply write $\mathcal{A}(m)$ instead of $\mathcal{A}^{(f_n),(g_n)}(m)$.

We cite without proof several results from [1].

Lemma II.2. *Let (f_n) and (g_n) be Bessel sequences with Bessel bounds B_F and B_G, respectively. If $(m_n) \in \ell^\infty$, then $\mathcal{A}(m)$ is a well-defined bounded operator on H with norm $\|\mathcal{A}(m)\|_{B(H)} \leq \sqrt{B_F B_G}\,\|m\|_\infty$.*

Lemma II.3. *Let (f_n) and (g_n) be Bessel sequences with Bessel bounds B_F and B_G, respectively. If $(m_n) \in \ell^1$, then $\mathcal{A}(m)$ is a trace class operator on H with norm $\|\mathcal{A}(m)\|_{\mathcal{S}^1} \leq \sqrt{B_F B_G}\,\|m\|_1$.*

Lemma II.4. *Let (f_n) and (g_n) be Bessel sequences with Bessel bounds B_F and B_G, respectively. If $\lim_{n\to\infty} m_n = 0$, i.e. $m \in c_0(\mathbb{N})$, then $\mathcal{A}(m)$ is a compact operator.*

From Lemma II.2 and Lemma II.3, the following is easily proved by interpolation:

Lemma II.5. *Let (f_n) and (g_n) be Bessel sequences with Bessel bounds B_F and B_G, repsectively. If $m \in \ell^p(\mathbb{N})$, $1 \leq p < \infty$, then $\mathcal{A}(m)$ is a Schatten p-class operator, and $\|\mathcal{A}(m)\|_{\mathcal{S}^p} \leq \sqrt{B_F B_G}\,\|m\|_p$.*

Table I summarizes these results.

Symbol	Bessel Multiplier
$\ell^\infty(\mathbb{N})$	$B(H) = \mathcal{S}^\infty(H)$
$c_0(\mathbb{N}) = \ell^0(\mathbb{N})$	$K(H) = \mathcal{S}^0(H)$, compact operator
$\ell^p(\mathbb{N})$, $1 \leq p < \infty$	$\mathcal{S}^p(H)$, Schatten class operator

TABLE I
Bessel multipliers with different symbols

See also the paper [3], which contains somewhat related results for Gabor multipliers.

III. Berezin Transform

Definition III.1. *Let (f_n) and (g_n) be Bessel sequences in H and $T \in B(H)$. The* Berezin transform *of T (associated to the sequences (f_n) and (g_n)) is defined as the function on $\mathbb{N}$ given by*

$$\mathcal{B}^{(f_n),(g_n)}(T)(n) := \langle Tf_n, g_n\rangle, \qquad n \in \mathbb{N}.$$

In order to simplify notation we will usually suppress the dependence on the Bessel sequences (f_n) and (g_n) and simply write $\mathcal{B}(T)$ instead of $\mathcal{B}^{(f_n),(g_n)}(T)$.

Lemma III.2. *Let (f_n) and (g_n) be Bessel sequences with Bessel bounds B_F and B_G, respectively. Then the Berezin transform $\mathcal{B}(T)$ is bounded, hence in $\ell^\infty(\mathbb{N})$, and*

$$\|\mathcal{B}(T)\|_\infty \leq \sqrt{B_F B_G}\,\|T\|_{B(H)}.$$

Proof: We have

$$|\mathcal{B}(T)(n)| \leq \|T\|_{B(H)}\|f_n\|\|g_n\| \leq \|T\|_{B(H)}\sqrt{B_F}\sqrt{B_G}$$

by Lemma I.2, for all $n \in \mathbb{N}$. ∎

For later use, we calculate the Berezin transform of a rank-one operator.

Corollary III.3. *Let $\phi, \psi \in H$ and $T : H \to H$, $h \mapsto \langle h, \phi\rangle\psi$ a rank-one operator. Then*

$$\mathcal{B}(T)(n) = \langle f_n, \phi\rangle\langle \psi, g_n\rangle.$$

We collect further mapping properties of the Berezin transform.

Lemma III.4. *Suppose $T \in B(H)$ is a compact operator and (f_n) and (g_n) are Bessel sequences. Then $\lim_{n\to\infty} |\mathcal{B}(T)(n)| = 0$, i.e. $\mathcal{B}(T) \in c_0(\mathbb{N})$.*

Proof: Since $f_n \xrightarrow{w} 0$ for $n \to \infty$, we have $\|Tf_n\| \to 0$ for $n \to \infty$. Together with Lemma I.2 this yields

$$|\langle Tf_n, g_n\rangle| \leq \|Tf_n\|\|g_n\| \to 0, \quad \text{for } n \to \infty.$$

∎

Lemma III.5. *Let (f_n) and (g_n) be Bessel sequences with Bessel bounds B_F and B_G, respectively. Let T be a Schatten class operator, with $1 \leq p < \infty$. Then the Berezin transform $\mathcal{B}(T)$ is in $\ell^p(\mathbb{N})$, and*

$$\|\mathcal{B}(T)\|_p \leq \sqrt{B_F B_G}\,\|T\|_{\mathcal{S}^p}.$$

Proof: Let (e_n) be an arbitrary orthonormal basis for H. By Lemma I.1, there are bounded operators R and S in $B(H)$ such that $f_n = Re_n$ and $g_n = Se_n$ for all n and $\|R\|_{B(H)} \leq \sqrt{B_F}$ and $\|S\|_{B(H)} \leq \sqrt{B_G}$. Hence

$$\langle Tf_n, g_n\rangle = \langle TRe_n, Se_n\rangle = \langle S^*TRe_n, e_n\rangle$$

for all n. The operator $\tilde{T} = S^*TR$ is again in $\mathcal{S}^p$, so

$$\left(\sum_n |\mathcal{B}(T)(n)|^p\right)^{\frac{1}{p}} = \left(\sum_n |\langle \tilde{T}e_n, e_n\rangle|^p\right)^{\frac{1}{p}} \leq \|\tilde{T}\|_{\mathcal{S}^p}.$$

Since

$$\|\tilde{T}\|_{\mathcal{S}^p} \le \|S^*\|_{B(H)}\|T\|_{\mathcal{S}^p}\|R\|_{B(H)} \le \|T\|_{\mathcal{S}^p}\sqrt{B_F B_G},$$

the proof is finished. ∎

Table II summarizes these results.

Operator	Berezin transform
$B(H) = \mathcal{S}^\infty(H)$	$\ell^\infty(\mathbb{N})$
$K(H) = \mathcal{S}^0(H)$, compact operator	$c_0(\mathbb{N}) = \ell^0(\mathbb{N})$
$\mathcal{S}^p(H)$, $1 \le p < \infty$, Schatten class	$\ell^p(\mathbb{N})$

TABLE II
BEREZIN TRANSFORM OF DIFFERENT OPERATORS

Suggested by the results given above, it becomes obvious that the concept of Bessel multiplier and the Berezin transform are dual to each other.

The following theorem gives the connection between the Berezin transform and Bessel multipliers.

Theorem III.6. *Let $m = (m_n) \in \ell^p(\mathbb{N})$, $1 \le p \le \infty$, and $T \in \mathcal{S}^q$, with q the conjugate exponent to p, i.e. $\frac{1}{p} + \frac{1}{q} = 1$. Then*

$$\langle \mathcal{A}(m), T\rangle_{\mathcal{S}^p,\mathcal{S}^q} = \langle m, \mathcal{B}(T)\rangle_{\ell^p,\ell^q}.$$

Proof: For the moment, let (e_k) be an arbitrary orthonormal basis of H. Then the left hand side can be written as

$$\langle \mathcal{A}(m), T\rangle = \sum_k \langle \mathcal{A}(m)(e_k), Te_k\rangle.$$

Inserting $\mathcal{A}(m) = \sum_n m_n\langle\cdot, f_n\rangle g_n$ yields

$$\langle \mathcal{A}(m), T\rangle = \sum_k\sum_n m_n\langle e_k, f_n\rangle\langle g_n, Te_k\rangle. \qquad (*)$$

The right hand side gives

$$\begin{aligned}\langle m, \mathcal{B}(T)\rangle &= \sum_n m_n\langle g_n, Tf_n\rangle \\ &= \sum_n m_n \sum_k \langle T^*g_n, e_k\rangle\langle e_k, f_n\rangle\end{aligned}$$

by Parseval's equality. Thus

$$\langle m, \mathcal{B}(T)\rangle = \sum_n\sum_k m_n\langle e_k, f_n\rangle\langle g_n, Te_k\rangle. \qquad (**)$$

Comparing $(*)$ and $(**)$, we see that the claimed equality is proved, if we can justify the change of order of summation in the double sum. In order to do so, we examine the corresponding double sum of the absolute values

$$S := \sum_n\sum_k |m_n||\langle e_k, f_n\rangle||\langle g_n, Te_k\rangle|.$$

Consider the case $p = 1$ (i.e. $m \in \ell^1$ and $T \in \mathcal{S}^\infty = B(H)$). Then

$$\begin{aligned}S &\le \sum_n |m_n|\Big(\sum_k |\langle e_k, f_n\rangle|^2\Big)^{\frac{1}{2}}\Big(\sum_k |\langle T^*g_n, e_k\rangle|^2\Big)^{\frac{1}{2}} \\ &\le \sum_n |m_n|\|f_n\|\|T^*g_n\| \\ &\le \sqrt{B_F B_G}\,\|m\|_1\|T\|_{B(H)} < \infty.\end{aligned}$$

If $1 < p \le \infty$, then $1 \le q < \infty$ and T is a compact operator in $\mathcal{S}^q$. As such, it has a spectral representation

$$T = \sum_k \lambda_k\langle\cdot, \sigma_k\rangle\tau_k$$

with orthonormal bases (σ_k) and (τ_k), and $\lambda_k \ge 0$ with $\sum_k \lambda_k^q = \|T\|_{\mathcal{S}^q}^q$. Choose the particular orthonormal basis $(e_k) = (\sigma_k)$. Then $Te_k = T\sigma_k = \lambda_k\tau_k$ for all k, and thus

$$\begin{aligned}S &= \sum_{n,k} |m_n||\lambda_k||\langle\sigma_k, f_n\rangle||\langle g_n, \tau_k\rangle| \\ &\le \Big(\sum_{n,k} |m_n|^p|\langle\sigma_k, f_n\rangle||\langle g_n, \tau_k\rangle|\Big)^{\frac{1}{p}} \times \\ &\qquad \Big(\sum_{n,k} |\lambda_k|^q|\langle\sigma_k, f_n\rangle||\langle g_n, \tau_k\rangle|\Big)^{\frac{1}{q}}.\end{aligned}$$

These two sums can be estimated, the first as

$$\begin{aligned}&\sum_{n,k} |m_n|^p|\langle\sigma_k, f_n\rangle||\langle g_n, \tau_k\rangle| \\ &\quad\le \sum_n |m_n|^p\Big(\sum_k |\langle\sigma_k, f_n\rangle|^2\Big)^{\frac{1}{2}}\Big(\sum_k |\langle g_n, \tau_k\rangle|^2\Big)^{\frac{1}{2}} \\ &\quad\le \sqrt{B_F B_G}\,\|m\|_p^p\end{aligned}$$

and the second similarly as

$$\begin{aligned}&\sum_{n,k} |\lambda_k|^q|\langle\sigma_k, f_n\rangle||\langle g_n, \tau_k\rangle| \\ &\quad\le \Big(\sum_k \lambda_k^q\Big)\sqrt{B_F}\|\sigma_k\|\sqrt{B_G}\|\tau_k\| \\ &\quad= \sqrt{B_F B_G}\,\|T\|_{\mathcal{S}^q}^q.\end{aligned}$$

So, finally, we have for $1 < p \le \infty$

$$S \le \sqrt{B_F B_G}\,\|m\|_p\|T\|_{\mathcal{S}^q} < \infty.$$

Since in every case $S < \infty$, Fubini's theorem yields the desired conclusion, the equality of $(*)$ and $(**)$. ∎

Corollary III.7. *(1) Let $\mathcal{A} : \ell^\infty \to B(H)$ and $\mathcal{B} : \mathcal{S}^1 \to \ell^1$. Then $\mathcal{A} = \mathcal{B}^*$ is the Banach space adjoint.*
(2) Let $\mathcal{A} : c_0 \to K(H)$ and $\mathcal{B} : \mathcal{S}^1 \to \ell^1$. Then $\mathcal{B} = \mathcal{A}^*$.
(3) Let $\mathcal{A} : \ell^p \to \mathcal{S}^p$, $1 \le p < \infty$, and $\mathcal{B} : \mathcal{S}^q \to \ell^q$, with $1 < q \le \infty$ the conjugate exponent. Then $\mathcal{B} = \mathcal{A}^*$.

Proof: Observe that $B(H) = (\mathcal{S}^1)^*$ and $\ell^\infty = (\ell^1)^*$ in case (1), $\mathcal{S}^1 = (K(H))^*$ and $c_0 \subseteq \ell^\infty$ with $\ell^1 = (c_0)^*$ in case (2), and $\mathcal{S}^q = (\mathcal{S}^p)^*$ and $\ell^q = (\ell^p)^*$ in case (3). The statements then follow immediately from Theorem III.6. ∎

IV. (NON-)DENSITY RESULTS

In this section we investigate whether a given operator on H can be approximated by a Gabor multiplier with respect to various norms. In particular, we would like to understand when the set of Gabor multipliers (associated to a fixed pair of Gabor systems) is dense in $B(H)$ or in $\mathcal{S}^p(H)$ (if ever).

In order to examine such density properties, we employ some well known results from functional analysis. Precisely,

we use the following facts ([5]):
Let X, Y be Banach spaces, $T : X \to Y$ be a bounded operator and let $T^* : Y^* \to X^*$ be the (Banach space) adjoint operator.

- T^* is one-to-one on Y^*, if and only if the range of T is dense in Y with respect to the norm topology on Y.
- T is one-to-one on X, if and only if the range of T^* is dense in X^* with respect to the weak* topology on X^*.

To understand when the mapping $a \to \mathcal{A}(a)$ has dense range, it suffices, in view of Theorem III.6 and its corollary, to check when the Berezin transform $\mathcal{B}$ is one-to-one.

Lemma IV.1. *Let $a, b > 0$ and assume that $(f_{n,m}) = (\pi(an, bm)f)$ and $(g_{n,m}) = (\pi(an, bm)g)$ are Gabor systems, $f, g \in L^2(\mathbb{R})$. Let $(z, \nu) \in \mathbb{R}^2$ and $T = \pi(z, \nu)$ be the corresponding time-frequency shift. Then*

$$\mathcal{B}(T)(n, m) = e^{2\pi i(an\nu - bmz)}\overline{V(g, f)(z, \nu)}$$

for all $n, m \in \mathbb{Z}$.

Corollary IV.2. *Let $(f_{n,m}) = (\pi(an, bm)f)$ and $(g_{n,m}) = (\pi(an, bm)g)$ be Bessel Gabor systems. If there exists a point $(z, \nu) \in \mathbb{R}^2$ such that $V(g, f)(z, \nu) = 0$, then the Berezin transform $\mathcal{B} : B(L^2) \to \ell^\infty$ is not one-to-one.*

Proof: We have $T = \pi(z, \nu) \neq 0$ in $B(L^2)$, but $\mathcal{B}(T) = 0$ by the preceding lemma. ∎

For the particular case of Hilbert-Schmidt operators, we have the following negative result:

Theorem IV.3. *Let $(f_{n,m})$ and $(g_{n,m})$ be Bessel Gabor systems. Then the range of $\mathcal{A} : \ell^2 \to \mathcal{S}^2$ is not a norm-dense subspace of Hilbert-Schmidt class. There are thus Hilbert-Schmidt operators on $L^2(\mathbb{R})$ that cannot be approximated in Hilbert-Schmidt norm by Gabor multipliers (with a given fixed pair of Gabor systems).*

Proof: In view of Corollary III.7, it suffices to show that $\mathcal{B} : \mathcal{S}^2 \to \ell^2$ is not one-to-one. Let $T \in \mathcal{S}^2$. Now note that there is a bijective correspondence between Hilbert-Schmidt operators and Weyl symbols in $L^2(\mathbb{R}^2)$. Thus there exists a unique Weyl symbol $\sigma \in L^2(\mathbb{R}^2)$ such that

$$\langle T\phi, \psi\rangle = \langle \sigma, W(\psi, \phi)\rangle$$

for all $\phi, \psi \in L^2(\mathbb{R})$. Thus

$$\begin{aligned}\mathcal{B}(T)(n, m) &= \langle \sigma, W(g_{n,m}, f_{n,m})\rangle \\ &= \langle \sigma, W(\pi(an, bm)g, \pi(an, bm)f)\rangle \\ &= \langle \sigma, T_{(an,bm)}W(g, f)\rangle.\end{aligned}$$

Observe that $W(g, f) \in L^2(\mathbb{R}^2)$. As is well-known, a discrete countable family of translates of a function $F \in L^2(\mathbb{R}^2)$ is never complete, thus

$$U := \overline{\text{span}\{T_{(an,bm)}W(g, f) \mid n, m \in \mathbb{Z}\}}$$

is a proper closed subspace of $L^2(\mathbb{R}^2)$. Choose $0 \neq \sigma \in U^\perp$. Then the corresponding Hilbert-Schmidt operator T satisfies $\mathcal{B}(T)(n, m) = \langle \sigma, W(g_{n,m}, f_{n,m})\rangle = 0$ for all $n, m \in \mathbb{Z}$, thus $\mathcal{B}(T) = 0$, but $T \neq 0$. Hence $\mathcal{B} : \mathcal{S}^2 \to \ell^2$ is not one-to-one. ∎

We can extend this result to the cases $1 \leq p < 2$.

Theorem IV.4. *Let $(f_{n,m})$ and $(g_{n,m})$ be Bessel Gabor systems and $1 \leq p < 2$. Then the range of $\mathcal{A} : \ell^p \to \mathcal{S}^p$ is not a norm-dense subspace of the Schatten class $\mathcal{S}^p$.*

Proof: Let $2 < q \leq \infty$ be the conjugate exponent to p. Observe that $\mathcal{S}^2 \subseteq \mathcal{S}^q \subseteq \mathcal{S}^\infty$. By Theorem IV.3, the Berezin transform $\mathcal{B} : \mathcal{S}^2 \to \ell^2$ is not one-to-one, hence, a fortiori, the Berezin transform $\mathcal{B} : \mathcal{S}^q \to \ell^q$ is not one-to-one, either. By Corollary III.7, this is equivalent to the range of $\mathcal{A} : \ell^p \to \mathcal{S}^p$ not being norm-dense. ∎

For the cases $2 < p < \infty$, we conjecture analogous results.

For the case $p = \infty$, we have the following result (whose proof we omit for lack of space):

Theorem IV.5. *Let $(f_{n,m})$ and $(g_{n,m})$ be Bessel Gabor systems. Then there exists an operator $R \in B(L^2)$ and a constant $\delta > 0$ such that*

$$\|R - \mathcal{A}(m)\|_{B(L^2)} \geq \delta$$

for all $m \in \ell^\infty$. In particular, the range of $\mathcal{A} : \ell^\infty \to B(L^2)$ is not a norm-dense subspace of $B(L^2)$.

One can take for R the Fourier transform, fractional Fourier transforms or any other operator that incorporates time-frequency shifts of arbitrarily large size.

V. Conclusion

Our results show that subsets of Gabor multipliers with symbols in ℓ^p-spaces are not dense in the respective Schatten classes, but span proper subspaces. There exist thus operators in these Schatten classes that cannot be approximated arbitrarily well by multipliers in the respective Schatten norm. This is in sharp contrast to the case of continuous (STFT) multipliers, as shown in [2]. For approximation of bounded operators in operator norm, however, the negative result shown in this paper also holds analogously in the continuous case.

VI. Acknowledgement

The first author would like to thank K. Gröchenig for his substantial help (in particular with [2]) and continual encouragement.

References

[1] P. Balazs. Basic definition and properties of Bessel multipliers. *J. Math. Anal. Appl.*, 325(1):571–585, 2007.
[2] D. Bayer. *Bilinear Time-Frequency Distributions and Pseudodifferential Operators*. PhD thesis, University of Vienna, 2010.
[3] J. Benedetto and G. Pfander. Frame expansions for Gabor multipliers. *Appl. Comp. Harm. Anal.*, 20(1):26–40, 2006.
[4] O. Christensen. *An Introduction to Frames and Riesz Bases*. Applied and Numerical Harmonic Analysis. Birkhäuser, 2003.
[5] J.B. Conway. *A Course in Functional Analysis*. Springer, New York, 2nd edition, 1990.
[6] K. Gröchenig. *Foundations of Time-Frequency Analysis*. Appl. Numer. Harmon. Anal. Birkhäuser Boston, Boston, MA, 2001.
[7] K. Zhu. *Operator Theory in Function Spaces*. Mathematical Surveys and Monographs. American Mathematical Society, 2nd edition, 2007.

Estimation of frequency modulations on wideband signals; applications to audio signal analysis

Harold Omer and Bruno Torrésani
Aix-Marseille Université, CNRS, Centrale Marseille, LATP, UMR 7353, 13453 Marseille, France
Email: harold.omer@latp.univ-mrs.fr bruno.torresani@latp.univ-mrs.fr

***Abstract*—The problem of joint estimation of power spectrum and modulation from realizations of frequency modulated stationary wideband signals is considered. The study is motivated by some specific signal classes from which departures to stationarity can carry relevant information and has to be estimated.**

The estimation procedure is based upon explicit modeling of the signal as a wideband stationary Gaussian signal, transformed by time-dependent, smooth frequency modulation. Under such assumptions, an approximate expression for the second order statistics of the transformed signal's Gabor transform is obtained, which leads to an approximate maximum likelihood estimation procedure.

The proposed approach is validated on numerical simulations.

I. Introduction

Usual time-frequency models for audio signals often rest upon expansions with respect to dictionaries of time-frequency waveforms, such as Gabor frames, wavelet frames, or more general families. Such descriptions are generally adequate for signal classes such as (voiced) speech, music,... where specific time-frequency localisation properties can be exploited. They are less effective for less structured signals, such as wideband sound signals.

We are concerned here with an alternative description of audio signals, aiming at describing different sound classes such as environmental noise, engine sound,... which are in addition non-stationary, in the sense that they carry information related to dynamics. As an example, think of an accelerating engine sound, where the acceleration can generally be perceived. This example suggests to study sound models, which we will term *timbre×dynamics*, in which a reference (stochastic) stationary signal, characterized by its timbre, is modulated by some dynamic deformation. Given such signals, a problem is to estimate the modulation (and possibly the underlying power spectrum). While many techniques have been developed for frequency modulation estimation for narrow band signals (see e.g. [1], the wideband case is more complex and has apparently received less attention.

A class of models based upon deformations of stationary processes has been proposed and studied in [2], motivated by the famous *shape from texture* image processing problem. A main aspect of the approach is based on the remark that a generic class of transformations can be represented by transport equations in a suitable representation space.

We adopt here a more explicit point of view, and limit to stationary Gaussian processes, transformed by a time-dependent modulation. We characterize the distribution of fixed time slices of a Gabor transform of such signals, and formulate the corresponding maximum likelihood estimation problem. As a result, we provide an estimation algorithm which is demonstrated on a small number of numerical examples.

II. Frequency modulation of stationary random signals

A. Notations and background

1) Random signals: We shall be concerned with complex Gaussian random signal models $\boldsymbol{X}$ of finite length L, which we shall assume zero-mean for the sake of simplicity. As is customary in finite-dimensional Gabor analysis, we shall also assume periodic boundary conditions, i.e. $\boldsymbol{X}_{t+L} = \boldsymbol{X}_t$. Given such a signal $\boldsymbol{X}$, we shall denote by $C_{\boldsymbol{X}}$ its covariance matrix, and by $R_{\boldsymbol{X}}$ its relation matrix (see [3] for details), defined as

$$C_{\boldsymbol{X}}(t,s) = \mathbb{E}\left\{X_t\overline{X}_s\right\}\ , \quad R_{\boldsymbol{X}}(t,s) = \mathbb{E}\left\{X_tX_s\right\}\ , \qquad (1)$$

and we will write $\boldsymbol{X} \sim \mathcal{CN}(0, C_{\boldsymbol{X}}, R_{\boldsymbol{X}})$. $\boldsymbol{X}$ is said to be *circular* if $R_{\boldsymbol{X}} = 0$.

2) Time-frequency representation: We shall use the following notations. Given a window function $\boldsymbol{g}$, the corresponding short time Fourier transform of a signal (STFT) $\boldsymbol{x} \in \mathbb{C}^L$ is defined by

$$\mathcal{V}_{\boldsymbol{g}}\boldsymbol{x}(m,n) = \sum_{t=0}^{L-1} x[t]\overline{g}[t-n]e^{-2i\pi m(t-n)/L}\ . \qquad (2)$$

Given lattice constants a and b (divisors of the signal length L) the corresponding Gabor transform reads

$$\mathcal{G}_{\boldsymbol{x}}[m,n] = \mathcal{V}_{\boldsymbol{g}}\boldsymbol{x}(mb,na)\ m = 0,\dots M-1,\ n = 0\dots N-1\ , \qquad (3)$$

with $M = L/b$ and $N = L/a$. $\mathcal{G}_{\boldsymbol{X}}$ is an $M \times N$ array. For suitably chosen $\boldsymbol{g}$, and a and b small enough, the Gabor transform is invertible (see [4], [5]); in finite dimensional situations, efficient algorithms have been developed and implemented (see [6]).

Remark 1 (Notations): As usual, summation bounds in the frequency domain depend of the parity of the signal length L. For the sake of simplicity, we introduce some notations and denote by I_L the integer interval $I_L = [1 - L/2, L/2]$ if L is even, and the integer interval $I_L = [-(L-1)/2, (L-1)/2]$ if L is odd. The corresponding positive frequencies interval will be denoted by $I_L^+ = [0, L/2]$ if L is even, and $I_L^+ = [0, (L-1)/2]$ if L is odd.

B. The model: definition and main estimates

We are concerned here in a simple model of signal transformation, which may be written as follows. We denote by $\boldsymbol{X}$ a zero-mean, wide sense stationary Gaussian random process, with covariance matrix $C_{\boldsymbol{X}}$, and by $\boldsymbol{Z}$ the associated analytic signal. We denote by $\mathcal{S}_{\boldsymbol{X}}$ the power spectrum of $\boldsymbol{X}$, and assume that $\mathcal{S}_{\boldsymbol{X}}(0) = 0$, and if L is even, that $\mathcal{S}_{\boldsymbol{X}}(L/2) = 0$. Under such an assumption, it is easy to show that $\boldsymbol{Z}$ is a circular complex Gaussian random vector (by a finite dimensional version of a standard argument, see e.g. [7]).

The observation is assumed to be the real part $\boldsymbol{Y}_r = \mathsf{Re}(\boldsymbol{Y})$ of a complex valued signal $\boldsymbol{Y}$; for the sake of simplicity we shall only work with the latter, assumed to be an USB (upper sideband) modulated version $\boldsymbol{Y}$ of a reference stationary signal $\boldsymbol{X}$, of the form

$$Y_t = Z_t e^{2i\pi\gamma(t)/L} + N_t \ , \tag{4}$$

where $\gamma \in C^2$ is an unknown smooth, slowly varying *modulation function*, and $\boldsymbol{N} = \{N_t,\, t = 0, \dots L-1\}$ is a real Gaussian white noise, with variance σ_0^2. Obviously, when γ is not a constant function, $\boldsymbol{Y}$ is not a wide sense stationary signal any more. The problem at hand is to estimate the unknown modulation γ and the original power spectrum $\mathcal{S}_{\boldsymbol{X}}$ from a single realization of $\boldsymbol{Y}$.

Clearly, $\boldsymbol{Z} \sim \mathcal{CN}(0, C_{\boldsymbol{Z}}, 0)$ is a circular complex Gaussian random signal, with covariance matrix

$$C_{\boldsymbol{Z}}(t,s) = \sum_{\nu \in I_L^+} \mathcal{S}_{\boldsymbol{X}}(\nu) e^{2i\pi\nu(t-s)/L} \ , \tag{5}$$

and is therefore wide-sense stationary.

In the proposed approach, we will base the estimation on a Gabor representation of the observed signal, and deliberately disregard correlations across time of the Gabor transform (hence focusing on time slices of the Gabor transform of the observation). The distribution of time slices of the analytic signal $\boldsymbol{Z}$ of the original signal is characterized in the following two results, which result from direct calculations.

Proposition 1: For fixed n, the Gabor transform $\mathcal{G}_{\boldsymbol{N}}[.,n]$ of the gaussian white noise is a stationary Gaussian random vector, with circular covariance matrix

$$C_{\mathcal{G}_{\boldsymbol{N}}}[m,m'] = \sigma_0^2 \sum_{k=0}^{L-1} \overline{\hat{g}}[k]\hat{g}[k-(m'-m)b] \tag{6}$$

Proposition 2: For fixed time index n, the Gabor transform $\mathcal{G}_{\boldsymbol{Z}}[.,n]$ of the analytic signal is a circular complex Gaussian random vector, with covariance matrix

$$C_{\mathcal{G}_{\boldsymbol{Z}}}[m,m'] = \sum_{k \in I_L^+} \mathcal{S}_{\boldsymbol{X}}[k]\overline{\hat{g}}[k-mb]\hat{g}[k-m'b] \tag{7}$$

The estimation of the modulation will be based upon an approximation of the covariance matrix of the observed signal. In a few words, the Gabor transform of the frequency modulated signal can be approximated by a deformed version of the Gabor transform of the original signal. The deformation takes the form of a time-varying frequency shift. A more precise argument, based upon first order approximation of the modulation function γ, leads to the following result.

Theorem 1: 1) For fixed time, the Gabor transform $\mathcal{G}_{\boldsymbol{Y}}$ may be approximated as

$$\mathcal{G}_{\boldsymbol{Y}}[m,n] = \mathbf{G}^{(n;\gamma'(na)/b)}[m] + R[m] \ , \tag{8}$$

where $\mathbf{G}^{(n;\delta)}$ is a frequency-shifted Gabor transform

$$\mathbf{G}^{(n;\delta)}[m] = \sum_{t=0}^{L-1} Z_t \overline{g}[t-na] e^{-2i\pi[m-\delta][t-an]/M} + \mathcal{G}_{\boldsymbol{N}}[m,n] \ , \tag{9}$$

and the remainder is bounded as follows: for all m, m',

$$\left|\mathbb{E}\left\{R[m]\overline{R}[m']\right\}\right| \leq \sigma_Z^2 \left(\frac{\pi e}{L}\|\gamma''\|_\infty \mu_2 + 2\mu_1\right)^2, \tag{10}$$

where σ_Z^2 is the variance of Z and with

$$\mu_1 = \sum_{t \in I_T^c} |g(t)| \ , \ \mu_2 = \sum_{t \in I_T} t^2 |g(t)| \ , \ T = \sqrt{\frac{L}{\pi\|\gamma''\|_\infty}} \tag{11}$$

where $I_T = [-T, T]$ and $I_T^c = I_L \backslash I_T$

2) Given δ, and for fixed n, $\mathbf{G}^{(n;\delta)}$ is distributed following a circular multivariate complex Gaussian law, with covariance matrix

$$C_{\mathbf{G}^{(n;\delta)}}[m,m'] = C_{\mathcal{G}_{\boldsymbol{Z}}}[m-\delta, m'-\delta] + \ C_{\mathcal{G}_{\boldsymbol{N}}}[m,m'] \ . \tag{12}$$

The estimation procedure described below is a maximum likelihood approach, which requires inverting the covariance matrix of vectors $\mathbf{G}^{(n;\delta)}$. The latter is positive semi-definite by construction, but not necessarily definite. The result below provides a sufficient condition on $\boldsymbol{g}$ and the noise for invertibility.

Proposition 3: Assume that the window $\boldsymbol{g}$ is such that

$$K_{\boldsymbol{g}} := \min_{t=0\dots L-1} \left(\sum_{k=0}^{b-1} |g[t+kM]|^2\right) > 0 \ . \tag{13}$$

Then for all $\boldsymbol{x} \in \mathbb{C}^M$,

$$\boldsymbol{x}^* C_{\mathbf{G}} \boldsymbol{x} \geq \sigma_0^2 K_{\boldsymbol{g}} \ , \tag{14}$$

and the covariance matrix is therefore boundedly invertible.

Remark 2: The condition may seem at first sight unnatural to Gabor frame experts. However, it simply expresses that the number M of frequency bins shouldn't be too large if one wants the covariance matrix to be invertible. However, reducing M also reduces the precision of the estimate, and a trade-off has to be found, as discussed in the next section.

C. Improving the frequency resolution

We propose here a method to improve the frequency resolution of our estimations. We have already seen that the invertibility of the covariance matrix requires that the number of frequency bins of the Gabor transform shouldn't be too large. As a result however, it may be convenient, as we shall

see later, to have access to the information contained in all the frequency frames of the short time Fourier transform defined in equation (2). For this purpose, we also consider alternative versions of the Gabor transform, associated with frequency-shifted sampling lattices:

$$\mathcal{G}^c_{\boldsymbol{x}}[m,n] = \mathcal{V}_{\boldsymbol{g}}\boldsymbol{x}(mb+c,na)\ , \quad m \in \mathbb{Z}_M,\ n \in \mathbb{Z}_N\ , \quad (15)$$

where $c \in [0, b-1]$. We now have at our disposal a collection of b Gabor transforms, which are all different subsampled versions of the STFT. The previous results and proofs remain valid with this new definition of the Gabor transform. Equations (8) and (9) now become

$$\mathcal{G}^c_{\boldsymbol{Y}}[m,n] = \mathbf{G}^{(n;\gamma'(na)/b+c/b)}[m] + R\ , \quad (16)$$

where

$$\begin{aligned}\mathbf{G}^{(n;\delta^c)}[m] &= \sum_{t=0}^{L-1} Z_t\overline{g}[t-na]e^{-2i\pi[m-\delta^c][t-an]/M} \\ &\quad + \ \mathcal{G}^c_{\boldsymbol{N}}[m,n]\ , \end{aligned} \quad (17)$$

and the associated Equation (12) now reads:

$$C_{\mathbf{G}^{(n;\delta^c)}}[m,m'] = C_{\mathcal{G}_{\boldsymbol{Z}}}[m-\delta^c, m'-\delta^c] + C_{\mathcal{G}_{\boldsymbol{N}}}[m,m']\,. \quad (18)$$

The rationale will be that a frequency shift δ^c can be estimated from each one of these thansforms, and the optimal one will be retained.

III. Estimation procedure

We now describe in some details the estimation procedure corresponding to our problem. The estimation problem is the following: from a single realization of the signal model (4), estimate the modulation function γ and the original power spectrum $\mathcal{S}_{\boldsymbol{X}}$. We first notice the indeterminacy in the problem, namely the fact that adding an affine function to γ is equivalent to shifting $\mathcal{S}_{\boldsymbol{X}}$. This has to be fixed by adding an extra constraint in the estimation procedure.

A. Maximum likelihood modulation estimation

We now turn to the estimation procedure, that exploits the above results. With the same notations as before, we fix a value of the time index n, and denote for simplicity by $\boldsymbol{\mathcal{G}} = \boldsymbol{\mathcal{G}}^{(n)}$ the corresponding fixed time slice of $\mathcal{G}^c_{\boldsymbol{Z}}$. Due to the multivariate complex Gaussian distribution of the signal and the fixed time Gabor transform slices, the log-likelihood of a slice takes the form

$$\mathcal{L}_\delta(\boldsymbol{\mathcal{G}}) = \boldsymbol{\mathcal{G}}^* \left(C_{\mathbf{G}^{(n;\delta^c)}}\right)^{-1} \boldsymbol{\mathcal{G}} + \ln\left(\pi^M \det(C_{\mathbf{G}^{(n;\delta^c)}})\right)\ . \quad (19)$$

Therefore, the maximum likelihood estimate for the frequency shift assumes the form

$$\hat{\delta}^c = \arg\min_{\delta^c}\left[\boldsymbol{\mathcal{G}}^*(C_{\mathbf{G}^{(n;\delta^c)}})^{-1}\boldsymbol{\mathcal{G}} + \ln\left(\pi^M \det(C_{\mathbf{G}^{(n;\delta^c)}})\right)\right]\ . \quad (20)$$

However, we notice that $\det(C_{\mathbf{G}^{(n;\delta^c)}})$ actually does not depend on the modulation parameter δ^c. Therefore the maximum likelihood estimate reduces to

$$\hat{\delta}^c = \arg\min_{\delta^c}\left[\boldsymbol{\mathcal{G}}^*\left(C_{\mathbf{G}^{(n;\delta^c)}}\right)^{-1}\boldsymbol{\mathcal{G}}\right]\ , \quad (21)$$

a problem to be solved numerically. Notice that this requires the knowledge of the covariance matrix $C_{\mathbf{G}^{(n;0)}}$ corresponding to the Gabor transform of the noisy stationary signal. The latter is generally not available, and has to be estimated as well.

As $\delta^c(n) \approx (\gamma'(an) - c)/b$, the estimates of δ for each n lead to an estimate of γ'. Since we solve the minimisation problem by an exhaustive search on the δ^c, the estimate of γ' is coarsely quantized (see Remark 2), as b is large and $\hat{\gamma}'(an) \in [c, b+c, 2b+c, .., (M-1)b+c]$. This problem is solved by using the family of frequency-shifted versions of Gabor transform described in subsection II-C and making a new exhaustive search on the $\hat{\delta}^c$

$$\hat{\delta} = \arg\min_{c}\left[\boldsymbol{\mathcal{G}}^*\left(C_{\mathbf{G}^{(n;\delta^c)}}\right)^{-1}\boldsymbol{\mathcal{G}}\right]\ . \quad (22)$$

The quantization effect on the final estimation of the modulation function is therefore attenuated, i.e. $\hat{\gamma}'(an) \in [0, L-1]$. Obtaining from this estimation a smoother estimate for the modulation function γ requires extra interpolation techniques.

Remark 3: As an alternative, one may also avoid exhaustive searches and seek minimizers in (21) using more elaborate numerical techniques, that would avoid quantization effects. This question is currently under study.

B. Estimation of the underlying covariance matrix

We now describe a method for estimating the covariance matrix $C_{\mathbf{G}^{(n;0)}}$. Suppose that an estimate $\hat{\gamma}$ of the modulation function γ is available. Then the signal $\boldsymbol{Y}$ can be demodulated by setting

$$\boldsymbol{U} = \boldsymbol{Y}e^{-2i\pi\hat{\gamma}/L}\ , \quad (23)$$

Clearly, $\boldsymbol{U}$ is an estimator of $\boldsymbol{Z} + \boldsymbol{N}e^{-2i\pi\gamma/L}$, the noisy stationary signal. We can now compute the covariance matrix $C_{\mathcal{G}_{\boldsymbol{U}}}$ of the Gabor transform of $\boldsymbol{U}$, which is an estimator of $C_{\mathcal{G}_{\boldsymbol{Z}}} + C_{\mathcal{G}_{\boldsymbol{N}}}$. Comparing with equation (9) we finally obtain an estimator for the covariance matrix

$$C_{\mathcal{G}_{\boldsymbol{U}}} \approx C_{\mathbf{G}^{(n;0)}} \quad (24)$$

Remark 4: The power spectrum $\mathcal{S}_{\boldsymbol{X}}$ of the stationary signal can be estimated from $\boldsymbol{U}$ using a standard Welch periodogram estimator, or by marginalizing the square modulus of the Gabor transform of the demodulated signal, as described in [4].

C. Summary of the estimation procedure

We now summarize an iterative algorithm to jointly estimate the covariance matrix $C_{\mathbf{G}^{(n;0)}}$ and the modulation function γ, that exploits alternatively the two procedures described above. The procedure is as follows, given a first estimation of the modulation function, we can perform a first estimation of the covariance matrix, which in turn allows us obtain a new estimation of the modulation function. The operation is repeated until the stopping criterion is satisfied.

For the initialization, we need a first modulation frequency estimate, for which we use the center of mass of the modulated signal Gabor transform

$$\hat{\delta}^{(0)}(n) = \frac{\sum_{m=0}^{M-1} m|\mathbf{G}^{(n;\delta)}|^2[m]}{\sum_{m=0}^{M-1}|\mathbf{G}^{(n;\delta)}|^2[m]}\ . \quad (25)$$

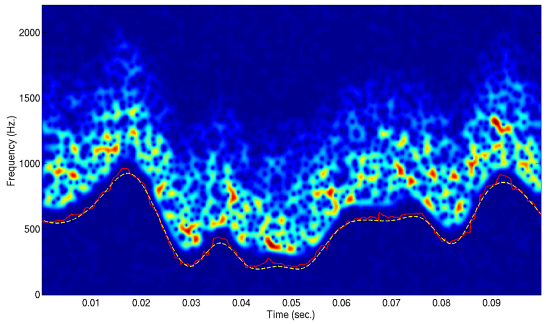


Fig. 1. Gabor transform of a frequency modulated synthetic stationary random signal, superimposed with the frequency modulation: estimate (red) and original (yellow).

The stopping criterion is based upon the evolution of the frequency modulation along the iterations. More precisely, we use the empirical criterion

$$\frac{||\hat{\delta}^{(k)} - \hat{\delta}^{(k+1)}||_2}{||\hat{\delta}^{(k+1)}||_2} < \epsilon \qquad (26)$$

The pseudo-code of the algorithm can be found below

Algorithm 1 Joint covariance and modulation estimation

Initialize as in (25)
while criterion (26) is false **do**
- Compute $\hat{\gamma}^{(k)}$ by interpolation from $\delta^{(k)}$.
- Demodulate $\boldsymbol{Y}$ using $\hat{\gamma}^{(k)}$ following (23)
- Compute the Gabor transform of the demodulated signal $\hat{\mathbf{G}}^{(k;n)}[m] = \mathcal{G}_{\boldsymbol{U}^{(k)}}[m,n]$
- Estimate $\hat{\delta}^{(k+1)}$ using the covariance matrix of $\hat{\mathbf{G}}^{(k;n)}$ from (21) and (22)
- $k := k+1$

end while

IV. Numerical results

The proposed estimation procedure has been implemented using MATLAB/OCTAVE, and relies on the LTFAT toolbox [8] for the time-frequency transforms.

We display in Fig. 1 an example of estimation result. The original signal was generated as pseudo-random stationary Gaussian signal with a smooth, wideband power spectrum, that was further modulated by a smooth frequency modulation function. Fig. 1 displays the Gabor transform of the modulated signal (positive frequencies only), together with the original and the estimate for the frequency modulation. For the sake of clarity, the frequency estimate has been displayed below the relevant part of the Gabor transform (remember that it is defined up to an additive constant). As can be seen, the result is fairly satisfactory, the estimated modulation follows closely the ground truth.

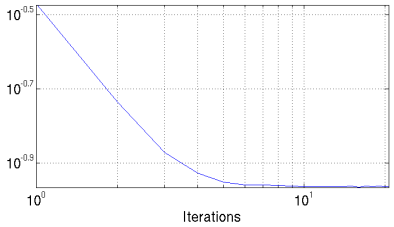


Fig. 2. Log-log plot of the evolution of the criterion proposed in (26) according on the number of iterations.

To asses the convergence properties of the proposed approach, the same experiment was run several times with the same modulation law and different seeds for the underlying stationary noise. We display in Fig. 2 the evolution of the criterion as a function of the iteration index, averaged over 20 realizations. Convergence appears to be fast, with power-law like decay speed.

V. Conclusion

We have presented in this paper a new approach for modulation frequency and power spectrum estimation from wideband signals, based upon explicit modeling. A main point that is exploited in our approach is the fact that modulations can be locally approximated by frequency shifts in the Gabor domain. The algorithm has been validated using numerical simulations, that show that when signals are generated according to the model of interest, very accurate results can be obtained.

Further developments include numerical tests on real signals, such as natural sounds generated by rolling bodies with variable speed,... We shall also consider extending this approach to other transformation models, such as time warping or more general transformations.

Acknowledgment

This work was supported by the ANR project Metason ANR-10-CORD-010.

References

[1] H. L. Van Trees, *Detection, estimation and modulation theory*. Wiley Interscience, 2003.
[2] M. Clerc and S. Mallat, "Estimating deformations of stationary processes," *Annals of Statistics*, vol. 31, no. 6, pp. 1772–1821, 2003.
[3] B. Picinbono, "Second-order complex random vectors and normal distributions," *IEEE Transactions on Signal Processing*, vol. 44, no. 10, pp. 2637–2640, 1996.
[4] R. Carmona, W. L. Hwang, and B. Torrésani, *Practical time-frequency analysis: Gabor and Wavelet Transforms With an Implementation in S*, C. K. Chui, Ed. Academic Press, 1998.
[5] K. Gröchenig, *Foundations of time-frequency analysis*, ser. Applied and Numerical Harmonic Analysis. Boston, MA: Birkhäuser Inc., 2001.
[6] P. L. Søndergaard, "Efficient Algorithms for the Discrete Gabor Transform with a long FIR window," *J. Fourier Anal. Appl.*, vol. 18, no. 3, pp. 456–470, 2012.
[7] B. Picinbono, "On circularity," *IEEE Transactions on Signal Processing*, vol. 42, no. 12, pp. 3473–3482, 1994.
[8] P. Søndergaard, B. Torrésani, and P. Balazs, "The linear time frequency analysis toolbox," *International Journal of Wavelets and Multiresolution Information Processing*, vol. 10, no. 4, pp. 1 250 032–1 – 1 250 032–27, 2012.

Gabor dual windows using convex optimization

Nathanaël Perraudin, Nicki Holighaus, Peter L. Søndergaard and Peter Balazs
Acoustics Research Institute, Austrian Academy of Sciences, Wohllebengasse 12–14, 1040 Vienna, Austria
Email: nathanael.perraudin@epfl.ch, nicki.holighaus@oeaw.ac.at, soender@kfs.oeaw.ac.at, peter.balazs@oeaw.ac.at

***Abstract*—Redundant Gabor frames admit an infinite number of dual frames, yet only the canonical dual Gabor system, constructed from the minimal ℓ^2-norm dual window, is widely used. This window function however, might lack desirable properties, such as good time-frequency concentration, small support or smoothness. We employ convex optimization methods to design dual windows satisfying the Wexler-Raz equations and optimizing various constraints. Numerical experiments show that alternate dual windows with considerably improved features can be found.**

I. Introduction

Time-frequency representations, in particular *Gabor transforms* [9], i.e. sampled Short-Time Fourier transforms, are ubiquitous in signal processing. Gabor transforms represent a signal as linear combination of translates and modulations of a single *window function*, which for best results should be chosen to be well-concentrated in time and frequency.

A signal can be reconstructed from its Gabor transform using a dual system with the same modulation and translation structure. Moreover, infinitely many such systems exist if the Gabor transform is redundant. Finding a dual system with desirable properties given a prescribed analysis window is the topic of this paper.

More explicitly, for $g \in \ell^2(\mathbb{Z})$, and $a, M \in \mathbb{Z}$, we define the Gabor system

$$\mathcal{G}(g,a,M) := \left(g_{m,n} = g[\cdot - na]e^{2\pi im\cdot/M}\right)_{n\in\mathbb{Z},\ m=0,\dots,M-1}. \tag{1}$$

If $\mathcal{G}$ is also a *frame* [5], we refer to the system as a *Gabor frame*. For $f \in \ell^2(\mathbb{Z})$, the corresponding Gabor transform is given by

$$(\mathbf{G}f)[m+nM] = \langle f, g_{m,n}\rangle = \sum_{l\in\mathbb{Z}} f[l]\overline{g_{m,n}[l]}, \tag{2}$$

with the analysis operator $\mathbf{G}$ as given by the infinite matrix $\mathbf{G}[m+nM,l] := \mathbf{G}_{g,a,M}[m+nM,l] := \overline{g_{m,n}[l]}$.

Gabor synthesis is performed by applying the adjoint of $\mathbf{G}$ to a coefficient sequence $c \in \ell^2(\mathbb{Z})$. The action of the synthesis operator can be equivalently described as

$$f_{syn}[l] = (\mathbf{G}^*c)[l] = \sum_{m,n} c[m+nM]g[l-na]e^{2\pi iml/M}. \tag{3}$$

The concatenation $\mathbf{S} = \mathbf{G}^*\mathbf{G}$ of the analysis and synthesis operators is called the *frame operator*.

Reconstruction can be realized using the so-called *canonical dual* system, obtained by inverting $\mathbf{S}$ and defined as

$$\tilde{g}_{m,n} = \mathbf{S}^{-1}g_{m,n}. \tag{4}$$

In the particular case of Gabor frames, the canonical dual system is again a Gabor frame, i.e. it equals $\mathcal{G}(\tilde{g}_{0,0},a,M)$. Therefore we refer to $\tilde{g} = \tilde{g}_{0,0} = \mathbf{S}^{-1}g$ as the canonical dual window.

The synthesis operator of $\tilde{g}$ coincides with the pseudo-inverse of the original analysis operator, i.e. $\mathbf{G}^*_{\tilde{g},a,M} = \mathbf{G}^\dagger$. So the inversion formula reads

$$f[l] = \sum_{m,n}\langle f, g_{m,n}\rangle\tilde{g}_{m,n}[l] = \mathbf{G}^\dagger\mathbf{G}f[l]. \tag{5}$$

There are several approaches for finding the canonical dual in an efficient way, e.g. [4], [11]. Only if the length of the window L_g is less than or equal to the number of channels M, is the canonical dual guaranteed to have the same length. This so-called *painless case* construction is omnipresent in signal processing, to the point where M and L_g are not distinguished.

Redundant Gabor frames possess infinitely many dual Gabor frames of the form $\mathcal{G}(h,a,M)$, any of which facilitates perfect reconstruction from unmodified coefficients. On the other hand, whenever the coefficient representation is processed, varying dual systems provide different reconstructions and the features of the chosen system suddenly play an important role. Some of the 'alternate duals' might possess properties preferable to those of the canonical dual, e.g. shorter support, better localization or smoothness.

For a Gabor frame $\mathcal{G}(h,a,M)$, the Wexler-Raz equations [17], [20] provide a necessary and sufficient condition to constitute a dual frame for $\mathcal{G}(g,a,M)$. Using this hard constraint, a convex optimization problem can be defined by adding functionals to be minimized that provide desired properties.

Recently, convex optimization in the context of audio signal processing has grown into a active field of research and in particular proximal splitting methods [6], [7], [8] have been used to great effect, e.g. in audio inpainting [2], [1] and sparse representation [12]. In those cases, optimization techniques are applied directly to the signal or its time-frequency representation. In this contribution, we apply optimization techniques to shape the building blocks of the time-frequency representation instead. Since a systematic evaluation of the available optimization techniques is beyond the scope of this contribution, we only present an exemplary realization.

Our method is a much more general approach than the construction of non-canonical dual windows found in [19] and optimizes several criteria at once. One particular application of the proposed approach is the construction of smooth dual windows satisfying a support constraint. To illustrate the viability of our method, we choose a Gabor frame $\mathcal{G}(g,a,M)$

with g being an FIR window, i.e. a window function supported on a finite interval I_g, and construct a smooth dual window h supported on an interval I_h.

II. Gabor frames

In this contribution, we consider Gabor systems $\mathcal{G}(g,a,M)$ in $\ell^2(\mathbb{Z})$. Such a system constitutes a frame if constants $0 < A \leq B < \infty$ exist, such that

$$A\|f\|_2^2 \leq \|\mathbf{G}f\|_2^2 \leq B\|f\|_2^2, \text{ for all } f \in \ell^2(\mathbb{Z}). \tag{6}$$

In that case, the closed linear span of its elements equals $\ell^2(\mathbb{Z})$ and every sequence $f \in \ell^2(\mathbb{Z})$ can be written as

$$f = \mathbf{G}^* c, \tag{7}$$

for some coefficient sequence $c \in \ell^2(\mathbb{Z})$. In particular, if $\mathcal{G}(h,a,M)$ is a dual Gabor frame, $c = \mathbf{G}_{h,a,M} f$ is one possible choice. Note that frames are "mutually dual", i.e. the role of $\mathcal{G}(g,a,M)$ and $\mathcal{G}(h,a,M)$ in the considerations above can be switched at will.

The Wexler-Raz equations [20], [17] for $\ell^2(\mathbb{Z})$ provide a necessary and sufficient condition for a function $h \in \ell^2(\mathbb{Z})$ to be a dual Gabor window for $\mathcal{G}(g,a,M)$. They are given by

$$\frac{M}{a}\left\langle h, g[\cdot - nM]e^{2\pi i m \cdot/a}\right\rangle = \delta[n]\delta[m], \tag{8}$$

for $m = 0, ..., a-1,\ n \in \mathbb{Z}$. In the equation above, $\delta[l]$ denotes the Kronecker delta at position l. In terms of the analysis matrix $\mathbf{G}^\circ = \mathbf{G}_{g,M,a}$, i.e. switching the role of a and M, they can be stated as

$$\mathbf{G}^\circ h = \frac{a}{M}\delta. \tag{9}$$

III. Proximal splitting methods

The convex optimization problems we consider are of the form

$$\underset{x\in\mathbb{R}^L}{\text{minimize}} \sum_{i=1}^{K} f_i(x), \tag{10}$$

where the f_i are convex functions. Note that if at least one function f_i is not differentiable, it is not possible to apply smooth optimization techniques. Proximal splitting methods [7], on the other hand may still apply. The term proximal splitting originates from the fact that each function f_i is minimized iteratively with the help of their corresponding *proximity operator*, a generalization of convex projection operators, defined as follows.

Definition 1. The proximity operator of a function $f \in \Gamma_0(\mathbb{R}^L)$ is defined by

$$\text{prox}_f(y) := \underset{x\in\mathbb{R}^L}{\text{argmin}} \left\{ \frac{1}{2}\|y - x\|_2^2 + f(x) \right\}. \tag{11}$$

Since f is convex, the minimization problem in (11) has a unique solution for every $y \in \mathbb{R}^L$ and consequently $\text{prox}_f : \mathbb{R}^L \to \mathbb{R}^L$ is well-defined.

More information on the properties of proximity operators can be found in [16], [13].

From now on, we will denote by i_C the indicator function [7], of a non-empty, closed and convex set $C \subset \mathbb{R}^L$ by

$$i_C : \mathbb{R}^L \to \{0, +\infty\} : x \mapsto \begin{cases} 0, & \text{if } x \in C \\ +\infty & \text{otherwise} \end{cases} \tag{12}$$

and by $\Gamma_0(\mathbb{R}^L)$ the class of functions

$$\Gamma_0(\mathbb{R}^L) = \big\{ f : \mathbb{R}^L \mapsto \mathbb{R} \ : \ f \text{ lower semi-continuous, convex and proper} \big\}.$$

Indicator functions can be used to add hard constraints, e.g. a set of linear equations that the solution must satisfy, to an optimization problem of the form (10). More explicitly,

$$\underset{x\in\mathcal{C}}{\text{argmin}} \sum_{i=1}^{K} \lambda_i f_i(x) = \underset{x\in\mathbb{R}^L}{\text{argmin}} \sum_{i=1}^{K} \lambda_i f_i(x) + i_C, \tag{13}$$

where $\mathcal{C} = \{x \in \mathbb{R}^L \ : \ x \text{ satisfies the hard constraints } \}$ is the *set of admissible points*. If $\mathcal{C}$ is non-empty and convex, Equation (13) has a solution for any given choice of regularization parameters λ_i, uniquely determined if at least one f_i is strictly convex.

Table I presents a list of commonly used regularizer functions f_i that can be combined to tune the solution x.

Table I
Some regularization functions

Function	Effect on the signal
$\|x\|_1$	sparse representation in time
$\|\mathcal{F}x\|_1$	sparse representation in frequency
$\|\nabla x\|_2^2$	smooth representation in time / concentrated in frequency
$\|\nabla \mathcal{F}x\|_2^2$	smooth representation in frequency / concentrated in time
$\|x\|_2^2$	spread values more evenly
$i_{\mathcal{C}}(x)$	force $x \in \mathcal{C}$

We decided to present a solution of (10) using the parallel proximal algorithm (PPXA, Algorithm 1). However, this contribution does not intend to propose the best method to solve (10), and other algorithms, e.g. generalized forward backward [15], might prove more efficient. Instead, we focus on a new formulation of the problem of finding dual Gabor windows.

In the next section we present one of the possible ways to solve (10). Optimality studies are beyond the scope of this paper and planned as future work.

IV. Methods

Utilizing the theory established in the previous sections, we can now describe our method in detail. We intend to compute non-canonical dual windows for a given Gabor frame $\mathcal{G}(g,a,M)$, where g is an analysis windows supported on some finite interval I_g. Furthermore, we want the dual window to

Algorithm 1 Parallel proximal algorithm (PPXA)

Initialize $\epsilon \in]0,1[$, $\tilde{\bar{g}} > 0$, $(\omega_i)_{1\leq i\leq K} \in]0,1]^K$ with $\sum_{i=1}^K \omega_i = 1$, $y_{1,0} \in \mathbb{R}^L$, ..., $y_{K,0} \in \mathbb{R}^L$
Fix $\theta \in [\epsilon, 2-\epsilon[$
$x_0 \leftarrow \sum_{i=1}^K \omega_i y_{i,0}$
for $n = 1, 2, \ldots$ **do**
  for $i = 1, ..., K$ **do**
    $p_{i,n} \leftarrow \mathrm{prox}_{\tilde{\bar{g}} f_i/\omega_i}(y_{i,n})$
  end for
  $p_n \leftarrow \sum_{i=1}^K \omega_i p_{i,n}$
  for $i = 1, ..., K$ **do**
    $y_{i,n+1} \leftarrow y_{i,n} + \theta\,(2p_n - x_n - p_{i,n})$
  end for
  $x_{n+1} \leftarrow x_n + \theta\,(p_n - x_n)$
end for

be supported on an interval I_h and denote the convex set of all signals satisfying this constraint by $\mathcal{C}_{\text{supp}}$.

Considering the support constraint, we see that all but a small subset of the Wexler-Raz equations are trivially satisfied. Without loss of generality we assume I_g and I_h to be centered around 0. Noting that $I_g \cap (I_h + nM) = \emptyset$ for $|n| \geq \frac{L_g+L_h}{2M}$, only the equations for

$$|n| < \frac{L_g + L_h}{2M}, \tag{14}$$

can possibly be non-zero. This makes a total of $2a(\lceil \frac{L_g+L_h}{2M} \rceil + 1)$ equations in L_h unknowns. As a consequence, we are not required to consider sequences of infinite length to compute the dual window, but we can equivalently work with signals in $\mathbb{C}^L$, where L is some multiple of a and M satisfying $L \geq L_g + L_h + 1$.

The solutions of the non-trivial equations from the Wexler-Raz equation system (8), numbered as in (14) form a convex set written $\mathcal{C}_{\text{dual}}$, providing the second hard constraint after the support condition.

Then, $\mathcal{C} = \mathcal{C}_{\text{dual}} \cap \mathcal{C}_{\text{supp}}$ is also convex and if non-empty[1] forms a legal set of admissible points for a problem of the form (13). To shape the resulting dual window towards some useful properties, we select suitable regularization functions (Table I) and parameters, employing PPXA to solve the resulting convex optimization problem, converging to the unique solution. The indicator functions $i_{\mathcal{C}_{\text{dual}}}$ and $i_{\mathcal{C}_{\text{supp}}}$ are used to realize the duality and support constraints.

Experience shows that PPXA needs a large number of iterations to perfectly satisfy the hard constraints. To speed up this process, a final projection is performed once the algorithm converges to a certain accuracy. If there is more than one regularization function to be minimized, the projection is realized by a POCS (Projection Onto Convex Set) algorithm [10], [21], governed by the updating rule

$$x_{n+1} = P_{\mathcal{C}_{\text{supp}}}\left(P_{\mathcal{C}_{\text{dual}}}(x_n)\right).$$

[1]To determine whether C is non-empty is a nontrivial task and investigating this set is planned for future work. In the experiments conducted so far, the support constraints and redundancy were determined heuristically.

A. Compactly supported duals by truncation

In [19], Strohmer proposed a simple algorithm for the computation of compactly supported dual windows, which we will call the *truncation method*. Strohmer proposed to truncate the Wexler-Raz equations as described in the previous section and then solve the resulting equation system by computing the Moore- Penrose inverse, obtaining the least-squares solution. While the resulting windows satisfy the duality conditions, they are not very smooth and indeed show some discontinuity-like behavior, see also Figure 1(e,f). One of the goals of this contribution is the improvement of these undesirable effects.

V. Numerical results

We present the construction of a smooth dual Gabor window with short support. Our setup considers $\mathcal{G}(g, 30, 60)$, i.e. a system with redundancy 2, where g is a "Nuttall" window [14] of length $L_g = 120$ samples, see Figure 1(a,b).

We aim at computing a dual that is supported on the same interval as the analysis prototype, yielding $\mathcal{C}_{\text{supp}} = \{x \in \mathbb{R}^L : \text{supp}(x) \subseteq \text{supp}(g)\}$. To further provide reasonable localization and smoothness, we select the regularization functions $f_1 = \|\cdot\|_1$, $f_2 = \|\mathcal{F}(\cdot)\|_1$, $f_3 = \|\nabla(\cdot)\|_2^2$ and $f_4 = \|\nabla\mathcal{F}(\cdot)\|_2^2$. The result shown in Figure 1(c,d) shows the optimal dual window with regards to the regularization parameters $\lambda_1 = \lambda_2 = 0.001$ and $\lambda_3 = \lambda_4 = 1$. Those values are choosen experimentally by considering that they are balancing the effect of the regularization functions as described in Table I. As reference, we included the least-squares solution provided by the truncation method, see Figure 1(e,f).

Minimizing the selected regularization functions improves upon the desired features, in particular smoothness (or frequency localization) with f_3 and time localization with f_4. The functions f_1 and f_2 avoid the solution to have a "M-shape", i.e. multiple peaks. This is unwanted as the temporal or frequency positions becomes ambiguous. Indeed, minimizing the l^1-norm will push all big coefficients to similar values.

The solution provided is assumed to perform perfect reconstruction on any signal with admissible length greater or equal to L. More precisely, by [11, Eq. (60)], the maximum relative reconstruction error can be shown to be of the order of the precision of the machine, more precisely at $4.5e^{-14}$.

Simulations were performed using the LTFAT [18] and the UNLocBoX matlab toolbox. A reproducible research addendum is available in http://unlocbox.sourceforge.net/rr/gdwuco.

In the experiment above, we constructed a smooth, well localized dual window, compactly supported with $L_h = 120$.

Considering the painless case, to guarantee the canonical dual window to be supported on $L_{\tilde{g}} = L_g$, enforces $M \geq 120$ therefore increasing the redundancy twofold, an unwanted side effect. Alternatively, in this setting, we could decide to keep the parameters $a = 30$, $M = 60$ fixed, but decrease the window size to $L_g \leq 60$. However, this construction provides a system with a more than 8 times larger frame bound ratio. Consequently, the resulting canonical dual window $\tilde{g}$, shown in Figure 2, shows bad frequency behavior and an undesirable, M-like shape in time. In contrast, the method proposed in

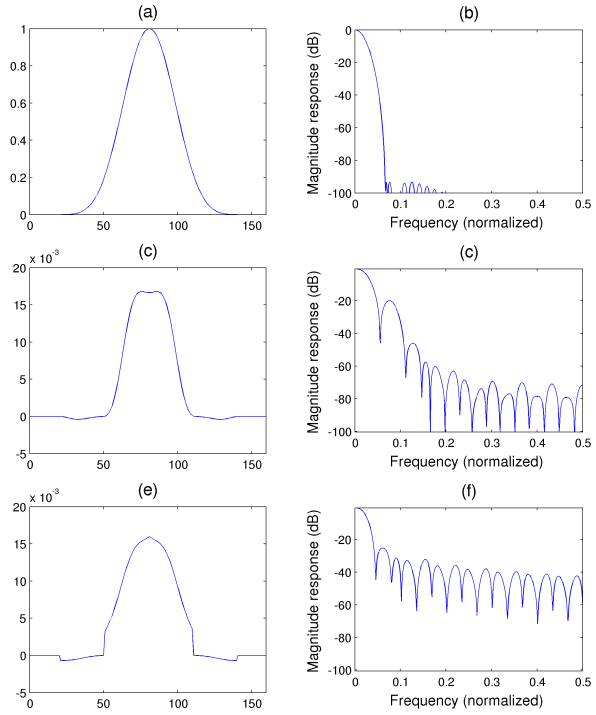


Figure 1. Experiments. (a) Analysis window in time. (b) Analysis window in frequency. (c) Synthesis window in time. (d) Synthesis window in frequency. (e) Truncation method in time. (f) Truncation method in frequency.

this manuscript allows the use of nicely shaped, compactly supported dual Gabor windows at low redundancies, without the strong restrictions of the painless case.

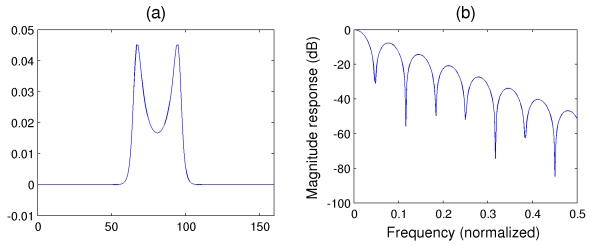


Figure 2. Half-overlap painless case construction ($\mathcal{G}(g, 30, 60)$, $L_g = 60$): Canonical dual window in time (a) and in frequency (b).

VI. Conclusion

We have proposed an algorithm for the design of non-canonical dual Gabor windows based on methods from convex optimization. Contrary to earlier methods, the algorithm discussed in this manuscript allows users to tune the dual window with regards to different desirable criteria. To illustrate the usefulness of the algorithm, we provided an example using a hard support constraint and shaped the window into a smooth shape using ℓ^1 priors on the window and its Fourier transform, as well as an ℓ^2 prior on its gradient. The result obtained considerably outperforms the result of an older method [19] that does not employ any smoothness constraints.

Our method can be applied in various situations to construct dual frames with properties more important for application than minimal ℓ^2-norm. Future work will further be concerned with applying the findings herein to frames with a different structure, e.g. nonstationary Gabor frames [3].

Acknowledgement

This work was supported by the Austrian Science Fund (FWF) START-project FLAME (“Frames and Linear Operators for Acoustical Modeling and Parameter Estimation”; Y 551-N13).

We would like to thank David Shuman for his useful comments about the paper.

References

[1] A. Adler, V. Emiya, M. Jafari, M. Elad, R. Gribonval, and M. Plumbley. A constrained matching pursuit approach to audio declipping. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 329–332. IEEE, 2011.

[2] A. Adler, V. Emiya, M. Jafari, M. Elad, R. Gribonval, and M. Plumbley. Audio inpainting. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(3):922–932, 2012.

[3] P. Balazs, M. Dörfler, N. Holighaus, F. Jaillet, and G. Velasco. Theory, implementation and applications of nonstationary Gabor frames. *Journal of Computational and Applied Mathematics*, 236(6):1481–1496, 2011.

[4] P. Balazs, H. G. Feichtinger, M. Hampejs, and G. Kracher. Double preconditioning for Gabor frames. *IEEE T. Signal. Proces.*, 54(12):4597–4610, December 2006.

[5] O. Christensen. *An Introduction to Frames and Riesz Bases*. Birkhäuser, 2003.

[6] P. Combettes and J. Pesquet. A Douglas–Rachford splitting approach to nonsmooth convex variational signal recovery. *IEEE Journal of Selected Topics in Signal Processing*, 1(4):564–574, 2007.

[7] P. Combettes and J. Pesquet. Proximal splitting methods in signal processing. *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, pages 185–212, 2011.

[8] P. Combettes and V. Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale Modeling & Simulation*, 4(4):1168–1200, 2005.

[9] H. G. Feichtinger and T. Strohmer, editors. *Gabor Analysis and Algorithms*. Boston, 1998.

[10] L. Gubin, B. Polyak, and E. Raik. The method of projections for finding the common point of convex sets. *USSR Computational Mathematics and Mathematical Physics*, 7(6):1–24, 1967.

[11] A. Janssen and P. L. Søndergaard. Iterative algorithms to approximate canonical Gabor windows: Computational aspects. *J. Fourier Anal. Appl.*, 13(2):211–241, 2007.

[12] M. Kowalski, K. Siedenburg, and M. Dörfler. Social sparsity! neighborhood systems enrich structured shrinkage operators. *Signal Processing, IEEE Transactions on*, 61(10):2498–2511, 2013.

[13] B. Martinet. Détermination approchée d’un point fixe d’une application pseudo-contractante. cas de l’application prox. *CR Acad. Sci. Paris Ser. AB*, 274:A163–A165, 1972.

[14] A. Nuttall. Some windows with very good sidelobe behavior. *IEEE Trans. Acoust. Speech Signal Process.*, 29(1):84–91, 1981.

[15] H. Raguet, J. Fadili, and G. Peyré. Generalized forward-backward splitting. *arXiv preprint arXiv:1108.4404*, 2011.

[16] R. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14(5):877–898, 1976.

[17] P. L. Søndergaard. Gabor frames by Sampling and Periodization. *Adv. Comput. Math.*, 27(4):355 –373, 2007.

[18] P. L. Søndergaard, B. Torrésani, and P. Balazs. The Linear Time Frequency Analysis Toolbox. *International Journal of Wavelets, Multiresolution Analysis and Information Processing*, 10(4), 2012.

[19] T. Strohmer. Numerical algorithms for discrete Gabor expansions. In Feichtinger and Strohmer [9], chapter 8, pages 267–294.

[20] J. Wexler and S. Raz. Discrete Gabor expansions. *Signal Process.*, 21(3):207–221, 1990.

[21] D. Youla and H. Webb. Image restoration by the method of convex projections: Part 1: Theory. *Medical Imaging, IEEE Transactions on*, 1(2):81–94, 1982.

Sparse Finite Gabor Frames for Operator Sampling

Götz E. Pfander
School of Engineering and Science,
Jacobs University Bremen,
28759 Bremen, Germany
Email: g.pfander@jacobs-university.de

David Walnut
Dept. of Mathematical Sciences,
George Mason University,
Fairfax, VA 22030 USA
Email: dwalnut@gmu.edu

***Abstract*—We derive some interesting properties of finite Gabor frames and apply them to the *sampling* or *identification* of operators with bandlimited Kohn-Nirenberg symbols, or equivalently those with compactly supported spreading functions. Specifically we use the fact that finite Gabor matrices are full Spark for an open, dense set of window vectors to show the existence of periodically weighted delta trains that identify simultaneously large operator classes. We also show that sparse delta trains exist that identify operator classes for which the spreading support has small measure.**

I. Introduction

A. Operator Sampling

The goal in *operator identification* is to determine an operator completely from its action on a single input function or distribution. If the operator models a linear (time-varying) communication channel, then the problem is one of *channel identification* and can be thought of as a generalization of the fact that the impulse response of a time-invariant communication channel modeled as a convolution operator can be determined from the response of the channel to a unit impulse. The question of determining which operators can be identified was addressed in foundational and pioneering work of T. Kailath ([3], [4], [5]) and P. Bello ([1]) who determined that the identifiability of a communication channel is characterized by the area of the support of its so-called *spreading function*. This work has been extended and placed on a firm mathematical footing in [6] and [8].

To be specific and to fix ideas for this paper, we restrict our attention to the class of Hilbert-Schmidt operators H on $L^2(\mathbf{R})$. Any such operator can be represented as a pseudodifferential operator as

$$Hf(x) = \int \sigma_H(x,\xi)\widehat{f}(\xi)\, e^{2\pi i x\xi}\, d\xi.$$

$\sigma_H(x,\xi) \in L^2(\mathbf{R}^2)$ is the *Kohn-Nirenberg* (KN) symbol of H. The spreading function $\eta_H(t,\nu)$ of the operator H is the *symplectic Fourier transform* of the KN symbol, viz.

$$\eta_H(t,\nu) = \iint \sigma_H(x,\xi)\, e^{-2\pi i(\nu x - \xi t)}\, dx\, d\xi$$

and we have the representation

$$Hf(x) = \iint \eta_H(t,\nu)\, T_t\, M_\nu f(x)\, d\nu\, dt$$

where $T_t f(x) = f(x-t)$ is the *time-shift operator* and $M_\nu f(x) = e^{2\pi i\nu x} f(x)$ is the *frequency-shift operator*. In this sense, an operator H whose spreading function has compact support can be said to have a *bandlimited symbol*. This motivates the following definition. Given a compact set $S \subseteq \mathbf{R}^2$, we define the *operator Paley-Wiener space* $OPW(S)$ to be the set of all Hilbert-Schmidt operators H on $L^2(\mathbf{R})$ with $\operatorname{supp}\eta_H \subseteq S$. Identifiability of an operator H therefore means informally that there exists a distribution g such that H is completely determined by Hg. To be more precise, suppose that $\mathcal{H}$ is some class of linear operators with common domain. We say that g *identifies* $\mathcal{H}$ if whenever $H_1, H_2 \in \mathcal{H}$ and $H_1 g = H_2 g$ (or equivalently $(H_1 - H_2)g = 0$) then $H_1 = H_2$. If $\mathcal{H}$ is a linear space, then g identifies $\mathcal{H}$ if and only if $H \in \mathcal{H}$ and $Hg = 0$ implies $H = 0$. However, these notions are not equivalent if $\mathcal{H}$ is not a linear space.

The following theorem was proved in [8] following Bello's work.

Theorem 1. If $|S| < 1$ then $OPW(S)$ is identifiable, and if $|S| > 1$ then $OPW(S)$ is not identifiable. In the former case an identifier has the form $g = \sum_n c_n\, \delta_{nT}$ for some $T > 0$ and periodic sequence $c = (c_n)$.

Since in this case, the operator is being "sampled" by a succession of evenly-spaced weighted impulses, and because the theory bears many formal analogies to the classical sampling of bandlimited functions, this procedure is called *operator sampling*. Indeed, it is shown in [9] that classical sampling is in fact a special case of operator sampling.

B. Gabor Matrices

Definition 2. Given $L \in \mathbf{N}$, let $\omega = e^{2\pi i/L}$ and define the *translation operator* T on $(x_0, \ldots, x_{L-1}) \in \mathbf{C}^L$ by

$$Tx = (x_{L-1}, x_0, x_1, \ldots, x_{L-2}),$$

and the *modulation operator* M on $\mathbf{C}^L$ by

$$Mx = (\omega^0 x_0, \omega^1 x_1, \ldots, \omega^{L-1} x_{L-1}).$$

Define the *full Gabor system matrix* $G(c)$ to be the $L \times L^2$ matrix given by

$$G(c) = [\, D_0\, W_L \mid D_1\, W_L \mid \cdots \mid D_{L-1}\, W_L \,]$$

where D_k is the diagonal matrix with diagonal $T^k c$, and where W_L is the $L \times L$ Fourier matrix $W_L = (e^{2\pi i nm/L})_{n,m=0}^{L-1}$.

The *finite Gabor system with window* c is the collection $\{M^p T^q c\}_{q,p=0}^{L-1}$.

For generic vectors $c \in \mathbf{C}^L$, the finite Gabor system with window c is a tight frame for $\mathbf{C}^L$, and in fact is a so-called *full Spark frame*. In particular the following holds.

Theorem 3. [7] Suppose that L is prime. Then there is an open, dense set of $c \in \mathbf{C}^L$ with the property that every square submatrix of $G(c)$ has nonzero determinant. In particular, this implies that every collection of columns in $G(c)$ has full rank.

Outline of Proof: Given any $N \times N$ submatrix, M, of $G(c)$, $\det(M)$ is a homogeneous polynomial of degree L in the variables $c_0, c_1, \ldots, c_{L-1}$, and it is sufficient to show that this polynomial does not vanish identically, and for that it suffices to show that there is a monomial in $\det(M)$ with a nonzero coefficient. In the proof such a monomial, p_M, is constructed recursively as follows.

If $N = 1$ then M is simply a multiple of a single variable c_j and we define $p_M = c_j$. If $N > 1$, let c_j be the variable of lowest index appearing in M. Choose any entry of M in which c_j appears, eliminate from M the row and column containing that entry, and call the remaining matrix M'. Define $p_M = c_j\, p_{M'}$. The coefficient of p_M in $\det M$ is a product of minors of W_L. Since L is prime, by Chebotarëv's Theorem, such minors never vanish.

C. Operator Sampling and Gabor Matrices

To illustrate the connection between operator sampling and Gabor matrices, we outline the proof of the sufficiency direction of Theorem 1 ([8], [9]).

Suppose that $|S| < 1$ is compact, and assume without loss of generality that S is contained in the first quadrant. Choose L prime so large that $S \subseteq [0, \sqrt{L}]^2$ and S meets no more than L rectangles of the form

$$R_{q,m} = [0, 1/\sqrt{L}]^2 + (q/\sqrt{L}, m/\sqrt{L}).$$

In the sequel, let $T\Omega = 1/\sqrt{L}$. For any sequence $c = (c_n)$ with period L, a straightforward calculation ([9], [8]) yields

$$\overline{Z}(t,\nu) = G(c)\,\overline{\eta}(t,\nu) \tag{1}$$

where

$$\overline{Z}(t,\nu) = [e^{-2\pi i pq/L}\, e^{-2\pi i \nu T p}\, (Z_{1/\Omega} \circ H) g(t + Tp, \nu)]_{p=0}^{L-1},$$

$Z_{1/\Omega} f(t,\nu) = \sum_{n \in \mathbf{Z}} f(t - n/\Omega)\, e^{2\pi i n\nu/\Omega}$ is the *Zak transform*, and

$$\overline{\eta}(t,\nu) = [e^{-2\pi i qm/L}\, e^{-2\pi i \nu T q}\, \eta_H(t + Tq, \nu + \Omega m)]_{q,m=0}^{L-1}.$$

Note that (1) is a linear system of L equations in L^2 unknowns, the coefficients of which are a discrete Gabor system. Because S meets no more than L rectangles $R_{q,m}$, (1) reduces to a system of L equations in L unknowns, with the reduced matrix $G_0(c)$ an $L \times L$ submatrix of $G(c)$. We now invoke Theorem 3 to assert that there is a choice of $c \in \mathbf{C}^L$ making $G_0(c)$ invertible.

II. Operators with Unknown Spreading Support

Theorem 3 says that the set of vectors c for which every square submatrix of $G(c)$ is invertible is dense and open. Since there are only finitely many such submatrices, there exists a dense, open set of $c \in \mathbf{C}^L$ such that all square submatrices of $G(c)$ are invertible. Hence c can be chosen independently of S, depending only on L.

Definition 4. Given $\Sigma > 0$ and $0 \le \Delta \le 1$, define the operator class $\mathcal{H}_\Sigma(\Delta)$ to be the collection of operators H in $OPW([-\Sigma, \Sigma]^2)$ such that $\operatorname{supp} \eta_H$ is contained in no more than Σ Jordan regions (that is, Jordan curves together with their interiors) with total area no more than $\Delta - 1/\Sigma$, and whose boundaries have total length no more than Σ.

Note that $\mathcal{H}_\Sigma(\Delta)$ is not a linear space, but has the property that the spreading supports admit uniformly good coverings by squares. A more general version of the following theorem appears in [9] (see [2] for a characterization in the case of fixed L.).

Theorem 5. Let $\Sigma > 0$ be given. Then for every sufficiently large prime L, there is a $c \in \mathbf{C}^L$ such that with $g = \sum_n c_n\, \delta_{n/\sqrt{L}}$, if $H \in \mathcal{H}_\Sigma(1)$ and $Hg = 0$, then $H = 0$. It follows immediately that if $\Delta \le 1/2$, then whenever $H_1, H_2 \in \mathcal{H}_\Sigma(\Delta)$, and $H_1 g = H_2 g$ then $H_1 = H_2$.

Proof: An argument in [9] shows that if $H \in \mathcal{H}_\Sigma(1)$, then the conditions in Definition 6 on $\operatorname{supp} \eta_H$ imply that as long as $1/\sqrt{L} + 1/L < 1/(4\Sigma^2)$, then $\operatorname{supp} \eta_H$ is guaranteed to meet at most L rectangles $R_{q,m}$. Since L is now fixed, we can choose $c \in \mathbf{C}^L$ with the property that all square submatrices of $G(c)$ are invertible. This combined with (1) implies the result.

The conclusion of Theorem 5 is not sufficient by itself to allow the recovery from Hg of the spreading function of H. However, it is shown in [9] and in [2] that if $\Delta \le 1/2$, and $H \in \mathcal{H}_\Sigma(\Delta)$, the support set of H can be determined and H can be stably recovered from Hg. Heckel and Boelcskei go further in [2] and show that for almost every operator $H \in \mathcal{H}_\Sigma(\Delta)$ with $\Delta \le 1$, the support set of H can be determined and H can be stably recovered from Hg. Once the support set is known, explicit formulas for reconstructing the spreading function and impulse response of H from Hg are given in [9].

III. Efficient Operator Sampling

It is easy to see that any c satisfying the conclusion of Theorem 3 must have full support, that is, $\|c\|_0 = L$ where $\|c\|_0$ is the number of nonzero entries in c. However, for a given operator class, there is an advantage to choosing a c that has minimal support. First, having some or most of the c_k vanish would mean that the matrix $G(c)$ in (1) would be sparse, and hence the reduced matrix $G_0(c)$ that must be inverted to recover the spreading function would be sparse as well. In fact, the quantity $\|c\|_0/L$ is the fraction of nonzero entries in $G(c)$. Second, a vector c with small support would mean that the identifier g would require fewer deltas to be

transmitted per unit time. This is analogous to reducing the "sampling rate" in operator sampling. Third, note that

$$\operatorname{supp} Hg(x) \subseteq \bigcup_{y \in supp(g)} \operatorname{supp} \kappa(x,y)$$

and hence that if $\|c\|_0$ is small, in particular if in each period c vanishes but for a few contiguous indices, then in each time interval of length TL, Hg would have small support thereby reducing the amount of time spent measuring the channel.

A. Invertibility of Gabor Submatrices.

Definition 6. Let $G(c)$ be an $L \times L^2$ Gabor system matrix, and let $G_0(c)$ be an $L \times N$ submatrix of $G(c)$ corresponding to a collection of $N \leq L$ columns of $G(c)$. Define

$$\mu = \min\{\|c\|_0 \colon G_0(c) \text{ has full rank}\}.$$

We associate to $G_0(c)$ the L-tuple $\tau = (\tau_0, \tau_1, \ldots, \tau_{L-1})$, where τ_k is the number of columns of $G_0(c)$ chosen from the submatrix $D_k W_L$. The total number of columns chosen is given by $\|\tau\|_1$, the number of submatrices $D_k W_L$ from which any columns are chosen by $\|\tau\|_0$, and the largest number of columns chosen from any submatrix $D_k W_L$ by $\|\tau\|_\infty$.

Part (1) of the following theorem is proved in [9].

Theorem 7 Suppose that the L-vector τ describes a collection of columns chosen from a full Gabor matrix.

(1) If L is prime then $\mu \leq (\|\tau\|_1 - \|\tau\|_0) + 1$.
(2) For any $L \in \mathbf{N}$, $\mu \geq \|\tau\|_\infty$.

Proof: (1) Let L be prime, and assume that columns are chosen from $G(c)$ according to the vector τ. By definition, there will be at least one column chosen from $\|\tau\|_0$ distinct submatrices $D_k W_L$ of $G(c)$. This means that there are exactly $\|\tau\|_0$ distinct rows in which the variable c_0 formally appears. Choose those rows and the remaining $\|\tau\|_1 - \|\tau\|_0$ rows arbitrarily, and let M be the resulting $\|\tau\|_1 \times \|\tau\|_1$ submatrix. Proceeding now with the construction of the monomial p_M defined above, it follows that p_M will contain exactly $\|\tau\|_0$ factors of c_0 plus at most $\|\tau\|_1 - \|\tau\|_0$ other distinct factors. Hence p_M will be a monomial with at most $\|\tau\|_1 - \|\tau\|_0 + 1$ distinct variables appearing. Hence the variables not chosen can be set to zero and the polynomial $\det M$ will still not vanish identically. Hence there is a choice of c with $\|c\|_0 \leq \|\tau\|_1 - \|\tau\|_0 + 1$ for which $\det M \neq 0$, and the result follows.

(2) Let $L \in \mathbf{N}$ be given and suppose that columns are chosen from $G(c)$ according to the vector τ. Let $\|\tau\|_1$ rows be chosen from the submatrix $G_0(c)$, and call the resulting $\|\tau\|_1 \times \|\tau\|_1$ matrix M. Any diagonal of M must have τ_k entries chosen from τ_k distinct rows of each submatrix $D_k W_L$. Hence every term in the expansion of $\det(M)$ is a multiple of a monomial with at least τ_k distinct variables appearing in it. Therefore, if more than $\|\tau\|_\infty$ of the c_k are zero, then the polynomial $\det(M)$ will vanish identically. Hence $\mu \geq \|\tau\|_\infty$.

Remark (a) The bounds on μ in Theorem 7 cannot be improved. For example, if one column is chosen from distinct submatrices $D_k W_L$, then the vector τ will have $\|\tau\|_1$ nonzero entries each of which is 1 and . Hence $\|\tau\|_1 = \|\tau\|_0$, and $\|\tau\|_\infty = 1$. Letting $c_0 = 1$, $c_1 = c_2 = \cdots = c_{L-1} = 0$, and letting the rows of M be those of $G_0(c)$ in which c_0 appears gives

$$\mu = \|\tau\|_\infty = (\|\tau\|_1 - \|\tau\|_0) + 1.$$

If all $\|\tau\|_1$ columns are chosen from one submatrix $D_k W_L$, then $\|\tau\|_0 = 1$ and $\|\tau\|_1 = \|\tau\|_\infty$. If fewer than $\|\tau\|_1$ of the c_k are nonzero, then any choice of $\|\tau\|_1$ rows of $G_0(c)$ will contain at least one identically zero row. This means that

$$\mu \geq (\|\tau\|_1 - \|\tau\|_0) + 1 = \|\tau\|_1 = \|\tau\|_\infty.$$

Moreover, if L is prime we once again have equality ([7]).

(b) The following example will show that for arbitrarily large L there are vectors τ that avoid both extremes, that is, for any choice of submatrix $G_0(c)$, $\|\tau\|_\infty < \mu < \|\tau\|_1 - \|\tau\|_0 + 1$. More specifically, the following theorem holds.

Theorem 8 For every $L \in \mathbf{N}$ large enough, there is an L-vector τ describing a choice of columns of a full Gabor matrix $G(c)$ such that $\|\tau\|_\infty < \mu$. Moreover, if L is prime, then also $\mu < \|\tau\|_1 - \|\tau\|_0 + 1$.

Proof: In order to construct this vector τ, first choose P, $R \in \mathbf{N}$ such that $P \leq R$ and

$$\frac{R+P-1}{RP} < \frac{1}{2}.$$

Note that these imply that at least $R \geq P \geq 3$. Given $L \in \mathbf{N}$ with $L \geq 9$, we can write $L = PR + j$ uniquely for some $0 \leq j \leq R-1$. Define the L-vector τ as follows. Let $\tau_k = 2$ for $0 \leq k \leq R-1$, and for $k = mR-1$, $2 \leq m \leq P$, and let $\tau_k = 0$ otherwise. Then $\|\tau\|_0 = R + P - 1$, $\|\tau\|_\infty = 2$, and $\|\tau\|_1 = 2(R+P-1)$. We will show that $\|\tau\|_\infty = 2 < 3 \leq \mu$ and that in case L is also prime, $\mu \leq R < R + P = \|\tau\|_1 - \|\tau\|_0 + 1$. We describe the matrix $G_0(c)$ pictorially in the figure below. Each column in the figure that starts with $N-k$ represents two columns chosen from the submatrix $D_k W_L$. A generalized diagonal of the matrix $G_0(c)$ corresponds to the choice of two indices from each column and one from each row.

$$\begin{array}{ccccccc}
0 & N-1 & \ldots & N-R+1 & N-2R+1 & \ldots & N-PR+1 \\
1 & 0 & \ldots & N-R+2 & N-2R+2 & \ldots & N-PR+2 \\
2 & 1 & \ldots & N-R+3 & N-2R+3 & \ldots & N-PR+3 \\
3 & 2 & \ldots & N-R+4 & N-2R+4 & \ldots & N-PR+4 \\
\vdots & \vdots & & \vdots & \vdots & & \vdots \\
R-2 & R-3 & \ldots & N-1 & N-R-1 & \ldots & N-(P-1)R-1 \\
R-1 & R-2 & \ldots & 0 & N-1 & \ldots & N-(P-2)R-1 \\
\vdots & \vdots & & \vdots & \vdots & & \vdots \\
2R-1 & 2R-2 & \ldots & R & 0 & \ldots & N-(P-3)R-1 \\
\vdots & \vdots & & \vdots & \vdots & & \vdots \\
3R-1 & 3R-2 & \ldots & 2R & R & \ldots & N-(P-4)R-1 \\
\vdots & \vdots & & \vdots & \vdots & & \vdots \\
PR-1 & PR-2 & \ldots & (P-1)R & (P-2)R & \ldots & 0 \\
\vdots & \vdots & & \vdots & \vdots & & \vdots \\
N-1 & N-2 & \ldots & N-R & N-2R & \ldots & N-PR
\end{array},$$

In order to see the first inequality, let $G_0(c)$ be an $L \times 2(R+P-1)$ matrix described by τ. Specifically, we choose 2 columns from each submatrix $D_k W_L$ of $G(c)$ for all those

k for which $\tau_k = 2$. Now suppose that $\|c\|_0 = 2$ and assume without loss of generality that c_0 and c_{k_0} are the only non-zero entries of c. We will show that any choice of $2(R+P-1)$ rows of $G_0(c)$ will contain a zero row, which will imply that $\mu \geq 3$.

Note that each of the variables c_0 and c_{k_0} appears in at most $R+P-1$ rows of $G_0(c)$. Therefore, if a choice of $2(R+P-1)$ rows of $G_0(c)$ were not to contain a zero row, then we must be able to choose $R+P-1$ rows containing c_0 and an additional $R+P-1$ rows containing c_{k_0}. We will show that this is not possible by showing that there must be at least one row of $G_0(c)$ in which both c_0 and c_{k_0} appear. Specifically, we will show that all of the variables $c_1, c_2, \ldots, c_{L-1}$ appear at least once in the first R rows of $G_0(c)$. Clearly, c_0 also appears in each of these rows.

In the pair of columns of $G_0(c)$ chosen from the matrix $D_0 W_L$, the variables $c_1, \ldots, c_{R-1}$ appear in the first R rows. Given $1 \leq m \leq P$, consider the pair of columns of $G_0(c)$ chosen from the matrix $D_{mR-1} W_L$. It is not hard to see that in the first R rows of these columns, the variables $c_{(P-m)R+j+1}, \ldots, c_{P-(m-1))R+j}$ appear. Consequently, as m runs from 1 through P, all of the variables $c_{j+1}, \ldots, c_{PR+j}$ will appear in the first R rows of $G_0(c)$. This completes the first part of the proof.

Now suppose that L is prime. We will show that $\mu \leq R$ by showing that we can choose $2(R+P-1)$ rows of $G_0(c)$ in such a way that the monomial p_M of the resulting square matrix M, as described in the proof of Theorem 3, will have no more than R distinct variables c_j appearing in it.

First, choose the $R+P-1$ rows of $G_0(c)$ in which c_0 appears. For all $1 \leq m \leq P-1$, note that c_1 appears in row $mR+1$, c_2 appears in row $mR+2$ and in general, c_k appears in row $mR+k$ for $k = 1, 2, \ldots, R-1$. Note also that c_0 does not appear in these rows. Therefore, choose those $(P-1)(R-1)$ rows of $G_0(c)$. Note that $(R+P-1)+(P-1)(R-1) = RP > 2(R+P-1)$ by our assumption at the beginning of the proof. This means that by choosing rows in this way, and eliminating some if necessary, we arrive at a square sub-matrix M of $G_0(c)$. The corresponding monomial p_M will have $R+P-1$ factors of c_0 and at most $P-1$ factors of $c_1, c_2, \ldots, c_{R-1}$, resulting in no more than R distinct variables appearing in p_M. Hence $\mu \leq R < R+P = \|\tau\|_1 - \|\tau\|_0 + 1$.

Theorem 9. Let L prime be fixed, and let $N \leq L$. There exists a $c \in \mathbf{C}^L$ with $\|c\|_0 \leq N$ such that for any vector τ with $\|\tau\|_1 = N$ and every $L \times N$ matrix $G_0(c)$ with associated distribution vector τ has full rank. In fact, the collection of all such c considered as vectors in $\mathbf{C}^N$ constitutes a dense, open subset of $\mathbf{C}^N$.

Proof: By Theorem 7, for every vector τ with $\|\tau\|_1 = N$, there is a $c \in \mathbf{C}^L$ with the property that $\|c\|_0 \leq N$ and that $G_0(c)$ has full rank. We will first show that such a c can always be chosen such that $c_N = c_{N+1} = \cdots = c_{L-1} = 0$. To see this, consider a matrix $G_0(c)$, and set c_N through c_{L-1} to zero. In this case, every column of $G_0(c)$ will have N nonvanishing entries. We can therefore follow the algorithm outlined in the proof of Theorem 3 and observe that at each step in the algorithm, there will always be a row of the remaining matrix in which a variable c_j with $0 \leq j \leq N-1$ appears, for if not, this would imply that one of the columns of $G_0(c)$ had fewer than N nonvanishing entries. Choosing now these N rows, and letting M denote the resulting $N \times N$ submatrix of $G_0(c)$, it follows that in the monomial p_M, only variables c_j with $0 \leq j \leq N-1$ will appear and hence the polynomial $\det M$ will be a homogeneous polynomial of degree N in the variables $c_0, c_1, \ldots, c_{N-1}$.

Therefore, any choice of $c_0, c_1, \ldots, c_{N-1}$ that avoids the zero set of the polynomial $\det M$ will ensure that $G_0(c)$ has full rank. The set of such choices constitutes a dense open set in $\mathbf{C}^N$. Since there are only finitely many vectors τ with $\|\tau\|_1 = N$ and only finitely many associated $L \times N$ matrices $G_0(c)$, the collection of such c is the intersection of finitely many dense open subsets of $\mathbf{C}^N$. Since this is also a dense open set, the result follows.

IV. Implications for Operator Sampling

Theorem 10. Let $\Sigma > 0$, $0 \leq \Delta < 1$ be given. Then for every sufficiently large prime L, there is a $c \in \mathbf{C}^L$ with $\|c\|_0/L \leq \Delta$ such that the operator class $\mathcal{H}_\Sigma(\Delta)$ is identifiable by $g = \sum_n c_n \, \delta_{n/\sqrt{L}}$.

Proof. As in the proof of Theorem 5, we can choose a prime L sufficiently large that for any $H \in \mathcal{H}_\Sigma(\Delta)$, $\operatorname{supp} \eta_H$ meets at most ΔL rectangles $R_{q,m}$. For this L, we have seen that it is possible to choose $c \in \mathbf{C}^L$ such that $\|c\|_0 \leq \Delta L$ and such that any collection of no more than ΔL columns of the Gabor matrix $G(c)$ is linearly independent. Hence the operator H is completely determined by Hg where $g = \sum_n c_n \delta_{n/\sqrt{L}}$ and $\|c\|_0/L \leq \Delta$.

References

[1] P.A. Bello. Measurement of random time-variant linear channels. 15:469–475, 1969.

[2] R. Heckel and H. Boelcskei. Compressive identification of linear operators. In *Proc. IEEE Int. Symp.on Inf. Theory (ISIT)*, pages 1412–1416, St. Petersburg, Russia, 2011.

[3] T. Kailath. Sampling models for linear time-variant filters. Technical Report 352, Massachusetts Institute of Technology, Research Laboratory of Electronics, 1959.

[4] T. Kailath. Measurements on time–variant communication channels. 8(5):229–236, Sept. 1962.

[5] T. Kailath. Time–variant communication channels. *IEEE Trans. Inform. Theory: Inform. Theory. Progress Report 1960–1963*, pages 233–237, Oct. 1963.

[6] W. Kozek and G.E. Pfander. Identification of operators with bandlimited symbols. *SIAM J. Math. Anal.*, 37(3):867–888, 2006.

[7] J. Lawrence, G.E. Pfander, and D. Walnut. Linear independence of Gabor systems in finite dimensional vector spaces. *J. Fourier Anal. Appl.*, 11(6):715–726, 2005.

[8] G.E. Pfander and D. Walnut. Measurement of time–variant channels. *IEEE Trans. Inform. Theory*, 52(11):4808–4820.

[9] G.E. Pfander and D. Walnut. Sampling and reconstruction of operators. submitted, 2012.

Optimal wavelet reconstructions from Fourier samples via generalized sampling

Ben Adcock
Purdue University, USA
Email: adcock@purdue.edu

Anders Hansen
University of Cambridge, UK
Email: a.hansen@damtp.cam.ac.uk

Clarice Poon
University of Cambridge, UK
Email: cmhsp2@cam.ac.uk

***Abstract*—We consider the problem of computing wavelet coefficients of compactly supported functions from their Fourier samples. For this, we use the recently introduced framework of generalized sampling in the context of compactly supported orthonormal wavelet bases. Our first result demonstrates that using generalized sampling one obtains a stable and accurate reconstruction, provided the number of Fourier samples grows linearly in the number of wavelet coefficients recovered. We also present the exact constant of proportionality for the class of Daubechies wavelets.**

Our second result concerns the optimality of generalized sampling for this problem. Under some mild assumptions generalized sampling cannot be outperformed in terms of approximation quality by more than a constant factor. Moreover, for the class of so-called perfect methods, any attempt to lower the sampling ratio below a certain critical threshold necessarily results in exponential ill-conditioning. Thus generalized sampling provides a nearly-optimal solution to this problem.

I. Generalized sampling

A fundamental problem of signal processing is the reconstruction of signals from a discrete set of measurements. This can be formulated in a Hilbert Space $\mathcal{H}$ with inner product $\langle \cdot, \cdot \rangle$, where one seeks to reconstruct a function $f \in \mathcal{H}$ from measurements of the form $\langle f, s_j \rangle$ for some $\{s_j\}_{j\in\mathbb{N}} \subseteq \mathcal{S} \subseteq \mathcal{H}$. A key development is the Shannon-Nyquist Sampling Theorem, which stated that bandlimited or compactly supported signals to be fully described via measurements $\left\langle f, e^{2\pi i \epsilon j \cdot} \right\rangle$, $j \in \mathbb{Z}$, for some appropriate $\epsilon > 0$. In particular, f and its Fourier transform $\hat{f}(\cdot) = \int f(x) e^{-ix\cdot} \mathrm{d}x$ can be approximated respectively as follows:

$$f_N(t) = \epsilon \sum_{|k|\leq N} \hat{f}(2\pi k\epsilon) e^{2\pi i \epsilon k t}, \qquad f_N \xrightarrow{L^2} f,$$

$$\hat{f}_N(t) = \sum_{|k|\leq N} \hat{f}(2\pi k\epsilon) \mathrm{sinc}\left(\frac{t + 2\pi k\epsilon}{2\epsilon}\right), \quad \hat{f}_N \xrightarrow{L^2, L^\infty} \hat{f}.$$

However, in many cases, such approximations are not used because the bases generated by the sinc-function or complex exponentials are generally considered inappropriate representation systems for the underlying signals [1]. In fact, many images and signals can be better represented in terms of a different basis (e.g. splines or wavelets) than the basis in which they are sampled (e.g. the Fourier basis). Consequently, there is much interest in generalising the Shannon-Nyquist Sampling Theorem to recover the coefficients of a signal or image in a particular basis from samples taken with respect to another basis[1], this problem is often referred to as generalized sampling.

The goal now is to reconstruct in an arbitrary space $\mathcal{W} \subseteq \mathcal{H}$ without placing any constraints on the type of input vectors. In practice, we seek an approximation of f in the finite dimensional space $\mathcal{W}_N = \mathrm{span}\,\{w_j : 1 \leq j \leq N\}$ such that $\overline{\bigcup_{j\in\mathbb{N}} \mathcal{W}_j} = \mathcal{W}$ from some finite set of measurements $\hat{f}_M = (\langle f, s_j \rangle)_{j=1}^M$.

A. Desirable qualities of the reconstruction algorithm

We will be primarily be concerned with perfect reconstruction algorithms, where the underlying signals can be perfectly reconstructed from our discrete measurement sets. So, if $f \in \mathcal{W}$, then the algorithm should be able to recover f exactly from its measurements. Note that if $\mathcal{W} \cap \mathcal{S}^\perp \neq \{0\}$, then there will exist some non-zero vector $g \in \mathcal{W} \cap \mathcal{S}^\perp$ such that $\langle g, s_j \rangle = 0$ for all j. So g is indistinguishable from 0, regardless of the reconstruction algorithm. Thus, when considering the reconstruction problem, we will require that $\mathcal{W}$ and $\mathcal{S}$ satisfy the following condition:

$$\mathcal{W} \cap \mathcal{S}^\perp = \{0\}, \quad \mathcal{W} + \mathcal{S}^\perp \text{ is closed in } \mathcal{H} \tag{1}$$

and will refer to this as the subspace condition. Let us now consider the desirable qualities of a 'good' reconstruction method: Any reconstruction method should be such that the approximation will converge to the true signal as the number of samples increases and the method should be robust to small perturbations in the input data. With this in mind, we consider the following two definitions :

Definition I.1. *[2] Let $F_{N,M} : \mathcal{H} \to \mathcal{W}_N$. The quasi-optimality constant $\mu = \mu(F_{N,M})$ is the least constant such that*

$$\|f - F_{N,M}(f)\| \leq \mu \|f - P_{\mathcal{W}_N} f\|, \quad \forall f \in \mathcal{H},$$

If no such constant exists, we write $\mu = \infty$. We say that $F_{N,M}$ is quasi-optimal if $\mu(F_{N,M})$ is small.

Note that $P_{\mathcal{W}_N} f$ is the best approximation in norm to f from $\mathcal{W}_N$. So quasi-optimality means that the difference in norm between f and $F_{N,M}(f)$ is at most a constant factor μ of the difference between f and its best approximation in the subspace $\mathcal{W}_N$.

We also define the condition number of a reconstruction:

Definition I.2. *[2] Let $F_{N,M} : \mathcal{H} \to \mathcal{W}_N$ be a mapping such that, for each $f \in \mathcal{H}$, $F_{N,M}(f)$ depends only on the samples $\{\hat{f}_j\}_{j=1}^M$. The condition number of $\kappa(F_{N,M})$ is given by*

$$\kappa(F_{N,M}) = \sup_{f\in\mathcal{H}} \lim_{\epsilon\to 0^+} \sup_{\substack{g\in\mathcal{H} \\ 0<\|\hat{g}\|\leq\epsilon}} \frac{\|F_{N,M}(f+g) - F_{N,M}(f)\|}{\|\hat{g}\|},$$

where $\hat{g} = \{\hat{g}_j\}_{j=1}^M \in \mathbb{C}^M$. The mapping $F_{N,M}$ is well-conditioned if $\kappa(F_{N,M})$ is small and ill-conditioned otherwise.

We say that the reconstruction $F_{N,M}$ is 'good' if it is stable and quasi-optimal. In other words, if the *reconstruction constant*

$$C(F_{N,M}) = \max\{\kappa(F_{N,M}), \mu(F_{N,M})\},$$

is small.

II. Reduced consistency sampling

This problem of generalized sampling is not new and has been extensively studied - important contributions include the consistent sampling scheme introduced by Aldroubi and Unser [3], [4], [5] and significantly extended by Eldar [6], [7].

For $f \in \mathcal{H}$, one seeks an approximation $F_N(f) \in \mathcal{W}_N$ which agrees with the given measurements, so it is such that

$$\langle F_N(f), s_j \rangle = \langle f, s_j \rangle, \quad j = 1, \dots, N. \tag{2}$$

This involves solving a linear system of N equations and $F_N(f)$ exists uniquely when $\mathcal{W}_N \oplus \mathcal{S}_N^\perp = \mathcal{H}$. However, this condition need not hold even if $\mathcal{W} \oplus \mathcal{S}^\perp = \mathcal{H}$ and there are important cases for which (2) has no solution, or the method F_N is unstable or nonconvergent, i.e. $\kappa(F_N) \to \infty$ or $F_N(f) \not\to f$ as $N \to \infty$ [8], [5].

To circumvent these problems, various authors have considered overdetermined systems, where the number of measurements exceeds the number of reconstruction coefficients to be recovered. We in particular mention the work of Pruessmann et al [9] in the recovery of voxel coefficients (which may be considered as Haar wavelet coefficients) from Fourier samples and [10] by Hrycak and Gröchenig for the recovery of polynomial coefficients from Fourier samples. To formalise these approaches, Adcock and Hansen introduced the reduced consistency sampling scheme [8], [11]. The task is then as follows: Given $N \in \mathbb{N}$, for some appropriate $M \in \mathbb{N}$, find $F_{N,M}(f) \in \mathcal{W}_N$ such that

$$\langle P_{\mathcal{S}_M} F_{N,M}(f), w_j \rangle = \langle P_{\mathcal{S}_M} f, w_j \rangle, j = 1, \dots, N. \tag{3}$$

So, $F_{N,M}(f)$ coincides with f on $P_{\mathcal{S}_M}(\mathcal{W}_N)$ rather than on $\mathcal{S}_M$. Under this framework, a stable and convergent scheme can always be devised. Indeed, for all $N \in \mathbb{N}$, there exists m_0 such that for all $M \geq m_0$, there exists a unique $F_{N,M}(f)$ satisfying (3), and such a solution is quasi-optimal in $\mathcal{W}_N$ and stable with reconstruction constant at most

$$D_{N,M} = \left(\inf_{g \in \mathcal{W}_N} \|P_{\mathcal{S}_M} g\| \right)^{-1}.$$

As both convergence and numerical stability are governed by the quantity $D_{N,M}$, the notion of a *stable sampling rate* was introduced:

Definition II.1. *[2] For $N \in \mathbb{N}$ and $\theta \in (1, \infty)$, the stable sampling rate is given by*

$$\Theta(N; \theta) = \min \{ M \in \mathbb{N} : D_{N,M} \leq \theta \}.$$

As demonstrated in [2], for any $N \in \mathbb{N}$, $\Theta(N; \theta)$ can be numerically calculated and determines the number of samples required to obtain a convergent and stable reconstructions in $\mathcal{W}_N$ as $N \to \infty$.

III. Optimality of generalized sampling

In [2], the reduced consistency scheme is shown to be optimal amongst all perfect methods, in that it is not possible improve upon its stability. The following result shows that the stable sampling rate is a universal property amongst perfect methods, since any perfect method must sample at a rate at least that of the stable sampling rate to achieve the same stability.

Theorem III.1. *[2] For $M \geq N$ let $G_{N,M} : \mathcal{H} \to \mathcal{W}_N$ be a perfect reconstruction method such that, for each $f \in \mathcal{H}$, $G_{N,M}(f)$ depends only on the samples $\{\hat{f}_j\}_{j=1}^M$. Then the condition number is such that $\kappa(G_{N,M}) \geq \kappa(F_{N,M})$, where $F_{N,M}$ is the generalized sampling reconstruction.*

For nonperfect methods, the following result holds:

Theorem III.2. *[2] Suppose that the stable sampling rate $\Theta(N; \theta)$ is linear in N for a particular sampling and reconstruction problem. Let $f \in \mathcal{H}$ be fixed, and suppose that there exists a sequence of mappings*

$$G_M : \{\hat{f}_j\}_{j=1}^M \mapsto G_M(f) \in \mathcal{W}_{\Psi_f(M)},$$

where $\Psi_f : \mathbb{N} \to \mathbb{N}$ with $\Psi_f(M) \leq cM$. Suppose also that there exist constants $c_1(f), c_2(f), \alpha_f > 0$ such that

$$c_1(f) N^{-\alpha_f} \leq \|f - P_{\mathcal{W}_N} f\| \leq c_2(f) N^{-\alpha_f}, \quad \forall N \in \mathbb{N}. \tag{4}$$

Then, given $\theta \in (1, \infty)$, there exist constants $c(\theta) \in (0, 1)$ and $c_f(\theta) > 0$ such that

$$\|f - F_{c(\theta)M,M}(f)\| \leq c_f(\theta) \|f - G_M(f)\|, \quad \forall M \in \mathbb{N}, \tag{5}$$

where $F_{N,M}$ is the generalized sampling reconstruction.

Thus, for problems with linear stable sampling rates, even if one is allowed to design a method that depends on f in a completely non-trivial way, it is still not possible to obtain a faster asymptotic rate of convergence than that of generalized sampling. In fact, we will show that the stable sampling rate is linear for wavelets, making this theorem directly applicable.

IV. Wavelet reconstructions from Fourier samples

Any implementation of the reduced consistency sampling scheme requires an understanding of the corresponding stable sampling rate. The case where the reconstruction space $\mathcal{W}$ is generated by compactly supported wavelets and the sampling space is the space of complex exponentials $\mathcal{S} = \overline{\text{span}} \{ e^{2\pi i \epsilon j \cdot} : j \in \mathbb{Z} \}$ for some appropriate $\epsilon > 0$ is particularly important, with applications in medical imaging. In this section, we present some results which show that the stable sampling rate is linear in this setting. We first describe the construction of the reconstruction and sampling spaces.

For the reconstruction space, we aim to create orthonormal subsets $\{\varphi_k\}_{k \in \mathbb{N}} \subseteq L^2(\mathbb{R})$ with the property that $L^2[0, a] \subseteq \overline{\text{span}} \{\varphi_k\}_{k \in \mathbb{N}}$ for some $a > 0$. Suppose that we are given an orthonormal mother wavelet ψ and an orthonormal scaling function ϕ such that $\text{supp}(\psi) = \text{supp}(\phi) = [0, a]$ for some $a \geq 1$.

The standard approach is to consider the following collection of functions

$$\begin{aligned} \Omega_a = \{ & \phi_k, \psi_{j,k} : \text{supp}(\phi_k)^o \cap [0, a] \neq \emptyset, \\ & \text{supp}(\psi_{j,k})^o \cap [0, a] \neq \emptyset, j \in \mathbb{Z}_+, k \in \mathbb{Z}, \}, \end{aligned}$$

where

$$\phi_k = \phi(\cdot - k), \qquad \psi_{j,k} = 2^{\frac{j}{2}} \psi(2^j \cdot - k).$$

(the notation K^o denotes the interior of a set $K \subseteq \mathbb{R}$). This now gives

$$L^2[0, a] \subseteq \text{cl}(\text{span}\{\varphi : \varphi \in \Omega_a\}) = \mathcal{W} \subseteq L^2[-T_1, T_2],$$

where $T_1 = \lceil a \rceil - 1$ and $T_2 = 2\lceil a \rceil - 1$ are such that $[-T_1, T_2]$ contains the support of all functions in Ω_a.

For the Fourier sampling space, we let $\epsilon \leq 1/(T_1 + T_2)$ be the *sampling density*. Note that $1/(T_1 + T_2)$ is the corresponding Nyquist criterion for functions supported on $[-T_1, T_2]$. We now define the sampling vectors by

$$s_l = \sqrt{\epsilon} e^{2\pi i l \epsilon \cdot} \chi_{[-T_1/(\epsilon(T_1+T_2)), T_2/(\epsilon(T_1+T_2))]},$$

and the sampling space by

$$\mathcal{S} = \overline{\text{span}\{s_l : l \in \mathbb{Z}\}} = \left\{ f \in L^2(\mathbb{R}) : \text{supp}(f) \subseteq \left[-\frac{T_1}{\epsilon(T_1+T_2)}, \frac{T_2}{\epsilon(T_1+T_2)} \right] \right\}$$

and the space spanned by the first M sampling vectors by

$$\mathcal{S}_M = \text{span}\left\{ s_l : -\left\lfloor \frac{M}{2} \right\rfloor \leq l \leq \left\lceil \frac{M}{2} \right\rceil - 1 \right\}.$$

Our main result on the stable sampling rate is as follows.

Theorem IV.1. *[12] For $R \in \mathbb{N}$, let N_R denote the number of elements in Ω_a of the form $\phi_{j,k}$ or $\psi_{j,k}$ with $j < R$, in particular, $N_R = 2^R\lceil a \rceil + (R+1)(\lceil a \rceil - 1)$. Then for $N \leq N_R$ and all $\theta \in (1,\infty)$, there exists $S_\theta \in \mathbb{N}$, independent of R, such that for $M = \left\lceil \frac{S_\theta 2^{R+1}}{\epsilon} \right\rceil$, we have $D_{NM} \leq \theta$. Hence, $\Theta(N,\theta) = \mathcal{O}(N)$ for any $\theta \in (1,\infty)$.*

So, the stable sampling rate is linear and in other words, given any $f \in \mathcal{H}$, for any $N \in \mathbb{N}$ and $\theta \in (1,\infty)$, there exists a constant r such that $r \cdot N$ samples will up to a factor of θ, yield the best possible approximation in the space $\mathcal{W}_N$ and the condition number of the method is no worse than θ as $N \to \infty$.

One may ask, how small can the ratio r be? The next result shows that there is a critical ratio, below which, the reconstruction will become exponentially ill posed.

Theorem IV.2. *[12] Let $F_{N,M}$ denote the reduced consistency sampling method and N_R be as in Theorem IV.1. Let $N = N_R$ and $M = c \cdot 2^R$, with $c < \epsilon^{-1}$. Then $\kappa(F_{N,M}) \to \infty$ exponentially as $N \to \infty$.*

The first consequence of this with regards to optimality is that this critical ratio is universal amongst perfect methods. It is not the case that a perfect method could reconstruct in $\mathcal{W}_{N_R}$ from less than $2^R/\epsilon$ samples and still only experience mild growth in its condition number - this method will inherently become exponentially ill posed. The second consequence for optimality is as explained at the end of Section III, any non-perfect method which has a lower sampling ratio for a particular function f satisfying (4) can only outperform generalized sampling by a constant factor.

A. Daubechies wavelets

Our next result examines the special case of Daubechies wavelets and asymptotically, the stable sampling ratio can be determined exactly.

Theorem IV.3. *[12] Let $\mathcal{W}$ be generated by a Daubechies wavelet, and recall N_R from Theorem IV.1. Then, there exists $\theta \in (1,\infty)$ and $R_0 \in \mathbb{N}$ such that for all $R \geq R_0$, $\Theta(N_R,\theta) = \lceil 2^R/\epsilon \rceil$. In particular, when $1/\epsilon \in \mathbb{Z}$ it suffices to let $\theta > \left(\inf_{\xi \in [-\pi,\pi]} \left| \hat{\phi}(\xi) \right| \right)^{-1}$. Moreover, in addition to this, for Haar wavelets, where $a = 1$, we have that $\Theta(N_R,\theta) \leq \lceil 2^R/\epsilon \rceil$ for all $R \in \mathbb{N}$.*

V. Numerical examples

In this section, we provide numerical simulations of three key ideas for generalized sampling in the context of wavelet reconstructions from Fourier samples. Firstly, generalized sampling can offer substantial improvements. Secondly, the stable sampling rate is linear for wavelet reconstructions from Fourier samples, moreover, our result for the Daubechies wavelet case is sharp. Finally, understanding of the stable sampling rate is crucial to the implementation of reduced consistency sampling and violation of it could lead to disastrous results.

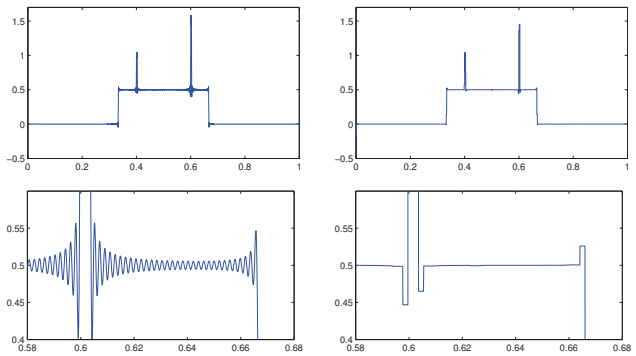


Fig. 1. The top row shows f_M (left) and $f^{[N,M]}$ (right). The bottom row shows f_M (left) and $f^{[N,M]}$ (right) on the interval $[0.58, 0.68]$.

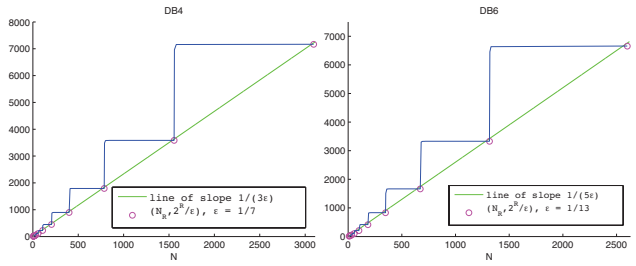


Fig. 2. The figure displays the stable sampling rate $\Theta(N,\theta_1)$ and $\Theta(N,\theta_2)$ in blue for the Daubechies-4 wavelet (left) and the Daubechies-6 wavelet (right) with Fourier samples at a sampling distance $\epsilon = 1/7$ and $\epsilon = 1/13$ respectively.

A. Signal recovery via generalized sampling

We consider the reconstruction of the following function

$$f = \frac{1}{2}\chi_{[1/3,2/3]} + \frac{1}{2}\chi_{[2/5,2/5+1/300]} + \chi_{[3/5,3/5+1/300]},$$

from $M = 1024$ Fourier samples of sampling density $\epsilon = 1/2$. Figure 1 shows the truncated Fourier series representation f_M as presented in the S-N Sampling Theorem as well as the reconstruction $f^{[N,M]}$ from implementing generalized sampling for a Haar wavelet reconstruction space. In this case, N is chosen to be 512. It is clear that $f^{[N,M]}$ is visually preferable to f_M with less oscillations at discontinuities. We remark that similar figures were generated in [13] to justify the use of wavelet encoding for MRI, which modifies an MR scanner to direct acquire wavelet coefficients rather than Fourier samples. In proving that the stable sampling rate is linear, we show that that wavelet coefficients can be accurately approximated via a post-processing and there is little to be gained in modifying the sampling process.

B. Sharpness of Theorem IV.3

To demonstrate the sharpness of this result, we consider the Daubechies-4 wavelet (supported in $[0,3]$), and the Daubechies-6 wavelet (supported in $[0,5]$). The graphs of Figure 2 plots the stable sampling rate $\Theta(N,\theta)$ against N, the number of reconstruction vectors to be recovered. In each case, we set $\theta > \left(\inf_{\xi \in [-\pi,\pi]} \left| \hat{\phi}(\xi) \right| \right)^{-1}$. Note that at the points N_R, $\Theta(N_R,\theta) = 2^R/\epsilon$ as predicted by Theorem IV.3.

M	$\|f - \tilde{f}_{M/c,M}\|_{L^2}$	$\|f - \tilde{f}_{M/c_1,M}\|_{L^2}$	Noise Level
482	7.3×10^{-7}	2.8×10^{-2}	0
934	1.4×10^{-7}	5.4×10^{-2}	0
1834	2.6×10^{-8}	1.4×10^{-2}	0
482)	9.6×10^{-6}	6.1	1.0×10^{-5}
934	9.5×10^{-6}	77.2	1.0×10^{-5}
1834	9.7×10^{-6}	85.9	1.0×10^{-5}

TABLE I
THE TABLE SHOWS THE ERROR OF THE GENERALIZED SAMPLING RECONSTRUCTIONS $\tilde{f}_{N,M}$ WITH $N = M/c$ AND $N = M/c_1$, WITH NOISELESS AND NOISY DATA.

Observe also from Theorem IV.3 that

$$\Theta(N_R, \theta) < \Theta(N, \theta) \leq \Theta(N_{R+1}, \theta), \quad N_R < N \leq N_{R+1}.$$

The staircase effect witnessed in the figure suggests that the upper bound is in fact an equality. Hence, although the stable sampling rate is linear for all N, from the point of view of the stable sampling rate at least, there is nothing to be gained from allowing $N \neq N_R$.

C. Importance of the stable sampling rate

We demonstrate, as predicted by Theorem IV.2, that failure of satisfying the stable sampling rate gives a completely unstable and non-convergent reconstruction. We compare the choices

$$M = cN, \quad c = \frac{1}{\epsilon \lceil a \rceil}, \qquad M = c_1 N, \quad c_1 = 0.95c.$$

for the recovery of the function $f = \sum_{j=1}^{3\times 10^3} j^{-3}\varphi_j$, where φ_j are Daubechies-4 wavelets. We will consider Fourier samples $\langle f, s_j\rangle$ for $|j| \leq M/2$ which are contaminated with noise and thus we observe $\xi = \{\langle f, s_1\rangle, \ldots, \langle f, s_M\rangle\} + v$ with $\|v\| = \varepsilon$ for some noise level $\varepsilon \geq 0$. As verified in Table I the latter choice of $M = c_1 N$ gives disastrous results as an incorrect choice of the sampling ratio causes the condition number of the algorithm to blow up exponentially.

VI. EXTENSION TO OTHER MRA WAVELET BASES

Although the theorems presented in the previous sections have been for orthonormal systems of MRA wavelets, the key property required for the proofs is the existence of an increasing sequence

$$0 < N_1 < \cdots < N_R < N_{R+1} < \cdots$$

such that $N_R = \mathcal{O}(2^R)$, $\overline{\bigcup_{R\in\mathbb{N}} \mathcal{W}_{N_R}} = \mathcal{W}$ and

$$\begin{aligned} &\mathcal{W}_{N_R} \subseteq \operatorname{span}\{\phi_{R,j} : A_{R,1} \leq j \leq A_{R,2}\}, \\ &A_{R,2} - A_{R,1} = \mathcal{O}(2^R). \end{aligned} \tag{6}$$

Consequently, the results of this paper can be readily extended to other compactly supported MRA wavelets such as the Semi-orthogonal spline wavelets of [14], [15] or the bi-orthogonal Cohen-Daubechies-Feauveau wavelets of [16]. We also remark that the construction of the wavelet reconstruction space in Section IV is the standard construction of wavelets on an interval with zero-padding which can lead to large wavelet coefficients at the end points of the interval. However, there are more sophisticated constructions of wavelets on the intervals to reduce this effect, such as the basis of Daubechies wavelets with special boundary wavelet and scaling functions as described in [17]. Their construction is such that the number of vanishing moments is preserved and the boundary scaling function can be written as a linear combination of finitely many elements in $\{\phi(\cdot - k) : k \in \mathbb{Z}\}$. Such a wavelet basis will also satisfy the requirements of (6) and the associated stable sampling rate is also linear. In combination with known results [18] about the characterization of the Sobolev space $W^s[0,1]$, $s > 0$ via the decay of wavelet coefficients from interval wavelets with $q > s$ vanishing moments, we have the following result.

Theorem VI.1. *Let $\mathcal{W}$ be the reconstructed space constructed from the Daubechies wavelet of q vanishing moments on the unit interval and let $\mathcal{S}$ be the Fourier sampling space with sampling density $\epsilon \leq 1$. Then, for any $\theta \in (1, \infty)$, the stable sampling rate $\Theta(N, \theta)$ is linear in N. Furthermore, given any $f \in W^s[0,1]$ with $s \in (0, q)$, the generalized sampling solution $F^{[N,M]}(f)$ implemented with $M = \Theta(N, \theta)$ samples satisfies*

$$\left\| f - F^{[N,M]}(f) \right\| = \mathcal{O}(M^{-s}).$$

Thus, another consequence of a linear stable sampling rate is as follows: given M Fourier samples of any $f \in W^s[0,1]$, it is well known that the Fourier representation cannot yield a convergence rate of $\mathcal{O}(M^{-s})$. However, this convergence rate can be attained from exactly these M Fourier measurements by reconstructing in an appropriate wavelet basis via generalized sampling.

REFERENCES

[1] M. Unser, "Sampling–50 years after Shannon," *Proc. IEEE*, vol. 88, no. 4, pp. 569–587, 2000.

[2] B. Adcock, A. C. Hansen, and C. Poon, "Beyond consistent reconstructions: optimality and sharp bounds for generalized sampling, and application to the uniform resampling problem," *Preprint*, 2012.

[3] M. Unser and A. Aldroubi, "A general sampling theory for nonideal acquisition devices," *IEEE Trans. Signal Process.*, vol. 42, no. 11, pp. 2915–2925, 1994.

[4] A. Aldroubi, "Oblique projections in atomic spaces," *Proceedings of the American Mathematical Society*, vol. 124, no. 7, pp. 2051–2060, 1996.

[5] A. Hirabayashi and M. Unser, "Consistent sampling and signal recovery," *IEEE Trans. Signal Process.*, vol. 55, no. 8, pp. 4104–4115, 2007.

[6] Y. C. Eldar, "Sampling with arbitrary sampling and reconstruction spaces and oblique dual frame vectors," *J. Fourier Anal. Appl.*, vol. 9, no. 1, pp. 77–96, 2003.

[7] ——, "Sampling without input constraints: Consistent reconstruction in arbitrary spaces," in *Sampling, Wavelets and Tomography*, A. I. Zayed and J. J. Benedetto, Eds. Boston, MA: Birkhäuser, 2004, pp. 33–60.

[8] B. Adcock and A. C. Hansen, "A generalized sampling theorem for stable reconstructions in arbitrary bases," *J. Fourier Anal. Appl.*, vol. 18, no. 4, pp. 685–716, 2012.

[9] K. P. Pruessmann, M. Weiger, M. B. Scheidegger, P. Boesiger *et al.*, "Sense: sensitivity encoding for fast MRI," *Magnetic Resonance in Medicine*, vol. 42, no. 5, pp. 952–962, 1999.

[10] T. Hrycak and K. Gröchenig, "Pseudospectral Fourier reconstruction with the modified inverse polynomial reconstruction method," *J. Comput. Phys.*, vol. 229, no. 3, pp. 933–946, 2010.

[11] B. Adcock and A. C. Hansen, "Stable reconstructions in Hilbert spaces and the resolution of the Gibbs phenomenon," *Appl. Comput. Harmon. Anal.*, vol. 32, no. 3, pp. 357–388, 2012.

[12] B. Adcock, A. C. Hansen, and C. Poon, "On optimal wavelet reconstructions from Fourier samples: linearity and universality of the stable sampling rate," *Preprint*, 2012.

[13] J. B. Weaver, Y. Xu, D. M. Healy, and J. R. Driscoll, "Wavelet-encoded MR imaging," *Magn. Reson. Med.*, vol. 24, pp. 275–287, 1992.

[14] C. Chui and J. Wang, "On compactly supported spline wavelets and a duality principle," *Trans. Amer. Math. Soc*, vol. 330, no. 2, pp. 903–915, 1992.

[15] M. Unser, A. Aldroubi, and M. Eden, "On the asymptotic convergence of B-spline wavelets to Gabor functions," *Information Theory, IEEE Transactions on*, vol. 38, no. 2, pp. 864–872, 1992.

[16] A. Cohen, I. Daubechies, and J. Feauveau, "Biorthogonal bases of compactly supported wavelets," *Communications on pure and applied mathematics*, vol. 45, no. 5, pp. 485–560, 2006.

[17] A. Cohen, I. Daubechies, and P. Vial, "Wavelets on the interval and fast wavelet transforms," *Applied and Computational Harmonic Analysis*, vol. 1, no. 1, pp. 54–81, 1993.

[18] S. Mallat, *A wavelet tour of signal processing*. Academic press, 1999.

Wavelet Signs: A New Tool for Signal Analysis

Martin Storath, Laurent Demaret, and Peter Massopust
Helmholtz Zentrum München and Technische Universität München, Germany
Email: {martin.storath, laurent.demaret, peter.massopust}@helmholtz-muenchen.de

***Abstract*—We propose a new analysis tool for signals, called signature, that is based on complex wavelet signs. The complex-valued signature of a signal at some spatial location is defined as the fine-scale limit of the signs of its complex wavelet coefficients. We show that the signature equals zero at sufficiently regular points of a signal whereas at salient features, such as jumps or cusps, it is non-zero. We establish that signature is invariant under fractional differentiation and rotates in the complex plane under fractional Hilbert transforms. We derive an appropriate discretization, which shows that wavelet signatures can be computed explicitly. This allows an immediate application to signal analysis.**

I. Introduction

The determination and classification of salient features, such as jumps or cusps, is an important task in signal processing. Classical approaches assume the interesting features of a signal to be points of low regularity. In this context, local regularity is measured in terms of the (fractional) differentiability order, e.g., in the sense of local Hölder, Sobolev or Besov regularity. Since such measures of smoothness only rely on the *modulus* of wavelet coefficients [5], [9], they do not take into account wavelet sign (or phase) information.

We may observe the shortcomings of a purely modulus based approach by considering the two functions $f(x) = \operatorname{sgn} x$ and $g(x) = 2\log|x|$. Since f and g are related by the Hilbert transform, their wavelet coefficients are equal with respect to the order of magnitude. Hence, the locally symmetric singularity of f and the locally antisymmetric singularity of g at the origin cannot be distinguished using a purely modulus-based signal analysis.

We present a new signal analysis tool, which exclusively uses the (complex) *sign* of the wavelet coefficients. To this end, we investigate the fine scale limits of the signs of the wavelet coefficients

$$\sigma f(b) := \lim_{a\to 0} \operatorname{sgn} \langle f, \kappa_{a,b}\rangle := \lim_{a\to 0} \frac{\langle f, \kappa_{a,b}\rangle}{|\langle f, \kappa_{a,b}\rangle|}, \qquad (1)$$

where κ is a complex-valued wavelet, $a > 0$ the scale, and $b \in \mathbb{R}$ the location. The *complex-valued* quantity $\sigma f(b)$ is called the *signature of f at location b*. We shall see that the signature allows the local analysis of isolated salient features. Hereby, the orientation of the signature within the complex plane may be interpreted as an indicator of local symmetry or antisymmetry. In particular, we show that the signature is purely imaginary at a jump, whereas it is purely real at a cusp. Moreover, the signature is invariant under fractional Laplacians, i.e.,

$$\sigma((-\Delta)^{\frac{r}{2}} f) = \sigma f,$$

and it serves as a multiplier when acting on the fractional Hilbert transform, in the sense that

$$\sigma(\mathcal{H}^\alpha f) = e^{i\alpha\frac{\pi}{2}}\sigma f.$$

Therefore, the signature may be interpreted as being "dual" to the local Sobolev regularity index, which is invariant under fractional Hilbert transforms but shifts under fractional Laplacians. We also establish that

$$\operatorname{sing\,supp} f \not\subset \operatorname{supp} \sigma f \quad \text{and} \quad \operatorname{supp} \sigma f \not\subset \operatorname{sing\,supp} f. \quad (2)$$

Thus, a singularity in the classical sense need not coincide with a signature-type singularity.

We further introduce a method to numerically compute the signature of digital or sampled signals and validate the theoretically developed concepts by numerical experiments. There are some connections between our discretization and *phase congruency* [6]. However, the approach undertaken in [6] tends to favor unwanted large coefficients, which our method avoids.

In this short communication, we omit the proofs which the interested reader may find in [11].

II. Definitions and basic properties

We define the Fourier transform of a Schwartz function $f \in \mathscr{S}(\mathbb{R};\mathbb{C})$ by

$$\mathscr{F}(f)(\omega) := \widehat{f}(\omega) := \int_{\mathbb{R}} e^{-i\omega x} f(x) dx.$$

Likewise, we use the above notation for the usual extension to the space of tempered distributions $\mathscr{S}'(\mathbb{R};\mathbb{C})$. Furthermore, $\mathscr{F}^{-1}(f)$ and $f^\vee$ denote the corresponding inverse Fourier transform of f. Let us introduce the class of complex wavelets we need for the definition of signature.

Definition 1. *We call a complex-valued non-zero function $\kappa \in \mathscr{S}(\mathbb{R};\mathbb{C})$ a* signature wavelet *if κ has a one-sided compactly supported Fourier transform, i.e.,*

$$\operatorname{supp}\widehat{\kappa} \subseteq [c,d], \quad 0 < c < d < \infty, \qquad (3)$$

and a non-negative frequency spectrum, i.e.,

$$\widehat{\kappa}(\omega) \geq 0, \quad \textit{for all } \omega \in \mathbb{R}. \qquad (4)$$

The wavelet system associated with a signature wavelet κ is defined as the family of functions

$$\kappa_{a,b}(x) := \frac{1}{\sqrt{a}}\,\kappa\left(\frac{x-b}{a}\right), \qquad \text{where } a > 0 \text{ and } b \in \mathbb{R}.$$

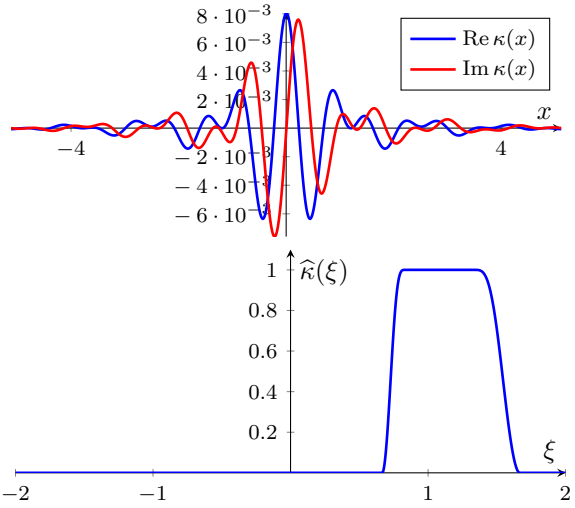


Fig. 1. The Meyer-type signature wavelet κ (left) and its Fourier transform $\widehat{\kappa}$ (right).

An example of a signature wavelet is given by the inverse Fourier transform of the (one-sided) Meyer window W, i.e.,

$$\kappa(x) = \mathscr{F}^{-1}(W)(x), \tag{5}$$

where W is a Meyer window function (see Figure 1). We refer to [11] for the definition of W.

Recall that the sign of a complex number $z \in \mathbb{C}$ is given by

$$\operatorname{sgn} z = \begin{cases} \frac{z}{|z|}, & \text{if } z \neq 0, \\ 0, & \text{if } z = 0. \end{cases}$$

The signature of a signal is then defined as follows.

Definition 2. *Let $f \in \mathscr{S}'(\mathbb{R};\mathbb{R})$. If there exists a $z \in \mathbb{C}$, such that for all signature wavelets κ,*

$$\lim_{a \to 0} \operatorname{sgn} \langle f, \kappa_{a,b} \rangle = z,$$

then we define the signature, σf, of f at $b \in \mathbb{R}$ by $\sigma f(b) := z$; otherwise, we set $\sigma f(b) := 0$.

Note that the signature $\sigma f(b)$ is either equal to zero or is a complex number of modulus 1. It follows directly from the definition that the signature is invariant under translations, i.e.,

$$\sigma(T_r f)(b) = (\sigma f)(b - r) \tag{6}$$

and under dilations, i.e.,

$$\sigma(D_\nu f)(b) = (\sigma f)(\nu b). \tag{7}$$

Here, the operator of translation by $r \in \mathbb{R}$, T_r, and dilation by $\nu \in \mathbb{R} \setminus \{0\}$, D_ν, are defined by

$$T_r f(x) := f(x - r) \quad \text{and} \quad D_\nu f(x) := \frac{1}{\sqrt{\nu}} f\left(\frac{x}{\nu}\right),$$

respectively.

Since the Fourier transform of a signature wavelet κ vanishes in a neighborhood of the origin, we have that

$$\langle p, \kappa \rangle = 0, \quad \text{for any polynomial } p. \tag{8}$$

Therefore, the signature is well defined on the space of tempered distributions modulo polynomials $\mathscr{S}'/\mathscr{P}$, where $\mathscr{P}$ denotes the space of all polynomials.

Our first result shows that a signal of polynomial growth has signature equal to zero at a point where all derivatives are equal to zero.

Theorem 3. *Let f be a real-valued, locally integrable function of polynomial growth. Further assume that f is smooth in a neighborhood of $b \in \mathbb{R}$. If $f^{(k)}(b) = 0$, for all $k \in \mathbb{N}_0$, then $\sigma f(b) = 0$. In particular, $\operatorname{supp} \sigma f \subseteq \operatorname{supp} f$.*

An interesting consequence of Theorem 3 is the case when f is locally a polynomial.

Corollary 4. *Let f be a real-valued, locally integrable function of polynomial growth which is smooth in a neighborhood of $b \in \mathbb{R}$. If for some $k_0 \in \mathbb{N}_0$, $f^{(k)}(b) = 0$, for all $k \geq k_0$, then $\sigma f(b) = 0$. In particular, if f coincides on an open set $U \subset \mathbb{R}$ with a polynomial then $\sigma f(b) = 0$, for every $b \in U$.*

In the following example, we consider the unit step function. Here, we can compute the signature at $b = 0$ explicitly. For $b \neq 0$, we can apply Corollary 4.

Example 5. Let U be the unit step function defined by

$$U(x) := \begin{cases} 1, & \text{if } x \geq 0, \\ 0, & \text{else.} \end{cases}$$

For any signature wavelet κ, we have that

$$\langle U, \kappa_{a,0} \rangle = \left\langle \widehat{U}, (\kappa_{a,0})^\vee \right\rangle = \frac{i\sqrt{a}}{\pi} \int_{\mathbb{R}} \frac{\widehat{\kappa}(a\xi)}{\xi} \, \mathrm{d}\xi. \tag{9}$$

Hence, since $\widehat{\kappa} \geq 0$ and $\operatorname{supp} \widehat{\kappa} \subset [0, \infty)$, we obtain that $\operatorname{sgn} \langle U, \kappa_{a,0} \rangle = i$, for all $a > 0$. For $b \neq 0$, we apply Corollary 4 yielding

$$\sigma U(b) = \begin{cases} i, & \text{if } b = 0, \\ 0, & \text{else.} \end{cases}$$

In our next example, we turn our attention to the signature of a pure cusp-type singularity.

Example 6. For a fixed x_0, consider the function

$$f(x) = |x - x_0|^\gamma, \qquad \text{where } \gamma > 0.$$

In [11] we proved that the wavelet signs are given by

$$\sigma f(x_0) = \begin{cases} 0, & \text{if } \gamma \in 2\mathbb{N}_0, \\ -1, & \text{if } \gamma \in]0,2[\cup]4,6[\cup \ldots, \\ +1, & \text{if } \gamma \in]2,4[\cup]6,8[\cup \ldots. \end{cases}$$

and $\sigma f(b) = 0$, for $b \neq x_0$.

Next we show that, in general, a jump discontinuity induces a purely imaginary signature at the jump location. A function f has a *jump (or step) discontinuity* at b if the left-hand and the right-hand limits $f(b+)$ and $f(b-)$ exist but are not equal.

Theorem 7. *Let f be a real-valued, locally integrable function of polynomial growth and let $b \in \mathbb{R}$. If there exists a neighborhood U of b such that f is continuous on $U \setminus \{b\}$ and has a jump discontinuity at b, then*

$$\sigma f(b) = \begin{cases} +i, & \text{if} \quad f(b-) < f(b+), \\ -i, & \text{if} \quad f(b-) > f(b+). \end{cases} \tag{10}$$

III. Fractional Laplacians and fractional Hilbert transforms

We now investigate the behavior of signature under the action of fractional powers of the Laplacian and the fractional Hilbert transform. We shall see that the former leaves the signature invariant whereas the latter acts on the signature by a rotation in the complex plane.

We recall that fractional powers of the Laplacian $(-\Delta)^{\frac{r}{2}}$, acting on $f \in \mathscr{S}'(\mathbb{R})/\mathscr{P}$, are defined by

$$\widehat{(-\Delta)^{\frac{r}{2}} f} := |\bullet|^r \cdot \widehat{f}, \quad \text{for } r \in \mathbb{R}. \tag{11}$$

We show that the signature is invariant under $(-\Delta)^{\frac{r}{2}}$. Again, note that the signature is well defined for $f \in \mathscr{S}'(\mathbb{R})/\mathscr{P}$.

Theorem 8. *Let $f \in \mathscr{S}'(\mathbb{R})/\mathscr{P}$ and $r \in \mathbb{R}$. Then,*

$$\sigma\left((-\Delta)^{\frac{r}{2}} f\right)(b) = \sigma f(b), \qquad \text{for all } b \in \mathbb{R}.$$

Now we turn to the fractional Hilbert transform, which was first introduced in [7]. We follow the definition given in [8]. For $\alpha \in \mathbb{R}$, the fractional Hilbert transform $\mathcal{H}^\alpha$ is defined on $\mathscr{S}'(\mathbb{R})/\mathscr{P}$ by

$$\widehat{\mathcal{H}^\alpha f} := e^{-i\alpha\frac{\pi}{2}\cdot \operatorname{sgn}(\bullet)} \cdot \widehat{f}. \tag{12}$$

The following theorem shows that the fractional Hilbert transform $\mathcal{H}^\alpha$ acts on the signature as multiplication by $e^{i\alpha\frac{\pi}{2}}$, i.e., as a rotation in the complex plane.

Theorem 9. *Let $f \in \mathscr{S}'(\mathbb{R})/\mathscr{P}$ and $b \in \mathbb{R}$. Then*

$$\sigma\left(\mathcal{H}^\alpha f\right)(b) = e^{i\alpha\frac{\pi}{2}} \cdot \sigma f(b). \tag{13}$$

See Table I for a comparison between local Sobolev regularity index of f, denoted by s_f (cf. e.g. [4]), and the signature under action of fractional Laplacians and fractional Hilbert transforms.

The next two examples show that the points of non-zero signature in general do not coincide with the singular support, cf. (2).

Example 10. Consider the Weierstraß function (see e.g. [2])

$$f(x) = \sum_{n=0}^{\infty} r^n \cos(t^n x), \qquad \text{where } 0 < r < 1 \text{ and } rt \geq 1;$$

	Sobolev regularity index	Signature
Fractional differentiation	$s_{(-\Delta)^{\frac{r}{2}} f} = s_f - r$	$\sigma((-\Delta)^{\frac{r}{2}} f) = \sigma(f)$
Fractional Hilbert transform	$s_{\mathcal{H}^\alpha f} = s_f$	$\sigma(\mathcal{H}^\alpha f) = e^{i\alpha\frac{\pi}{2}}\sigma(f)$

TABLE I
The action of fractional Laplacians and fractional Hilbert transforms to the Sobolev regularity index s_f and the signature σf.

As f is nowhere differentiable, it follows that $\operatorname{sing\,supp} f = \mathbb{R}$. In [11], we have proved that $\sigma f(b) = 0$, for all $b \in \mathbb{R}$. Therefore, we see that in general $\operatorname{sing\,supp} f \not\subseteq \operatorname{supp} \sigma f$.

Example 11. Let $f = e^{-\gamma x^2}$ be a Gaussian function with $\gamma > 0$, and let κ be any signature wavelet. The singular support of f is empty because f is smooth. On the other hand, as the support of $\widehat{\kappa}$ is not empty,

$$\langle f, \kappa_{a,0}\rangle = \left\langle \widehat{f}, (\kappa_{a,0})^\vee \right\rangle = \sqrt{\pi} \int_{\mathbb{R}} e^{-\frac{\omega^2}{4\gamma}} (\kappa_{a,0})^\vee(\omega)\, \mathrm{d}\omega > 0,$$

for all $a > 0$, implying that the signature equals 1 at $b = 0$. Thus, in general, $\operatorname{supp} \sigma f \not\subseteq \operatorname{sing\,supp} f$. This shows that the converse inclusion does not hold either.

IV. Discretization and numerical experiments

Now we turn our attention to the practical computation of wavelet signs for sampled signals. In practice, only a finite number of wavelet scales $\{a_j\}_{j=1}^N$ is available. Furthermore, since we cannot test for convergence in (1) using every signature wavelet, we have to choose a suitable signature wavelet κ. Thus, we have to estimate the signature from the finite set of samples $\{\operatorname{sgn}\langle f, \kappa_{a_j,b}\rangle\}_{j=1}^N$.

To motivate our numerical approach, we begin by considering the following elementary convergence result for discrete samples.

Proposition 12. *Let f be a tempered distribution, $\{a_j\}_{j\in\mathbb{N}}$ a sequence such that $a_j \to 0$, and $b \in \mathbb{R}$. If $\sigma f(b) \neq 0$, then*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{j=1}^{N} \operatorname{sgn}\langle f, \kappa_{a_j,b}\rangle = \sigma f(b) \tag{14}$$

for all signature wavelets κ.

Proposition 12 suggests the Cesàro limit (14) as an alternative to computing a non-zero signature $\sigma f(b)$. Note that $\sigma f(b)$ is of modulus 1 and so is the Cesàro limit (14). Furthermore, the elements of the Cesàro sequence

$$\frac{1}{N} \sum_{j=1}^{N} \operatorname{sgn}\langle f, \kappa_{a_j,b}\rangle, \quad N \in \mathbb{N},$$

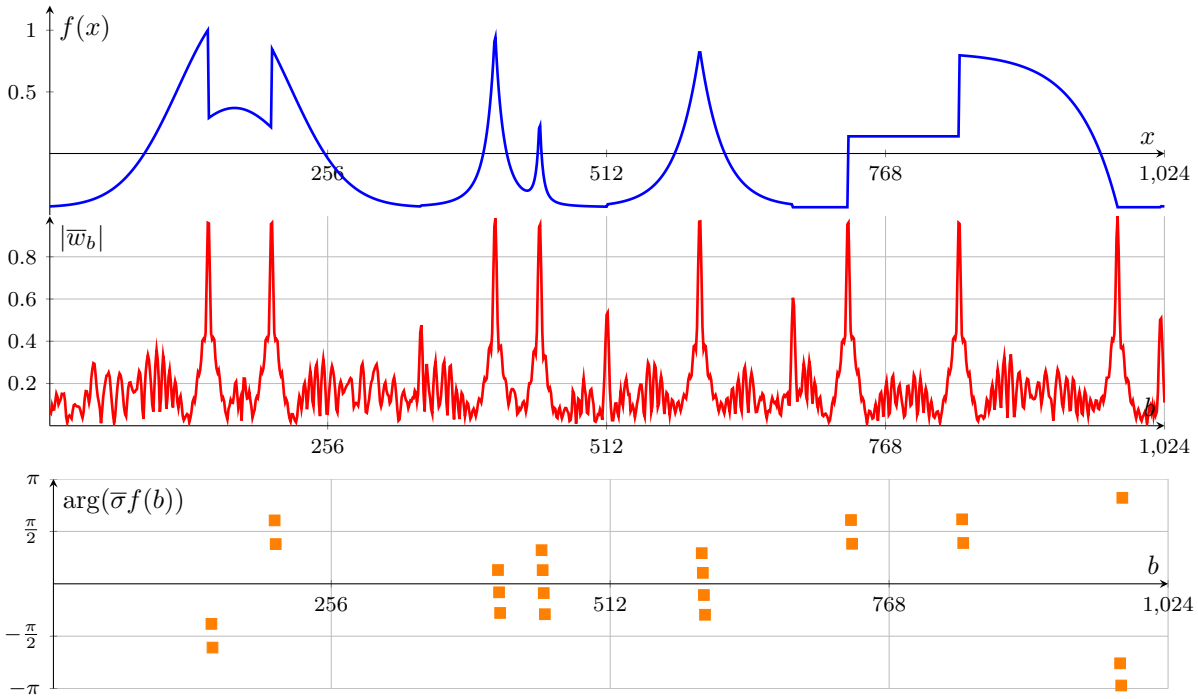


Fig. 2. The discrete signature of a sample signal (top) taken from Wavelab [1]. We observe that the absolute value of the mean $|\overline{w}_b|$ is large at the feature points and much lower at the other points (center). The bottom plot depicts the discrete signature in phase angle representation. We see that the signature clusters around the angles $\pm\frac{\pi}{2}$ at the step-like points, and around π and 0 at cusp-like points. The threshold τ is set equal to 0.7 in this experiment. If we choose a lower threshold parameter, say $\tau = 0.4$, then the discrete signature would also catch the subtle feature points, like the small jump at $x = 512$. However, in that case, we would require a non-maximum suppression to maintain the sharp localization of the pronounced feature points.

are not necessarily of modulus one, but their moduli converge to 1 as N goes to infinity. This observation motivates the following procedure for the numerical estimation of the signature.

Given a finite number of scale samples $\{a_j\}_{j=1}^N$, we interpret the mean of the sequence of discrete signs, given by

$$\overline{w}_b := \frac{1}{N}\sum_{j=1}^{N} \operatorname{sgn}\left\langle f, \kappa_{a_j,b}\right\rangle, \tag{15}$$

as the N-th element of a Cesàro sequence. If the absolute value $|\overline{w}_b|$ is close to 1, we consider the Cesàro sequence as being convergent, with $\operatorname{sgn}\overline{w}_b$ giving an estimate of $\sigma f(b)$. On the other hand, a small value of $|\overline{w}_b|$ suggests a vanishing signature. More precisely, we consider $|\overline{w}_b|$ to be non zero if it exceeds some empirical threshold parameter τ between 0 and 1. Hence, we propose a discrete estimate $\overline{\sigma} f(b)$ of the signature of the form

$$\overline{\sigma} f(b) := \begin{cases} \operatorname{sgn}\overline{w}_b, & \text{if } |\overline{w}_b| \geq \tau, \\ 0, & \text{elsewhere.} \end{cases} \tag{16}$$

In Figure 2, we see a numerical experiment based on the above procedure. We observe that the modulus of the mean $|\overline{w}_b|$ is large at the salient points. Furthermore, we see that the discrete signature is oriented towards the imaginary axis for jump discontinuities and oriented to the real axis for cusp singularities. This experiment illustrates that the procedure proposed above yields a reasonable way to compute the signature numerically. We used the Meyer-type signature wavelet (5) and scale samples of the form $a_j = 2^{-\frac{j}{3}}$, with $j = 0, 1, \ldots, 15$. The threshold parameter was set to $\tau = \frac{1}{2}\sqrt{2} \approx 0.7$.

In [10], a generalization of the discrete signature to higher dimensions is proposed, which can be applied directly for sign based edge detection and edge analysis. That generalization bases on monogenic wavelets similar to those of [3].

References

[1] D. Donoho, A. Maleki, and M. Shahram, *Wavelab 850*, Software toolkit for time-frequency analysis, (2006).

[2] G. Hardy, *Weierstrass's non-differentiable function*, Trans. Amer. Math. Soc, 17 (1916), pp. 301–325.

[3] S. Held, M. Storath, P. Massopust, and B. Forster, *Steerable wavelet frames based on the Riesz transform*, IEEE Transactions on Image Processing, 19 (2010), pp. 653–667.

[4] L. Hörmander, *The analysis of linear partial differential operators III: Pseudo-differential operators*, vol. 3, Springer verlag, 2007.

[5] S. Jaffard, *Pointwise smoothness, two-microlocalization and wavelet coefficients*, Publicacions Matematiques, 35 (1991), pp. 155–168.

[6] P. Kovesi, *Image features from phase congruency*, Videre: Journal of Computer Vision Research, 1 (1999), pp. 1–26.

[7] A. Lohmann, D. Mendlovic, and Z. Zalevsky, *Fractional hilbert transform*, Optics letters, 21 (1996), pp. 281–283.

[8] Y. Luchko, H. Matrínez, and J. Trujillo, *Fractional fourier transform and some of its applications*, Fract. Calc. Appl. Anal, 11 (2008), pp. 457–470.

[9] S. Mallat, *A wavelet tour of signal processing: The sparse way*, Academic, Burlington, (2009).

[10] M. Storath, *Amplitude and sign decompositions by complex wavelets – Theory and applications to image analysis*, PhD thesis, Technische Universität München, Germany, 2013.

[11] M. Storath, L. Demaret, and P. Massopust, *Signal analysis based on complex wavelet signs*, arXiv:1208.4578.

Balayage and short time Fourier transform frames

Enrico Au-Yeung
Pacific Institute for the Mathematical Sciences
Vancouver, BC V6T 1Z4 Canada
Email: enricoauy@math.ubc.ca

John J. Benedetto
Norbert Wiener Center, Department of Mathematics
University of Maryland, College Park, MD 20742 USA
Email: jjb@math.umd.edu

***Abstract*—Using his formulation of the potential theoretic notion of balayage and his deep results about this idea, Beurling gave sufficient conditions for Fourier frames in terms of balayage. The analysis makes use of spectral synthesis, due to Wiener and Beurling, as well as properties of strict multiplicity, whose origins go back to Riemann. In this setting and with this technology, we formulate and prove non-uniform sampling formulas in the context of the short time Fourier transform (STFT).**

I. Introduction

A. Background and theme

Frames provide a natural tool for dealing with signal reconstruction in the presence of noise in the setting of overcomplete sets of atoms, and with the goals of numerical stability and robust signal representation. Fourier frames were originally studied in the context of non-harmonic Fourier series by Duffin and Schaeffer [1], with a history going back to Paley and Wiener [2] (1934) and farther, and with significant activity in the 1930s and 1940s, e.g., see [3]. Since [1], there have been significant contributions by Beurling (unpublished 1959-1960 lectures), [4], [5], Beurling and Malliavin [6], [7], Kahane [8], Landau [9], Jaffard [10], and Seip [11], [12].

Definition I.1. (Frame) Let H be a separable Hilbert space. A sequence $\{x_n\}_{n\in\mathbb{Z}} \subseteq H$ is a *frame* for H if there are positive constants A and B such that

$$\forall\ f \in H, \quad A\|f\|^2 \leq \sum_{n\in\mathbb{Z}} |\langle f, x_n\rangle|^2 \leq B\|f\|^2.$$

The constants A and B are lower and upper frame bounds, respectively.

Our overall goal is to formulate a general theory of Fourier frames and non-uniform sampling formulas parametrized by the space $M_b(\mathbb{R}^d)$ of bounded Radon measures, see [13]. This formulation provides a natural way to generalize non-uniform sampling to the setting of short time Fourier transforms (STFTs) [14], Gabor theory [15], [16], [17], and pseudo-differential operators [14], [18], The techniques are based on Beurling's methods from 1959-1960, [5], [4], which incorporate balayage, spectral synthesis, and strict multiplicity. In this short paper, we show how to achieve this goal for STFTs.

B. Definitions

We define the Fourier transform $\mathcal{F}(f)$ of $f \in L^2(\mathbb{R}^d)$ and its inverse Fourier transform $\mathcal{F}^{-1}(f)$ by

$$\mathcal{F}(f)(\gamma) = \widehat{f}(\gamma) = \int_{\mathbb{R}^d} f(x)e^{-2\pi i x\cdot\gamma}\ dx,$$

and

$$\mathcal{F}^{-1}(\widehat{f})(\gamma) = f(x) = \int_{\widehat{\mathbb{R}}^d} \widehat{f}(\gamma)e^{2\pi i x\cdot\gamma}\ d\gamma.$$

$\widehat{\mathbb{R}}^d$ denotes $\mathbb{R}^d$ considered as the spectral domain. We write $F^\vee(x) = \int_{\widehat{\mathbb{R}}^d} F(\gamma)e^{2\pi i x\cdot\gamma}\, d\gamma$. The notation "$\int$"" designates integration over $\mathbb{R}^d$ or $\widehat{\mathbb{R}}^d$. When f is a bounded continuous function, its Fourier transform is defined in the sense of distributions. If $X \subseteq \mathbb{R}^d$, where X is closed, then $M_b(X)$ is the space of bounded Radon measures μ with the support of μ contained in X. $C_b(\mathbb{R}^d)$ denotes the space of complex-valued bounded continuous functions on $\mathbb{R}^d$.

Definition I.2. (Fourier frame) Let $E \subseteq \mathbb{R}^d$ be a sequence and let $\Lambda \subseteq \widehat{\mathbb{R}}^d$ be a compact set. Notationally, let $e_x(\gamma) = e^{2\pi i x\cdot\gamma}$. The sequence $\mathcal{E}(E) = \{e_{-x} : x \in E\}$ is a *Fourier frame* for $L^2(\Lambda)$ if there are positive constants A and B such that

$$\forall\ F \in L^2(\Lambda),$$
$$A\|F\|^2_{L^2(\Lambda)} \leq \sum_{x\in E} |\langle F, e_{-x}\rangle|^2 \leq B\|F\|^2_{L^2(\Lambda)}.$$

Define the *Paley-Wiener space*,

$$PW_\Lambda = \{f \in L^2(\mathbb{R}^d) : \text{supp}(\widehat{f}) \subseteq \Lambda\}.$$

Clearly, $\mathcal{E}(E)$ is a Fourier frame for $L^2(\Lambda)$ if and only if the sequence,

$$\{(e_{-x}\ \mathbb{1}_\Lambda)^\vee : x \in E\} \subseteq PW_\Lambda,$$

is a frame for PW_Λ, in which case it is called a *Fourier frame* for PW_Λ. Note that $\langle F, e_{-x}\rangle = f(x)$ for $f \in PW_\Lambda$, where $\widehat{f} = F \in L^2(\widehat{\mathbb{R}}^d)$ can be considered an element of $L^2(\Lambda)$.

Beurling introduced the following definition in his 1959-1960 lectures.

Definition I.3. (Balayage) Let $E \subseteq \mathbb{R}^d$ and $\Lambda \subseteq \widehat{\mathbb{R}}^d$ be closed sets. *Balayage* is possible for $(E, \Lambda) \subseteq \mathbb{R}^d \times \widehat{\mathbb{R}}^d$ if

$$\forall \mu \in M_b(\mathbb{R}^d),\ \exists \nu \in M_b(E) \text{ such that } \widehat{\mu} = \widehat{\nu} \text{ on } \Lambda.$$

Balayage originated in potential theory, where it was introduced by Christoffel (early 1870s) and by Poincaré (1890). Kahane formulated balayage for the harmonic analysis of restriction algebras. The set, Λ, of group characters (in this case $\mathbb{R}^d$) is the analogue of the original role of Λ in balayage as a set of potential theoretic kernels.

Let $\mathcal{C}(\Lambda) = \{f \in C_b(\mathbb{R}^d) : \text{supp}(\widehat{f}) \subseteq \Lambda\}$.

Definition I.4. (Spectral synthesis) A closed set $\Lambda \subseteq \widehat{\mathbb{R}}^d$ is a set of *spectral synthesis (S-set)* if

$$\forall f \in \mathcal{C}(\Lambda) \text{ and } \forall \mu \in M_b(\mathbb{R}^d),$$
$$\widehat{\mu} = 0 \text{ on } \Lambda \Rightarrow \int f \, d\mu = 0,$$

see [19].

Closely related to spectral synthesis is the ideal structure of L^1, which can be thought of as the Nullstellensatz of harmonic analysis. As examples of sets of spectral synthesis, polyhedra are S-sets, and the middle-third Cantor set is an S-set which contains non-S-sets. Laurent Schwartz (1947) showed that $S^2 \subseteq \widehat{\mathbb{R}}^3$ is not an S-set; and, more generally, Malliavin (1959) proved that every non-discrete locally compact abelian group contains non-S sets. See [19] for a unified treatment of this material.

Definition I.5. (Strict multiplicity) A closed set $\Gamma \subseteq \widehat{\mathbb{R}}^d$ is a set of *strict multiplicity* if

$$\exists \mu \in M_b(\Gamma) \setminus \{0\} \text{ such that } \lim_{\|x\| \to \infty} |\mu^\vee(x)| = 0.$$

The notion of strict multiplicity was motivated by Riemann's study of sets of uniqueness for trigonometric series. Menchov (1906) showed that there exists a closed set $\Gamma \subseteq \widehat{\mathbb{R}}/\mathbb{Z}$ and $\mu \in M(\Gamma) \setminus \{0\}$, such that $|\Gamma| = 0$ and $\mu^\vee(n) = O((\log|n|)^{-1/2}), |n| \to \infty$. There have been intricate refinements of Menchov's result by Bary (1927), Littlewood (1936), Beurling, et al., see [19].

The above concepts are used in the deep proof of the following theorem.

Theorem I.6. *Assume that Λ is an S-set of strict multiplicity, and that balayage is possible for (E, Λ). Let $\Lambda_\epsilon = \{\gamma \in \widehat{\mathbb{R}}^d : dist\,(\gamma, \Lambda) \leq \epsilon\}$. There is $\epsilon_0 > 0$ such that if $0 < \epsilon < \epsilon_0$, then balayage is possible for (E, Λ_ϵ).*

Definition I.7. A sequence $E \subseteq \mathbb{R}^d$ is *separated* if

$$\exists\, r > 0 \text{ such that } \inf\{\|x - y\| : x, y \in E \text{ and } x \neq y\} \geq r.$$

The following theorem, due to Beurling, gives a sufficient condition for Fourier frames in terms of balayage. Its history and structure are analyzed in [13] as part of a more general program. Theorem I.6 is used in its proof.

Theorem I.8. *Assume that $\Lambda \subseteq \widehat{\mathbb{R}}^d$ is an S-set of strict multiplicity and that $E \subseteq \mathbb{R}^d$ is a separated sequence. If balayage is possible for (E, Λ), then $\mathcal{E}(E)$ is a Fourier frame for $L^2(\Lambda)$, i.e., $\{(e_{-x}\, \mathbb{1}_\Lambda)^\vee : x \in E\}$ is a Fourier frame for PW_Λ.*

See [9], [20], [21] (SampTA 1999), and [22].

II. Short time Fourier transform (STFT) frame inequalities

Definition II.1. Let $f, g \in L^2(\mathbb{R}^d)$. The *short time Fourier transform* (STFT) of f with respect to g is the function $V_g f$ on $\mathbb{R}^{2d}$ defined as

$$V_g f(x, \omega) = \int f(t)\overline{g(t-x)}\; e^{-2\pi i t\cdot\omega}\; dt,$$

see [14], [18] (chapter 8).

The STFT is uniformly continuous on $\mathbb{R}^{2d}$. Further, for a fixed "window" $g \in L^2(\mathbb{R}^d)$ with $\|g\|_2 = 1$, we can recover the original function $f \in L^2(\mathbb{R}^d)$ from its STFT $V_g f$ by means of the vector-valued integral inversion formula,

$$f = \int\int V_g f(x, \omega)\; e_\omega \tau_x g \; d\omega \; dx,$$

where $(\tau_x g)(t) = g(t - x)$.

Theorem II.2. *Let $E = \{x_n\} \subseteq \mathbb{R}^d$ be a separated sequence, that is symmetric about $0 \in \mathbb{R}^d$; and let $\Lambda \subseteq \mathbb{R}^d$ be an S-set of strict multiplicity that is compact, convex, and symmetric about $0 \in \widehat{\mathbb{R}}^d$. Assume balayage is possible for $(E, \Lambda.)$ Further, let $g \in L^2(\mathbb{R}^d), \widehat{g} = G$, have the property that $\|g\|_2 = 1$.*

a. We have that

$$\exists\; A > 0, \quad \textit{such that} \quad \forall f \in PW_\Lambda \backslash \{0\}, \widehat{f} = F,$$

$$A\|f\|_2^2 \leq \sum_{x\in E} \int |V_G F(\omega, x)|^2 \; d\omega$$
$$= \sum_{x\in E} \int |V_g f(x, \omega)|^2 \; d\omega.$$

b. Let $G_0(\lambda) = 2^{d/4} e^{-\pi\|\lambda\|^2}$ so that $\|G_0\|_2 = 1$; and assume $\|V_{G_0} G\|_1 < \infty$. We have that

$$\exists\; B > 0, \quad \textit{such that} \quad \forall f \in PW_\Lambda \backslash \{0\}, \widehat{f} = F,$$

$$\sum_{x\in E} \int |V_g f(x, \omega)|^2 \; d\omega = \sum_{x\in E} \int |V_G F(\omega, -x)|^2 \; d\omega$$
$$\leq B\|f\|_2^2,$$

where B can be taken as $C\|V_{G_0} G\|_1$ and where

$$C = \sup_{y,\gamma} \left\{ \sum_{x\in E} \int |V_{G_0} G_0(\gamma + \omega, y + x)| \; d\omega \right\}.$$

The technique of using G_0 goes back to Feichtinger and Zimmermann [23] (Lemma 3.2.15) for a related type of problem, see also [16] (Lemma 3.2).

We next consider balayage being possible for (E, Λ), where $E = \{(s_m, t_n)\} \subseteq \mathbb{R}^{2d}$ and $\Lambda \subseteq \widehat{\mathbb{R}}^{2d}$. This allows us to express the STFT $V_g f$ of f as

$$V_g f(y, \omega) = \sum_m \sum_n a_{mn}(y, \omega) h(s_m - y, t_n - \omega) V_g f(s_m, t_n),$$

where

$$\sum_m \sum_n |a_{mn}(y, \omega)| < \infty.$$

The following result and others like it, including Theorem II.2, can be formulated in terms of (X, μ) frames, [24], [25]. [26].

Theorem II.3. *Assume balayage is possible for (E, Λ), where $E = \{(s_m, t_n)\} \subseteq \mathbb{R}^{2d}$ is separated, and $\Lambda \subseteq \widehat{\mathbb{R}}^{2d}$ is an S-set*

that is compact, convex, and symmetric about $0 \in \widehat{\mathbb{R}}^{2d}$. *Fix a window function* $g \in L^2(\mathbb{R}^d)$ *such that* $\|g\|_2 = 1$. *There are constants* $A, B > 0$, *such that if* $f \in L^2(\mathbb{R}^d)$ *satisfies the conditions,*
(1) $V_g f \in L^1(\mathbb{R}^{2d})$ *and*
(2) $\mathcal{F}(V_g f)(\zeta_1, \zeta_2)$ *has support* $\subseteq \Lambda \subseteq \widehat{\mathbb{R}}^{2d}$,
then

$$A \int |f(x)|^2 \, dx \leq \sum_m \sum_n |V_g f(s_m, t_n)|^2 \leq B \int |f(x)|^2 \, dx.$$

The hypothesis that $V_g f \in L^1(\mathbb{R}^{2d})$ means that f belongs to the Feichtinger algebra $\mathcal{S}_0(\mathbb{R}^d)$. It is the smallest Banach space that is invariant under translations and modulations. There are other equivalent characterizations of $\mathcal{S}_0(\mathbb{R}^d)$, see [27], [23]. Fix a function $\mathcal{S}_0(\mathbb{R}^d)$ and define the vector space $\mathcal{M}_1^1$ of all non-uniform Gabor expansions

$$f = \sum_{n=1}^{\infty} c_n \tau_{x_n} e_{\omega_n} g,$$

where $\{(x_n, \omega_n) \in \mathbb{R}^{2d}, n \in \mathbb{N}\}$ is an arbitrary countable set of numbers and $\sum_{n=1}^{\infty} |c_n| < \infty$. For this space, the norm is taken to be $\inf \sum_{n=1}^{\infty} |c_n|$, where the infimum is taken over all possible representations. Then the vector space $\mathcal{M}_1^1$ coincides with $\mathcal{S}_0(\mathbb{R}^d)$. For functions in $\mathcal{S}_0(\mathbb{R}^d)$, Theorem II.3 should be compared to the following theorem of Gröchenig [15], [14] (Chapter 12):

Theorem II.4. *Given any* $g \in \mathcal{S}_0(\mathbb{R}^d)$. *There is* $r = r(g) > 0$ *such that if* $E = \{(s_n, \sigma_n)\} \subseteq \mathbb{R}^d \times \widehat{\mathbb{R}}^d$ *is a separated sequence with the property that*

$$\bigcup_{n=1}^{\infty} \overline{B((s_n, \sigma_n), r(g))} = \mathbb{R}^d \times \widehat{\mathbb{R}}^d,$$

then the frame operator, $S = S_{g,E}$, *defined by*

$$S_{g,E} f = \sum\nolimits_{n=1}^{\infty} \langle f, \tau_{s_n} e_{\sigma_n} g \rangle \, \tau_{s_n} e_{\sigma_n} g,$$

is invertible on $\mathcal{S}_0(\mathbb{R}^d)$.

Moreover, every $f \in \mathcal{S}_0(\mathbb{R}^d)$ *has a non-uniform Gabor expansion,*

$$f = \sum\nolimits_{n=1}^{\infty} \langle f, \tau_{x_n} e_{\omega_n} g \rangle S_{g,E}^{-1} (\tau_{x_n} e_{\omega_n} g),$$

where the series converges unconditionally in $\mathcal{S}_0(\mathbb{R}^d)$.
(E *depends on* g.*)*

A critical, thorough comparison of Theorems II.3 and II.4 is given in [13].

ACKNOWLEDGMENT

The first named author gratefully acknowledges the support of MURI-AFOSR Grant FA9550-05-1-0443. The second named author gratefully acknowledges the support of MURI-ARO Grant W911NF-09-1-0383 and NGA Grant HM-1582-08-1-0009. Both authors also benefitted from insightful observations by Professors Carlos Cabrelli, Matei Machedon, Ursula Molter, and Kasso Okoudjou, as well as from Dr. Henry J. Landau, the grand master of Fourier frames. Finally, we would like to thank the referees for their insightful observations.

REFERENCES

[1] R. J. Duffin and A. C. Schaeffer, "A class of nonharmonic Fourier series," *Trans. Amer. Math. Soc.*, vol. 72, pp. 341–366, 1952.
[2] R. E. A. C. Paley and N. Wiener, *Fourier Transforms in the Complex Domain*, ser. Amer. Math. Society Colloquium Publications. Providence, RI: American Mathematical Society, 1934, vol. XIX.
[3] N. Levinson, *Gap and Density Theorems*, ser. Amer. Math. Society Colloquium Publications. Providence, RI: American Mathematical Society, 1940, vol. XXVI.
[4] A. Beurling, *The Collected Works of Arne Beurling. Vol. 2. Harmonic Analysis.* Boston: Birkhäuser, 1989.
[5] ——, "Local harmonic analysis with some applications to differential operators," *Some Recent Advances in the Basic Sciences, Vol. 1 (Proc. Annual Sci. Conf., Belfer Grad. School Sci., Yeshiva Univ., New York, 1962–1964)*, pp. 109–125, 1966.
[6] A. Beurling and P. Malliavin, "On Fourier transforms of measures with compact support," *Acta Mathematica*, vol. 107, pp. 291–309. [Online]. Available: http://dx.doi.org/10.1007/BF02545792
[7] ——, "On the closure of characters and the zeros of entire functions," *Acta Mathematica*, vol. 118, pp. 79–93. [Online]. Available: http://dx.doi.org/10.1007/BF02392477
[8] J.-P. Kahane, "Sur certaines classes de séries de Fourier absolument convergentes," *J. Math. Pures Appl. (9)*, vol. 35, pp. 249–259, 1956.
[9] H. J. Landau, "Necessary density conditions for sampling and interpolation of certain entire functions," *Acta Mathematica*, vol. 117, pp. 37–52, 1967. [Online]. Available: http://dx.doi.org/10.1007/BF02395039
[10] S. Jaffard, "A density criterion for frames of complex exponentials," *Michigan Math. J.*, vol. 38, pp. 339–348, 1991.
[11] K. Seip, "On the connection between exponential bases and certain related sequences in $L^2(-\pi, \pi)$," *J. Funct. Anal.*, vol. 130, pp. 131–160, 1995.
[12] J. Ortega-Cerdà and K. Seip, "Fourier frames," *Ann. of Math.*, vol. 155, no. 3, pp. 789–806, 2002.
[13] J. J. Benedetto and E. Au-Yeung, "Generalized Fourier frames in terms of balayage," *forthcoming*.
[14] K. Gröchenig, *Foundations of Time-Frequency Analysis*, ser. Applied and Numerical Harmonic Analysis. Boston, MA: Birkhäuser Boston Inc., 2001.
[15] ——, "Describing functions: atomic decompositions versus frames," *Monatsh. Math.*, vol. 112, pp. 1–42, 1991.
[16] H. G. Feichtinger and W. Sun, "Stability of Gabor frames with arbitrary sampling points," *Acta Mathematica Hungarica*, vol. 113, pp. 187–212, 2006. [Online]. Available: http://dx.doi.org/10.1007/s10474-006-0099-4
[17] D. Labate, G. Weiss, and E. Wilson, "An Approach to the Study of Wave Packet Systems," *Contemporary Mathematics, Wavelets, Frames, and Operator Theory*, vol. 345, pp. 215–235, 2004.
[18] K. Gröchenig, "A pedestrian's approach to pseudodifferential operators," in *Harmonic Analysis and Applications*, ser. Appl. Numer. Harmon. Anal. Boston: Birkäuser, 2006, pp. 139–169.
[19] J. J. Benedetto, *Spectral Synthesis*. New York-London: Academic Press, Inc., 1975.
[20] J. J. Benedetto and H.-C. Wu, "A multidimensional irregular sampling algorithm and applications," *IEEE-ICASSP*, 1999.
[21] J. J. Benedetto and H. Wu, "A Beurling covering theorem and multidimensional irregular sampling," in *SampTA*, Loen, 1999.
[22] J. J. Benedetto and H.-C. Wu, "Non-uniform sampling and spiral MRI reconstruction," *SPIE*, 2000.

[23] H. G. Feichtinger and G. Zimmermann, "A Banach space of test functions for Gabor analysis." in *Gabor analysis and algorithms,*, ser. Appl. Numer. Harmon. Anal. Boston: Birkhäuser, 1998, pp. 123–170.
[24] S. T. Ali, J.-P. Antoine, and J.-P. Gazeau, *Coherent states, wavelets and their generalizations*, ser. Graduate Text in Contemporary Physics. New York: Springer-Verlag, 2000.
[25] J.-P. Gabardo and D. Han, "Frames associated with measurable spaces. frames." *Adv. Comput. Math.*, vol. 18, no. 2-4, pp. 127–147, 2003.
[26] M. Fornasier and H. Rauhut, "Continuous frames, function spaces, and the discretization problem," *J. Fourier Anal. Appl.*, vol. 11, no. 3, pp. 245–287, 2005.
[27] H. G. Feichtinger, "On a new Segal algebra," *Monatsh. Math.*, vol. 92, pp. 269–289, 1981.

Fundamental Limits of Phase Retrieval

Afonso S. Bandeira
Program in Applied and Computational Mathematics
Princeton University
Princeton, New Jersey 08544
Email: ajsb@math.princeton.edu

Jameson Cahill
Department of Mathematics
University of Missouri
Columbia, Missouri 65211
Email: jameson.cahill@gmail.com

Dustin G. Mixon and Aaron A. Nelson
Department of Mathematics and Statistics
Air Force Institute of Technology
Wright-Patterson AFB, Ohio 45433
Email: dustin.mixon@afit.edu
aaron.nelson@afit.edu

***Abstract*—Recent advances in convex optimization have led to new strides in the phase retrieval problem over finite-dimensional vector spaces. However, certain fundamental questions remain: What sorts of measurement vectors uniquely determine every signal up to a global phase factor, and how many are needed to do so? This paper presents several results that address these questions, specifically in the less-understood complex case. In particular, we characterize injectivity, we identify that the complement property is indeed necessary, we pose a conjecture that $4M - 4$ generic measurement vectors are necessary and sufficient for injectivity in M dimensions, and we describe how to prove this conjecture in the special cases where $M = 2, 3$. To prove the $M = 3$ case, we leverage a new test for injectivity, which can be used to determine whether any 3-dimensional measurement ensemble is injective.**

I. Introduction

Phase retrieval is the problem of recovering a signal from absolute values (squared) of linear measurements, called intensity measurements. However, non-injectivity is inherent to many measurement processes. For instance, intensity measurements with the identity basis effectively discard all phase information contained in the signal's entries. As a result, many researchers invoke a priori knowledge of the desired signal in order to restrict to a smaller signal class over which the intensity measurements might be injective. To avoid the various ad hoc methods that invariably follow, an alternative approach to phase retrieval, as introduced in 2006 by Balan, Casazza and Edidin [3], seeks injectivity by designing a larger ensemble of intensity measurements. Using this approach, Balan et al. [3] characterized injectivity in the real case and further leveraged algebraic geometry to show that $4M-2$ intensity measurements suffice for injectivity over M-dimensional complex signals. This has since sparked a search for practical phase retrieval guarantees. For example, viewing intensity measurements as Hilbert-Schmidt inner products between rank-1 operators, Candès, Strohmer and Voroninski [7] applied certain intuition from convex optimization to reconstruct the desired M-dimensional signal with semidefinite programming using only $\mathcal{O}(M \log M)$ random measurements. Another approach uses the polarization identity to discern relative phases between certain intensity measurements using $\mathcal{O}(M \log M)$ random measurements in concert with an expander graph [1].

Despite these recent advances in phase retrieval algorithms, there remains a lack of understanding about the fundamental requirements for intensity measurements to be injective. For example, it was widely believed that $3M - 2$ intensity measurements sufficed for injectivity, until recently disproved by Heinosaari, Mazzarella and Wolf [10] using embedding theorems from differential geometry. Heinosaari et al. were able to establish the necessity of $(4 + o(1))M$ measurements for injectivity, but the following problem still remains:

Problem 1. *What are the necessary and sufficient conditions for measurement vectors to lend injective intensity measurements?*

This paper addresses this problem by first providing the only known characterization of injectivity in the complex case (Theorem 4). Next, we make a rather surprising identification: that intensity measurements are injective in the complex case precisely when the corresponding phase-only measurements are injective in some sense (Theorem 5). We then use this identification to establish the necessity of the complement property for injectivity (Theorem 7). Later, we conjecture that $4M - 4$ intensity measurements are necessary and sufficient for injectivity in the complex case, which we validate in the cases where $M = 2, 3$ (Theorems 10 and 12). We also introduce a new test for injectivity, which we then use to verify the injectivity of a certain quantum-mechanics-inspired measurement ensemble; with this ensemble, we conclude by suggesting a refinement of Wright's conjecture from [12] (see Conjecture 13). The proofs of the presented results are provided in [4].

Before we begin, let $\Phi = \{\varphi_n\}_{n=1}^N$ in $V = \mathbb{R}^M$ or $\mathbb{C}^M$ be a given collection of measurement vectors and consider the intensity measurement process defined by

$$(\mathcal{A}(x))(n) := |\langle x, \varphi_n \rangle|^2.$$

Note that $\mathcal{A}(x) = \mathcal{A}(y)$ whenever $y = cx$ for some scalar c of unit modulus. Thus, the mapping $\mathcal{A}\colon V \to \mathbb{R}^N$ is necessarily not injective. To resolve this issue, we consider sets of the form V/S, where S is a multiplicative subgroup of the field of scalars. This notation means to identify vectors $x, y \in V$ which satisfy $y = cx$ for some scalar $c \in S$, and we write $y \equiv x \bmod S$ to convey this identification. Most of the time, V/S is either $\mathbb{R}^M/\{\pm 1\}$ or $\mathbb{C}^M/\mathbb{T}$ (here, $\mathbb{T}$ is the complex unit circle), and the intensity measurement process is viewed

as a mapping $\mathcal{A}\colon V/S \to \mathbb{R}^N$. Injectivity of the measurement process is considered with respect to this mapping.

II. Injectivity and the complement property

Phase retrieval is impossible without injective intensity measurements. Balan, Casazza and Edidin [3] first analyzed injectivity by introducing the *complement property*, which we define in the following:

Definition 2. *We say $\Phi = \{\varphi_n\}_{n=1}^N$ in $\mathbb{R}^M$ ($\mathbb{C}^M$) satisfies the complement property (CP) if for every $S \subseteq \{1,\dots,N\}$, either $\{\varphi_n\}_{n\in S}$ or $\{\varphi_n\}_{n\in S^c}$ spans $\mathbb{R}^M$ ($\mathbb{C}^M$).*

The complement property is characteristic of injectivity in the real case:

Theorem 3. *Consider $\Phi = \{\varphi_n\}_{n=1}^N \subseteq \mathbb{R}^M$ and the mapping $\mathcal{A}\colon \mathbb{R}^M/\{\pm 1\} \to \mathbb{R}^N$ defined by $(\mathcal{A}(x))(n) := |\langle x, \varphi_n\rangle|^2$. Then $\mathcal{A}$ is injective if and only if Φ satisfies the complement property.*

This result was demonstrated in [3]. However, it was also erroneously used as justification for the necessity of CP for injectivity in the complex case. Although this statement is indeed true, the proof of Theorem 3 overlooks the peculiarity of equivalence modulo $\mathbb{T}$ and so cannot be used in the complex setting. We will address this issue, but first we characterize injectivity in the complex case:

Theorem 4. *Consider $\Phi = \{\varphi_n\}_{n=1}^N \subseteq \mathbb{C}^M$ and the mapping $\mathcal{A}\colon \mathbb{C}^M/\mathbb{T} \to \mathbb{R}^N$ defined by $(\mathcal{A}(x))(n) := |\langle x, \varphi_n\rangle|^2$. Viewing $\{\varphi_n\varphi_n^* u\}_{n=1}^N$ as vectors in $\mathbb{R}^{2M}$, denote $S(u) := \operatorname{span}_{\mathbb{R}}\{\varphi_n\varphi_n^* u\}_{n=1}^N$. Then the following are equivalent:*

(a) *$\mathcal{A}$ is injective.*
(b) *$\dim S(u) \geq 2M-1$ for every $u \in \mathbb{C}^M \setminus \{0\}$.*
(c) *$S(u) = \operatorname{span}_{\mathbb{R}}\{\mathrm{i}u\}^\perp$ for every $u \in \mathbb{C}^M \setminus \{0\}$.*

Note that unlike in the real case, it is not clear whether this characterization can be tested in finite time; instead of being a statement about all (finitely many) partitions of $\{1,\dots,N\}$, it is a statement about all nonzero vectors $u \in \mathbb{C}^M$. We can, however, view this characterization as an analog to the real case, in which the complement property is equivalent to having $\operatorname{span}\{\varphi_n\varphi_n^* u\}_{n=1}^N = \mathbb{R}^M$ for all nonzero $u \in \mathbb{R}^M$. The fact that more information is lost with phase in the complex case is what causes $\{\varphi_n\varphi_n^* u\}_{n=1}^N$ to fail to span all of $\mathbb{R}^{2M}$. As a result, it is still not intuitively apparent what it takes for an ensemble of complex vectors to yield injective intensity measurements. The following bizarre characterization was established while working toward a clearer understanding:

Theorem 5. *Consider $\Phi = \{\varphi_n\}_{n=1}^N \subseteq \mathbb{C}^M$ and the mapping $\mathcal{A}\colon \mathbb{C}^M/\mathbb{T} \to \mathbb{R}^N$ defined by $(\mathcal{A}(x))(n) := |\langle x, \varphi_n\rangle|^2$. Then $\mathcal{A}$ is injective if and only if the following statement holds: If for every $n = 1,\dots,N$, either $\arg(\langle x, \varphi_n\rangle^2) = \arg(\langle y, \varphi_n\rangle^2)$ or one of the sides is not well-defined, then $x = 0$, $y = 0$, or $y \equiv x \bmod \mathbb{R}\setminus\{0\}$.*

Theorem 5 is a consequence of a more general statement about the geometric properties of complex numbers: For $a, b \in \mathbb{C}$, $\operatorname{Im} a\bar{b} = 0$ if and only if $\arg(a^2) = \arg(b^2)$, $a = 0$, or $b = 0$. The proof leverages this fact within a restatement of part (c) of Theorem 4. This seemingly unrelated result is actually useful in correctly establishing the necessity of CP for injectivity in the complex case. Specifically, Theorem 5, leads to the following lemma, which in turn is used to prove necessity (Theorem 7).

Lemma 6. *Consider $\Phi = \{\varphi_n\}_{n=1}^N \subseteq \mathbb{C}^M$ and the mapping $\mathcal{A}\colon \mathbb{C}^M/\mathbb{T} \to \mathbb{R}^N$ defined by $(\mathcal{A}(x))(n) := |\langle x, \varphi_n\rangle|^2$. If $\mathcal{A}$ is injective, then the mapping $\mathcal{B}\colon \mathbb{C}^M/\{\pm 1\} \to \mathbb{R}^N$ defined by $(\mathcal{B}(x))(n) := \langle x, \varphi_n\rangle^2$ is also injective.*

Theorem 7. *Consider $\Phi = \{\varphi_n\}_{n=1}^N \subseteq \mathbb{C}^M$ and the mapping $\mathcal{A}\colon \mathbb{C}^M/\mathbb{T} \to \mathbb{R}^N$ defined by $(\mathcal{A}(x))(n) := |\langle x, \varphi_n\rangle|^2$. If $\mathcal{A}$ is injective, then Φ satisfies the complement property.*

The problem alluded to earlier concerning the proof of Theorem 3 is the reason that Theorem 7 is stated separately. This issue is resolved by using the injectivity of $\mathcal{B}$ modulo $\{\pm 1\}$. The proof is eerily similar to that of the necessity of CP for injectivity in Theorem 3, only using $\mathcal{B}$ in place of $\mathcal{A}$.

We emphasize here that the complement property is necessary but not sufficient for injectivity in the complex case. To see this, consider the ensemble $(1,0)$, $(0,1)$ and $(1,1)$. These certainly satisfy the complement property, but $\mathcal{A}((1,\mathrm{i})) = (1,1,2) = \mathcal{A}((1,-\mathrm{i}))$, despite the fact that $(1,\mathrm{i}) \not\equiv (1,-\mathrm{i}) \bmod \mathbb{T}$; in general, real frames fail to lend injective intensity measurements for the complex case. Indeed, a sufficient condition for injectivity in the complex case has yet to be found. As an analogy for what we really want, consider the notion of *full spark*: An ensemble $\{\varphi_n\}_{n=1}^N \subseteq \mathbb{R}^M$ is said to be full spark if every subcollection of M vectors spans $\mathbb{R}^M$. Full spark ensembles with $N \geq 2M-1$ necessarily satisfy the complement property, and the notion of full spark is simple enough to admit deterministic constructions [2], [11]. Because such constructions are particularly desirable for the complex case, finding a good sufficient condition for injectivity is an important problem that remains open.

III. Introducing the $4M-4$ Conjecture

Thinking of a matrix Φ as being built one column at a time, the rank-nullity theorem states that each column contributes to either the column space or the null space. If these columns are then used as linear measurement vectors, then the subspace that is actually sampled is described by the column space of Φ, while the null space captures the algebraic nature of redundancy in the measurements. An efficient sampling of an entire vector space would therefore apply a matrix Φ having a small null space and large column space. Although we are not dealing with linear measurements in our case, we would like to build our ensemble of intensity measurements so as to sample as much of the space as possible. More precisely, we are faced with the following question:

Problem 8. *For any dimension M, what is the smallest number $N^*(M)$ of injective intensity measurements, and how do we design such measurement vectors?*

To be clear, this problem was completely solved in the real case by Balan, Casazza and Edidin [3]. Indeed, Theorem 3 immediately implies that $2M-2$ intensity measurements are necessarily not injective, and furthermore that $2M-1$ measurements are injective if and only if the measurement vectors are full spark.

In the complex case, Problem 8 has some history in the quantum mechanics literature. For example, [12] presents *Wright's conjecture* that any pure state is uniquely determined by three observables. In other words, the conjecture states that there exist unitary matrices U_1, U_2 and U_3 such that $\Phi = [U_1\ U_2\ U_3]$ lends injective intensity measurements. Note that Wright's conjecture actually implies that $N^*(M) \leq 3M-2$, since U_1 determines the norm of the signal, rendering the last column of both U_2 and U_3 unnecessary. Finkelstein [8] later proved that $N^*(M) \geq 3M-2$ which, combined with Wright's conjecture, has led many to believe that $N^*(M) = 3M-2$. However, this was recently disproved by Heinosaari, Mazzarella and Wolf [10], who used embedding theorems from differential geometry to prove that $N^*(M) \geq 4M - 2\alpha(M-1) - 3$, where $\alpha(M-1) \leq \log_2(M)$ is the number of 1's in the binary representation of $M-1$. Combined with Balan, Casazza and Edidin's result that $N^*(M) \leq 4M-2$, we at least have the asymptotic expression $N^*(M) = (4+o(1))M$.

The lemma that follows will help to refine our intuition for $N^*(M)$. Before stating the result, however, we must first define the *super analysis operator* $\mathbf{A}\colon \mathbb{H}^{M\times M} \to \mathbb{R}^N$. Given an ensemble of measurement vectors $\{\varphi_n\}_{n=1}^N \subseteq \mathbb{C}^M$, this operator acts on the real M^2-dimensional vector space of self-adjoint $M \times M$ matrices, $\mathbb{H}^{M\times M}$, and is defined by $(\mathbf{A}H)(n) = \langle H, \varphi_n\varphi_n^*\rangle_{\mathrm{HS}}$, where $\langle\cdot,\cdot\rangle_{\mathrm{HS}}$ denotes the Hilbert-Schmidt inner product. Note that the super analysis operator is a linear operator which satisfies

$$(\mathbf{A}xx^*)(n) = \langle xx^*, \varphi_n\varphi_n^*\rangle_{\mathrm{HS}} = |\langle x, \varphi_n\rangle|^2 = (\mathcal{A}(x))(n).$$

To clarify, $x \bmod \mathbb{T}$ can be "lifted" to xx^*, a process which linearizes the intensity measurement process at the price of squaring the dimension of the vector space. This identification is not new, and as the following lemma shows, it can also be used to characterize injectivity:

Lemma 9. *$\mathcal{A}$ is not injective if and only if there exists a matrix of rank 1 or 2 in the null space of $\mathbf{A}$.*

Lemma 9 indicates that we want the null space of $\mathbf{A}$ to avoid nonzero matrices of rank ≤ 2. This is easier when the "dimension" of this set of matrices is small. As an exercise in intuition, we count real degrees of freedom to get an idea of this dimension: By the spectral theorem, almost every matrix in $\mathbb{H}^{M\times M}$ of rank ≤ 2 can be uniquely expressed in the form $\lambda_1 u_1u_1^* + \lambda_2 u_2u_2^*$. The pair of coefficients (λ_1,λ_2) introduces two degrees of freedom, while the vector u_1, which can be any vector in $\mathbb{C}^M$ of unit norm and is unique up to global phase, has a total of $2M-2$ real degrees of freedom. Finally, u_2 has the same norm and phase constraints as u_1, with the additional requirement that it must be orthogonal to u_1, (i.e., $\mathrm{Re}\langle u_2,u_1\rangle = \mathrm{Im}\langle u_2,u_1\rangle = 0$). Thus, u_2 has $2M-4$ real degrees of freedom. In this way we expect the set of matrices in question to have $2+(2M-2)+(2M-4) = 4M-4$ real dimensions.

If the set S of matrices of rank ≤ 2 formed a subspace of $\mathbb{H}^{M\times M}$, then we could expect it to have a nontrivial intersection with the null space of $\mathbf{A}$ whenever

$$\dim \mathrm{null}(\mathbf{A}) + (4M-4) > \dim(\mathbb{H}^{M\times M}) = M^2.$$

By the rank-nullity theorem, this would indicate that injectivity requires

$$N \geq \mathrm{rank}(\mathbf{A}) = M^2 - \dim \mathrm{null}(\mathbf{A}) \geq 4M-4.$$

Of course, this logic is not valid since S is not a subspace of $\mathbb{H}^{M\times M}$. It is, however, a special kind of set: a real projective variety (a real algebraic variety which is closed under scalar multiplication). If S were a projective variety over an *algebraically closed* field, then the projective dimension theorem (Theorem 7.2 of [9]) would imply that it intersects $\mathrm{null}(\mathbf{A})$ nontrivially whenever the dimensions are large enough: $\dim \mathrm{null}(\mathbf{A}) + \dim S > \dim \mathbb{H}^{M\times M}$, and so injectivity would require $N \geq 4M-4$. Unfortunately, this theorem is not valid when the field is $\mathbb{R}$; for example, the cone defined by $x^2+y^2-z^2=0$ in $\mathbb{R}^3$ is a projective variety of dimension 2, but its intersection with the 2-dimensional xy-plane is trivial, despite the fact that $2+2>3$.

In the absence of a proof, we pose the natural conjecture:

The $4M-4$ Conjecture. *Consider $\Phi = \{\varphi_n\}_{n=1}^N \subseteq \mathbb{C}^M$ and the mapping $\mathcal{A}\colon \mathbb{C}^M/\mathbb{T} \to \mathbb{R}^N$ defined by $(\mathcal{A}(x))(n) := |\langle x,\varphi_n\rangle|^2$. If $M \geq 2$, then the following statements hold:*

(a) *If $N < 4M-4$, then $\mathcal{A}$ is not injective.*
(b) *If $N \geq 4M-4$, then $\mathcal{A}$ is injective for generic Φ.*

For the sake of clarity, we state what is meant by the word "generic." A real algebraic variety is the set of common zeros of a finite set of polynomials with real coefficients. Taking all such varieties in $\mathbb{R}^n$ to be closed sets then defines the *Zariski topology* on $\mathbb{R}^n$. If we view Φ as a member of $\mathbb{R}^{2MN}$, we then say a *generic* Φ is any member of some nonempty Zariski-open subset of $\mathbb{R}^{2MN}$. Since Zariski-open sets are either empty or dense with full measure, genericity is a particularly strong property. As such, another way to state part (b) of the $4M-4$ conjecture is "If $N \geq 4M-4$, then there exists a real algebraic variety $V \subseteq \mathbb{R}^{2MN}$ such that $\mathcal{A}$ is injective for every $\Phi \notin V$." The work of Balan, Casazza and Edidin [3] already proves this for $N \geq 4M-2$. Furthermore, Bodmann and Hammen [5] establish that whenever $N \geq 4M-4$, there exists Φ such that $\mathcal{A}$ is injective, so for (b), it only remains to show that generic Φ make $\mathcal{A}$ injective.

The following results are given in the interest of resolving the $4M-4$ conjecture:

Theorem 10. *The $4M-4$ Conjecture is true when $M=2$.*

Since in this case injectivity is equivalent to having a full-rank super analysis operator (see Lemma 9), Theorem 10 can be established by defining the real algebraic variety

Algorithm 1 The HMW test for injectivity when $M = 3$

Input: Measurement vectors $\{\varphi_n\}_{n=1}^N \subseteq \mathbb{C}^3$
Output: Whether $\mathcal{A}$ is injective

```
Define A: H^{3×3} → R^N such that AH = {⟨H, φ_n φ_n^*⟩_HS}_{n=1}^N   {assemble the super analysis operator}
if dim null(A) = 0 then
    "INJECTIVE"                                                      {if A is injective, then A is injective}
else
    Pick H ∈ null(A), H ≠ 0
    if dim null(A) = 1 and det(H) ≠ 0 then
        "INJECTIVE"                                                  {if A only maps nonsingular matrices to zero, then A is injective}
    else
        "NOT INJECTIVE"                                              {in the remaining case, A maps differences of rank-1 matrices to zero}
    end if
end if
```

$V = \{\mathbf{A} : \operatorname{Re}\det\mathbf{A} = \operatorname{Im}\det\mathbf{A} = 0\}$ and showing that V^{c} is nonempty, and therefore dense with full measure. Before stating the analogous result for $M = 3$, we introduce the *HMW test* for injectivity (see Algorithm 1); we name it after Heinosaari, Mazarell and Wolf, who implicitly introduce this algorithm in their paper [10].

Theorem 11. *When $M = 3$, the HMW test correctly determines whether $\mathcal{A}$ is injective.*

The proof of Theorem 11 relies heavily on Lemma 9. For the case of $\dim\operatorname{null}(\mathbf{A}) = 2$, an application of the intermediate value theorem shows that a singular matrix of rank 1 or 2 can always be constructed from matrices in the null space of $\mathbf{A}$.

Theorem 12. *The $4M - 4$ Conjecture is true when $M = 3$.*

The proof of Theorem 12 first constructs the real algebraic variety V of matrices U, each gotten by a generalized cross product of a basis for the range of the adjoint of some $\mathbf{A}$, and further satisfying $\det U = 0$; the first part ensures that U spans the null space of $\mathbf{A}$, while at the same time being defined using polynomials of the entries of the matrix representation of $\mathbf{A}$. The HMW test is then used to show that V^{c} is nonempty.

Note that the HMW test can be used to test for injectivity in three dimensions regardless of the number of measurement vectors. Thus, it can be used to evaluate ensembles of 3×3 unitary matrices for quantum mechanics. For example, consider the 3×3 fractional discrete Fourier transform, defined in [6] using discrete Hermite-Gaussian functions:

$$F^\alpha = \frac{1}{6}\begin{bmatrix} 3+\sqrt{3} & \sqrt{3} & \sqrt{3} \\ \sqrt{3} & \frac{3-\sqrt{3}}{2} & \frac{3-\sqrt{3}}{2} \\ \sqrt{3} & \frac{3-\sqrt{3}}{2} & \frac{3-\sqrt{3}}{2} \end{bmatrix} + \frac{e^{\alpha \mathrm{i}\pi}}{6}\begin{bmatrix} 3-\sqrt{3} & -\sqrt{3} & -\sqrt{3} \\ -\sqrt{3} & \frac{3+\sqrt{3}}{2} & \frac{3+\sqrt{3}}{2} \\ -\sqrt{3} & \frac{3+\sqrt{3}}{2} & \frac{3+\sqrt{3}}{2} \end{bmatrix} + \frac{e^{\alpha \mathrm{i}\pi/2}}{2}\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & -1 & 1 \end{bmatrix}$$

It can be shown by the HMW test that $\Phi = [I\ F^{1/2}\ F\ F^{3/2}]$ lends injective intensity measurements. This leads to the following refinement of Wright's conjecture:

Conjecture 13. *Let F denote the $M \times M$ discrete fractional Fourier transform defined in [6]. Then for every $M \geq 3$, $\Phi = [I\ F^{1/2}\ F\ F^{3/2}]$ lends injective intensity measurements.*

Acknowledgments

The authors thank Profs. Bernhard G. Bodmann, Matthew Fickus, Thomas Strohmer and Yang Wang for insightful discussions, and the Erwin Schrödinger International Institute for Mathematical Physics for hosting a workshop on phase retrieval that helped solidify some of the ideas in this paper. The views expressed in this article are those of the authors and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the U.S. Government.

References

[1] B. Alexeev, A. S. Bandeira, M. Fickus, D. G. Mixon, Phase retrieval with polarization, Available online: arXiv:1210.7752

[2] B. Alexeev, J. Cahill, D. G. Mixon, Full spark frames, J. Fourier Anal. Appl. 18 (2012) 1167–1194.

[3] R. Balan, P. Casazza, D. Edidin, On signal reconstruction without phase, Appl. Comp. Harmon. Anal. 20 (2006) 345–356.

[4] A. Bandeira, J. Cahill, D. G. Mixon, A. A. Nelson, Saving phase: Injectivity and stability for phase retrieval, in preparation.

[5] B. G. Bodmann, N. Hammen, Stable phase retrieval with low-redundancy frames, Available online: arXiv:1302.5487

[6] Ç. Candan, M. A. Kutay, H. M. Ozaktas, The discrete fractional Fourier transform, IEEE Trans. Signal. Process. 48 (2000) 1329–1337.

[7] E. J. Candès, T. Strohmer, V. Voroninski, PhaseLift: Exact and stable signal recovery from magnitude measurements via convex programming, Available online: arXiv:1109.4499

[8] J. Finkelstein, Pure-state informationally complete and "really" complete measurements, Phys. Rev. A, 70 (2004) 052107.

[9] R. Hartshorne, Algebraic Geometry, Graduate Texts in Mathematics, Springer, New York, 1977.

[10] T. Heinosaari, L. Mazzarella, M. M. Wolf, Quantum tomography under prior information, Available online: arXiv:1109.5478

[11] M Püschel, J. Kovačević, Real, tight frames with maximal robustness to erasures, Proc. Data Compression Conf. (2005) 63–72.

[12] A. Vogt, Position and momentum distributions do not determine the quantum mechanical state, In: A. R. Marlow, ed., Mathematical Foundations of Quantum Theory, Academic Press, New York, 1978.

On transformations between Gabor frames and wavelet frames

Ole Christensen
Department of Applied Mathematics and Computer Science
Technical University of Denmark
Denmark
Email: Ole.Christensen@mat.dtu.dk

Say Song Goh
Department of Mathematics
National University of Singapore
Singapore 119260
Email: matgohss@nus.edu.sg

***Abstract*—We describe a procedure that enables us to construct dual pairs of wavelet frames from certain dual pairs of Gabor frames. Applying the construction to Gabor frames generated by appropriate exponential B-splines gives wavelet frames generated by functions whose Fourier transforms are compactly supported splines with geometrically distributed knot sequences. There is also a reverse transform, which yields pairs of dual Gabor frames when applied to certain wavelet frames.**

I. Introduction

In this note we will discuss a procedure that allows us to construct dual pairs of wavelet frames based on certain dual pairs of Gabor frames, and vice versa. Applying this to Gabor frames generated by exponential B-splines produces a class of attractive dual wavelet frame pairs generated by functions whose Fourier transform are compactly supported splines with geometrically distributed knots. Our main purpose here is to demonstrate the usefulness of the method; the proofs of the theoretical results are given in [2].

Let $\mathcal{H}$ be a separable Hilbert space. A sequence $\{f_i\}_{i\in I}$ in $\mathcal{H}$ is called a *frame* if there exist constants $A, B > 0$ such that

$$A\,||f||^2 \leq \sum_{i\in I} |\langle f, f_i\rangle|^2 \leq B\,||f||^2,\ \forall f \in \mathcal{H}. \qquad \text{(I.1)}$$

The constants A and B are *frame bounds*. The sequence $\{f_i\}_{i\in I}$ is a *Bessel sequence* if at least the upper bound in (I.1) is satisfied. A frame is *tight* if we can choose $A = B$ in (I.1). For any frame $\{f_i\}_{i\in I}$ there exists at least one *dual frame,* i.e., a frame $\{\widetilde{f}_i\}_{i\in I}$ for which

$$f = \sum_{i\in I} \langle f, f_i\rangle \widetilde{f}_i,\ \forall f \in \mathcal{H}.$$

We will consider Gabor frames and wavelet frames in the Hilbert space $L^2(\mathbb{R})$. A *Gabor system* in $L^2(\mathbb{R})$ has the form $\{e^{2\pi imbx} g(x-na)\}_{m,n\in\mathbb{Z}}$ for some parameters $a, b > 0$ and a given function $g \in L^2(\mathbb{R})$. Using the *translation operators* $T_a f(x) := f(x-a), a \in \mathbb{R}$, and the *modulation operators* $E_b f(x) := e^{2\pi ibx} f(x), b \in \mathbb{R}$, both acting on $L^2(\mathbb{R})$, we will denote a Gabor system by $\{E_{mb}T_{na}g\}_{m,n\in\mathbb{Z}}$. On the other hand, a *wavelet system* in $L^2(\mathbb{R})$ has the form $\{a^{j/2}\psi(a^j x - kb)\}_{j,k\in\mathbb{Z}}$ for some parameters $a > 1, b > 0$ and a given function $\psi \in L^2(\mathbb{R})$. Introducing the *scaling operators* $(D_a f)(x) := a^{1/2} f(ax)$, $a > 0$, acting on $L^2(\mathbb{R})$, the wavelet system can be written as $\{D_{a^j}T_{kb}\psi\}_{j,k\in\mathbb{Z}}$.

The duality conditions for a pair of Gabor systems were obtained by Ron & Shen [9], [10]. We state the formulation due to Janssen [8]:

Theorem 1.1: Given $b, \alpha > 0$, two Bessel sequences $\{E_{mb}T_{n\alpha}g\}_{m,n\in\mathbb{Z}}$ and $\{E_{mb}T_{n\alpha}\tilde{g}\}_{m,n\in\mathbb{Z}}$, where $g, \tilde{g} \in L^2(\mathbb{R})$, form dual Gabor frames for $L^2(\mathbb{R})$ if and only if for all $n \in \mathbb{Z}$,

$$\sum_{j\in\mathbb{Z}} \overline{g(x+j\alpha)}\tilde{g}(x+j\alpha+n/b) = b\delta_{n,0},\ a.e.\ x \in \mathbb{R}.$$

There are also characterizing equations for dual wavelet frames; see [5]. They are formulated in terms of the Fourier transform, for $f \in L^1(\mathbb{R})$ defined by $\widehat{f}(\gamma) := \int_{-\infty}^{\infty} f(x)e^{-2\pi i\gamma x}dx$, and extended to $L^2(\mathbb{R})$ in the usual way.

Theorem 1.2: Given $a > 1$, $b > 0$, two Bessel sequences $\{D_{a^j}T_{kb}\psi\}_{j,k\in\mathbb{Z}}$ and $\{D_{a^j}T_{kb}\widetilde{\psi}\}_{j,k\in\mathbb{Z}}$, where $\psi, \widetilde{\psi} \in L^2(\mathbb{R})$, form dual wavelet frames for $L^2(\mathbb{R})$ if and only if the following two conditions hold:

(i) $\sum_{j\in\mathbb{Z}} \overline{\widehat{\psi}(a^j\gamma)}\widehat{\widetilde{\psi}}(a^j\gamma) = b$ for a.e. $\gamma \in \mathbb{R}$.

(ii) For any number $\alpha \neq 0$ of the form $\alpha = m/a^j$, $m, j \in \mathbb{Z}$,

$$\sum_{(j,m)\in I_\alpha} \overline{\widehat{\psi}(a^j\gamma)}\widehat{\widetilde{\psi}}(a^j\gamma + m/b) = 0, \ a.e.\ \gamma \in \mathbb{R},$$

where $I_\alpha := \{(j,m) \in \mathbb{Z}^2 \mid \alpha = m/a^j\}$.

For more information on fundamental results of Gabor frames and wavelet frames, see, e.g., [1], [7], and [6].

II. From Gabor frames to wavelet frames

The goal of this section is to show how we can construct dual wavelet frame pairs based on certain dual Gabor frame pairs. The key is the following transform that allows us to move the Gabor structure into the wavelet structure.

Let $\theta > 1$ be given. Associated with a function $g \in L^2(\mathbb{R})$ for which $g(\log_\theta |\cdot|) \in L^2(\mathbb{R})$, we define a function $\psi \in L^2(\mathbb{R})$ by

$$\widehat{\psi}(\gamma) = \begin{cases} g(\log_\theta(|\gamma|)), & \text{if } \gamma \neq 0, \\ 0, & \text{if } \gamma = 0. \end{cases} \tag{II.1}$$

Note that by (II.1), for any $a > 0, j \in \mathbb{Z}$ and $\gamma \in \mathbb{R}\backslash\{0\}$,

$$\widehat{\psi}(a^j\gamma) = g(j\log_\theta(a) + \log_\theta(|\gamma|)). \tag{II.2}$$

Also, if $g \in L^2(\mathbb{R})$ is a bounded function with support in the interval $[M, N]$ for some $M, N \in \mathbb{R}$, then

$$\operatorname{supp}\widehat{\psi} \subseteq [-\theta^N, -\theta^M] \cup [\theta^M, \theta^N].$$

Note that (II.2) gives a convenient way to obtain functions ψ with the partition of unity property

$$\sum_{j\in\mathbb{Z}} \widehat{\psi}(a^j\gamma) = 1, \ \gamma \in \mathbb{R}. \tag{II.3}$$

Indeed, just take any function g satisfying the partition of unity condition

$$\sum_{j\in\mathbb{Z}} g(x + j) = 1, \ x \in \mathbb{R}, \tag{II.4}$$

and apply the construction in (II.1) with $\theta := a$. Comparing the corresponding conditions in Theorem 1.2(i) and Theorem 1.1, (II.3) provides a possible starting point for constructing dual wavelet frames, similar to (II.4) for dual Gabor frames, see, e.g., [3].

If g has compact support and is smooth, then the function $\widehat{\psi}$ in (II.1) is also smooth. Thus, by taking smooth functions g we obtain functions ψ with fast decay in the time domain.

A. Construction of dual pairs of wavelet frames

For fixed parameters $b, \alpha > 0$ we will consider two bounded compactly supported functions $g, \tilde{g} \in L^2(\mathbb{R})$ and the associated Gabor systems $\{E_{mb}T_{n\alpha}g\}_{m,n\in\mathbb{Z}}$ and $\{E_{mb}T_{n\alpha}\tilde{g}\}_{m,n\in\mathbb{Z}}$. For a fixed $\theta > 1$, define the functions $\psi, \widetilde{\psi} \in L^2(\mathbb{R})$ by (II.1) from $g, \tilde{g}$ respectively.

Theorem 2.1: Let $b > 0$, $\alpha > 0$, and $\theta > 1$ be given. Assume that $g, \tilde{g} \in L^2(\mathbb{R})$ are bounded functions with support in the interval $[M, N]$ for some $M, N \in \mathbb{R}$ and that $\{E_{mb}T_{n\alpha}g\}_{m,n\in\mathbb{Z}}$ and $\{E_{mb}T_{n\alpha}\tilde{g}\}_{m,n\in\mathbb{Z}}$ form dual frames for $L^2(\mathbb{R})$. With $a := \theta^\alpha$, if $b \leq \frac{1}{2\theta^N}$, then $\{D_{a^j}T_{kb}\psi\}_{j,k\in\mathbb{Z}}$ and $\{D_{a^j}T_{kb}\widetilde{\psi}\}_{j,k\in\mathbb{Z}}$ are dual frames for $L^2(\mathbb{R})$.

The proof follows from (II.2) and the characterizations of duality for Gabor frames and wavelet frames in Theorem 1.1 and Theorem 1.2.

If $g = \tilde{g}$ in Theorem 2.1, then $\psi = \widetilde{\psi}$, i.e., the result enables a tight wavelet frame to be constructed from a tight Gabor frame.

B. Explicit constructions

Based on Theorem 2.1, the rich theory for construction of dual pairs of Gabor frames enables us to provide explicit constructions of wavelet frame pairs.

Proposition 2.2: Let $g \in L^2(\mathbb{R})$ be a bounded real-valued function with support in the interval $[M, N]$ for some $M, N \in \mathbb{Z}$. Suppose that g satisfies the partition of unity condition (II.4). Let $a > 1$ and $b \in (0, \min(\frac{1}{2(N-M)-1}, 2^{-1}a^{-N})]$ be given, and take any real sequence $\{c_n\}_{n=-N+M+1}^{N-M-1}$ such that

$$c_0 = b, c_n + c_{-n} = 2b, \ n = 1, \ldots, N - M - 1.$$

Then the functions $\psi, \widetilde{\psi} \in L^2(\mathbb{R})$ defined by (II.1) and

$$\widehat{\widetilde{\psi}}(\gamma) = \sum_{n=-N+M+1}^{N-M-1} c_n g(\log_a(|\gamma|) + n), \ \gamma \neq 0, \tag{II.5}$$

generate dual wavelet frames $\{D_{a^j}T_{kb}\psi\}_{j,k\in\mathbb{Z}}$ and $\{D_{a^j}T_{kb}\widetilde{\psi}\}_{j,k\in\mathbb{Z}}$ for $L^2(\mathbb{R})$.

Proof. It follows from Theorem 3.1 in [3] that $\{E_{mb}T_n g\}_{m,n\in\mathbb{Z}}$ and the Gabor system generated by $\tilde{g}(x) = \sum_{n=-N+M+1}^{N-M-1} c_n g(x + n)$ form dual Gabor frames for $L^2(\mathbb{R})$ (the condition $b \leq \frac{1}{2(N-M)-1}$ is assumed in that result). Now the result follows from Theorem 2.1 with $\theta := a$. □

We will now consider a class of exponential B-splines that yields attractive dual pairs of wavelet frames, for which the Fourier transform of the generators are compactly supported splines with geometrically distributed knots and desired smoothness. These exponential splines are of the form

$$\mathcal{E}_N(\cdot) := e^{\beta_1(\cdot)}\chi_{[0,1]}(\cdot) * \cdots * e^{\beta_N(\cdot)}\chi_{[0,1]}(\cdot),$$

where $\beta_k = (k-1)\beta$, $k = 1, \ldots, N$, for some $\beta > 0$. Similar to the classical B-splines given by the choice $\beta_k = 0, k = 1, \ldots, N$, the exponential B-spline $\mathcal{E}_N$ is $N-2$ times differentiable (for $N \geq 2$) and its support is $[0, N]$. An explicit formula for $\mathcal{E}_N$ is given by Theorem 2.2 in [4] (note that there is a typo in the expression for $\mathcal{E}_N(x)$ for $x \in [k-1, k]$ on page 304 of [4]: the expression $e^{a_{j_1}} + \cdots + e^{a_{j_{k-1}}}$ should be $e^{a_{j_1}+\cdots+a_{j_{k-1}}}$). In Theorem 3.1 in the same paper, it is shown that for $N \geq 2$,

$$\sum_{k\in\mathbb{Z}} \mathcal{E}_N(x-k) = \frac{\prod_{m=1}^{N-1}\left(e^{\beta m}-1\right)}{\beta^{N-1}(N-1)!}. \qquad \text{(II.6)}$$

For the partition of unity constraint (II.4) to hold, we apply (II.6) and consider the function

$$g(x) := \frac{\beta^{N-1}(N-1)!}{\prod_{m=1}^{N-1}\left(e^{\beta m}-1\right)}\ \mathcal{E}_N(x).$$

Furthermore, let $a := e^{\beta}$. For $\gamma \neq 0$, using that $e^{\beta k \log_{e^\beta}(|\gamma|)} = |\gamma|^k$, we obtain from (II.1) an expression that identifies $\widehat{\psi}$ explicitly as a geometric spline, i.e., as a spline with geometrically distributed knots. Now the formula (II.5) yields a dual wavelet frame generator $\widetilde{\psi}$. Note that $\widehat{\widetilde{\psi}}$ is also a geometric spline.

Example 2.3: Consider the exponential B-spline $\mathcal{E}_3$ with $N = 3$ and $\beta = 1$. Then

$$\mathcal{E}_3(x) = \begin{cases} \frac{1-2e^x+e^{2x}}{2}, & x \in [0,1], \\ \frac{-(e+e^2)+2(e^{-1}+e)e^x-(e^{-2}+e^{-1})e^{2x}}{2}, & x \in [1,2], \\ \frac{e^3-2e^x+e^{-3}e^{2x}}{2}, & x \in [2,3], \\ 0, & x \notin [0,3]. \end{cases}$$

By (II.6) we have

$$\sum_{k\in\mathbb{Z}} \mathcal{E}_3(x-k) = \frac{1}{2}(e-1)(e^2-1),\ x \in \mathbb{R},$$

so we consider $g(x) := 2(e-1)^{-1}(e^2-1)^{-1}\mathcal{E}_3(x)$.

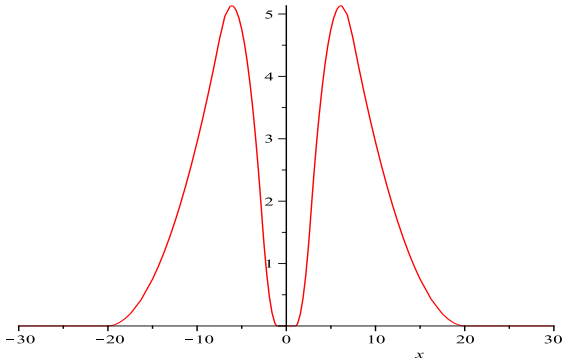


Fig. 1. Plot of the geometric spline $\widehat{\psi}$ in Example 2.3.

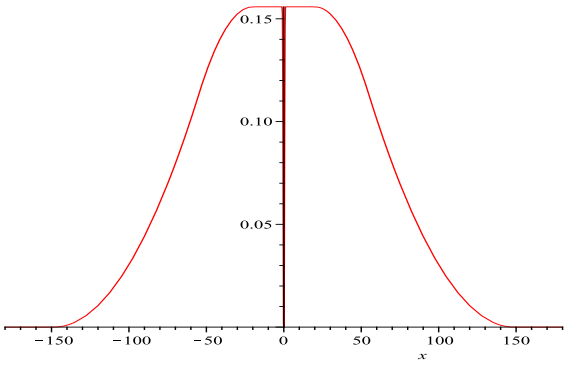


Fig. 2. Plot of the geometric spline $\widehat{\widetilde{\psi}}$ in Example 2.3.

Let $a := e^{\beta} = e$, and define the function ψ by

$$\widehat{\psi}(\gamma) = \begin{cases} \frac{1-2|\gamma|+\gamma^2}{(e-1)(e^2-1)}, & |\gamma| \in [1,e], \\ \frac{-(e+e^2)+2(e^{-1}+e)|\gamma|-(e^{-2}+e^{-1})\gamma^2}{(e-1)(e^2-1)}, & |\gamma| \in [e,e^2], \\ \frac{e^3-2|\gamma|+e^{-3}\gamma^2}{(e-1)(e^2-1)}, & |\gamma| \in [e^2,e^3], \\ 0, & |\gamma| \notin [1,e^3]. \end{cases}$$

The function $\widehat{\psi}$ is a geometric spline with knots at the points $\pm 1, \pm e, \pm e^2, \pm e^3$.

The construction in Proposition 2.2 works for $b \leq 2^{-1}e^{-3}$. Taking $b = 41^{-1}$ and $c_n = 41^{-1}$ for $n = -2, \ldots, 2$, it follows from (II.2) and (II.5) that the resulting dual frame generator $\widetilde{\psi}$ satisfies

$$\widehat{\widetilde{\psi}}(\gamma) = \frac{1}{41}\sum_{n=-2}^{2} \widehat{\psi}(e^n\gamma),\ \gamma \in \mathbb{R}.$$

The function $\widehat{\widetilde{\psi}}$ is a geometric spline with knots at the points $\pm e^{-2}, \pm e^{-1}, \pm 1, \pm e^3, \pm e^4, \pm e^5$.

Figures 1–3 show the graphs of the functions $\widehat{\psi}$ and $\widehat{\widetilde{\psi}}$, where Figure 3 re-plots part of the graph in Figure

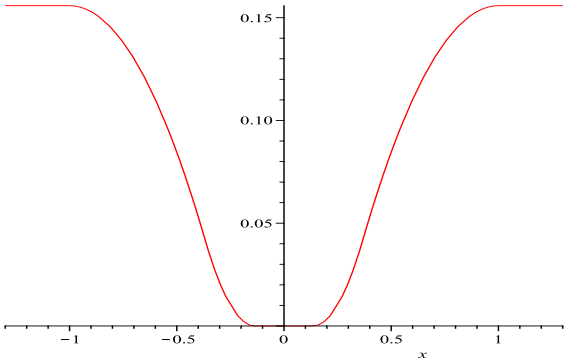


Fig. 3. Plot of the geometric spline $\widehat{\widetilde{\psi}}$ in Example 2.3 on the interval $[-1.3, 1.3]$.

2 on a smaller interval to better depict the behavior of $\widehat{\widetilde{\psi}}$ around 0. Note that $\widehat{\widetilde{\psi}}$ is constant on the support of $\widehat{\psi}$ and decays to zero outside this set. This is due to (II.6) and the special structure of $\widehat{\widetilde{\psi}}$ in (II.5). In fact, the same will occur when the construction is applied to any function whose integer-translates form a partition of unity. If higher order smoothness in $\widehat{\psi}$ and $\widehat{\widetilde{\psi}}$ is desired, this can be achieved if we use higher order exponential B-splines in the construction. □

III. From wavelet frames to Gabor frames

It is possible to reverse the process discussed so far, and obtain a way to obtain Gabor frames based on certain wavelet frames. Assume the functions $\psi, \widetilde{\psi} \in L^2(\mathbb{R})$ to be given. For a parameter $\theta > 1$ we define the functions $g, \tilde{g}$ by

$$g(x) := \widehat{\psi}(\theta^x), \quad \tilde{g}(x) := \widehat{\widetilde{\psi}}(\theta^x), \ x \in \mathbb{R}. \qquad \text{(III.1)}$$

The conditions below imply that $g, \tilde{g} \in L^2(\mathbb{R})$.

Theorem 3.1: Let $a > 1$ and $b > 0$. Assume that $\{D_{a^j} T_{kb} \psi\}_{j,k\in\mathbb{Z}}$ and $\{D_{a^j} T_{kb} \widetilde{\psi}\}_{j,k\in\mathbb{Z}}$ are dual frames for $L^2(\mathbb{R})$ and that the functions $\widehat{\psi}$ and $\widehat{\widetilde{\psi}}$ are supported in $[-L, -K] \cup [K, L]$ for some $K, L > 0$. Take $\theta > 1$ and $\alpha > 0$ such that $a = \theta^\alpha$. If $b \leq \frac{1}{\log_\theta(L/K)}$, then $\{E_{mb} T_{n\alpha} g\}_{m,n\in\mathbb{Z}}$ and $\{E_{mb} T_{n\alpha} \tilde{g}\}_{m,n\in\mathbb{Z}}$ form dual frames for $L^2(\mathbb{R})$.

Theorem 3.1 is proved using the characterizations of dual pairs of Gabor frames and wavelet frames in Theorem 1.1 and Theorem 1.2. Again, the result has an immediate consequence for construction of tight Gabor frames via tight wavelet frames.

The result can, e.g., be applied to the Meyer wavelet, which yields a construction of a tight Gabor frame generated by a $C^\infty(\mathbb{R})$, compactly supported function. Details of this are provided in [2].

Let us end this note with a short explanation of why we speak about (III.1) being a reverse transform of (II.1). If we start with a sufficiently well behaving function ψ and use the transform (III.1), we obtain the function $g(x) = \widehat{\psi}(\theta^x)$. Going "back" with the procedure in (II.1) applied on the function g, we arrive at the function

$$\widehat{\phi}(\gamma) = g(\log_\theta(|\gamma|)) = \widehat{\psi}(\theta^{\log_\theta(|\gamma|)}) = \widehat{\psi}(|\gamma|), \ \gamma \neq 0.$$

So, if the function $\widehat{\psi}$ is symmetric, we have that $\phi = \psi$.

On the other hand, starting with a function g and using (II.1), we obtain the function ψ, given by $\widehat{\psi}(\gamma) = g(\log_\theta(|\gamma|))$, $\gamma \neq 0$; applying the approach in (III.1) on $\widehat{\psi}$ leads to the function

$$h(x) = \widehat{\psi}(\theta^x) = g(\log_\theta(|\theta^x|)) = g(x), \ x \in \mathbb{R}.$$

Thus, we get the original function back.

References

[1] O. Christensen, An Introduction to Frames and Riesz Bases, Birkhäuser, Boston, 2003.
[2] O. Christensen, S.S. Goh, From dual pairs of Gabor frames to dual pairs of wavelet frames and vice versa, Appl. Comput. Harmon. Anal., to appear.
[3] O. Christensen, R.Y. Kim, On dual Gabor frame pairs generated by polynomials, J. Fourier Anal. Appl. 16 (2010) 1–16.
[4] O. Christensen, P. Massopust, Exponential B-splines and the partition of unity property, Adv. Comput. Math. 37 (2012) 301–318.
[5] C.K. Chui, X. Shi, Orthonormal wavelets and tight frames with arbitrary real dilations, Appl. Comput. Harmon. Anal. 9 (2000) 243–264.
[6] I. Daubechies, A. Grossmann, Y. Meyer, Painless nonorthogonal expansions, J. Math. Phys. 27 (1986) 1271–1283.
[7] K. Gröchenig, Foundations of Time-Frequency Analysis, Birkhäuser, Boston, 2000.
[8] A.J.E.M. Janssen, The duality condition for Weyl-Heisenberg frames, in: H.G. Feichtinger, T. Strohmer (Eds.), Gabor Analysis and Algorithms: Theory and Applications, Birkhäuser, Boston, 1998, pp. 33–84.
[9] A. Ron, Z. Shen, Frames and stable bases for shift-invariant subspaces of $L^2(\mathbb{R}^d)$, Can. J. Math. 47 (1995) 1051–1094.
[10] A. Ron, Z. Shen, Weyl-Heisenberg frames and Riesz bases in $L^2(\mathbb{R}^d)$, Duke Math. J. 89 (1997) 237–282.

Perfect Preconditioning of Frames by a Diagonal Operator

Gitta Kutyniok
Technische Universität Berlin
Institut für Mathematik
10623 Berlin, Germany
Email: kutyniok@math.tu-berlin.de

Kasso A. Okoudjou
University of Maryland
Department of Mathematics
College Park, MD 20742, USA
Email: kasso@math.umd.edu

Friedrich Philipp
Technische Universität Berlin
Institut für Mathematik
10623 Berlin, Germany
Email: philipp@math.tu-berlin.de

***Abstract*—Frames which are tight might be considered optimally conditioned in the sense of their numerical stability. This leads to the question of perfect preconditioning of frames, i.e., modification of a given frame to generate a tight frame. In this paper, we analyze prefect preconditioning of frames by a diagonal operator. We derive various characterizations of functional analytic and geometric type of the class of frames which allow such a perfect preconditioning.**

I. Introduction

Frames are nowadays a common methodology in applied mathematics, computer science, and engineering, see [7], when non-unique, but stable decompositions and expansions are required. They are utilized in various applications which can roughly be subdivided into two categories. One type of applications utilize frames for decomposing data. In this case, typical goals are erasure-resilient transmission, data analysis or processing, and compression, with the advantage of frames being their robustness as well as their flexibility in design. The other type of applications requires frames for expanding data. This approach is extensively used in sparsity methodologies such as Compressed Sensing (see [9]), but also, for instance, as systems generating trial spaces for PDE solvers. Again, it relies on non-uniqueness of the expansion which promotes sparse expansions and on the flexibility in design.

A crucial requirement for all such applications is the numerical stability of the associated algorithms, which is optimally ensured by the subclass of tight frames. Thus, urgent questions are: When can a given frame be modified to become a tight frame? Obviously, the most careful modification – which also retains properties such as providing sparse representations for a class of data – is to rescale each frame vector. Thus, in this paper, we consider the question: When can the vectors of a given frame be rescaled to obtain a tight frame?

A. Tight Frames

Before continuing, let us first fix the notions we will use throughout. Letting $\mathcal{H}$ be a real or complex separable Hilbert space and letting J be a subset of $\mathbb{N}$, a set of vectors $\Phi = \{\varphi_j\}_{j\in J} \subset \mathcal{H}$ is called a *frame* for $\mathcal{H}$, if there exist positive constants $A, B > 0$ (the *lower* and *upper frame bound*) such that

$$A\|x\|^2 \leq \sum_{j\in J} |\langle x, \varphi_j\rangle|^2 \leq B\|x\|^2 \quad \text{for all } x \in \mathcal{H}. \tag{1}$$

A frame Φ is referred to as *A-tight* or just *tight*, if $A = B$ is possible in (1), and *Parseval*, if $A = B = 1$ is possible. Moreover, if $|J| < \infty$ (which implies that $\mathcal{H} = \mathbb{K}^N$ with $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$), the frame Φ is called *finite*.

Let $\Phi = \{\varphi_j\}_{j\in J} \subset \mathcal{H}$ now be a frame for $\mathcal{H}$. Signals are analyzed using a frame by application of the associated *analysis operator* $T_\Phi : \mathcal{H} \to \ell^2(J)$ defined by $T_\Phi x := \big(\langle x, \varphi_j\rangle\big)_{j\in J}$. Its adjoint T_Φ^*, the *synthesis operator* of Φ, maps then $\ell^2(J)$ surjectively onto $\mathcal{H}$. Concatenating both operators leads to the *frame operator* $S_\Phi := T_\Phi^* T_\Phi$ of Φ, given by

$$S_\Phi x = \sum_{j\in J} \langle x, \varphi_j\rangle \varphi_j, \quad x \in \mathcal{H},$$

which is a bounded and strictly positive selfadjoint operator in $\mathcal{H}$. These properties imply that Φ admits the reconstruction formula

$$x = \sum_{j\in J} \langle x, \varphi_j\rangle S_\Phi^{-1}\varphi_j \quad \text{for all } x \in \mathcal{H}.$$

To avoid numerical stability issues, it seems desirable to have $S_\Phi = const \cdot I_\mathcal{H}$ ($I_\mathcal{H}$ denoting the identity on $\mathcal{H}$, for $\mathcal{H} = \mathbb{K}^N$ we will use I_N). And in fact, this equation characterizes tight frames. Thus an A-tight frame admits the numerically optimally stable reconstruction given by

$$x = A^{-1} \cdot \sum_{j\in J} \langle x, \varphi_j\rangle \varphi_j \quad \text{for all } x \in \mathcal{H}.$$

B. Generating Parseval Frames

Since applications typically require specific frames, which might not automatically form a tight frame, an important problem is to introduce approaches for modifying a given frame in order to generate a tight frame. We might restrict our attention to generating Parseval frames, since this just requires a renormalization once we derived a tight frame. One key issue in this whole process is to modify the frame as careful as possible to not disturb its frame properties – which might be crucial for the application at hand – too much. As an example,

think for instance of a frame which sparsifies a given test data set; a property one might want to keep.

A very common approach to generate a tight frame is to apply $S_\Phi^{-1/2}$ to each frame vector of a frame Φ, which in fact even yields a Parseval frame. However, this modification also changes the frame properties significantly, and to date it is still entirely unclear in which way. In particular, a sparse representation property would be completely destroyed.

The most careful modification of a frame is scaling its frame vectors. For instance, this procedure even preserves any sparse representation properties of the frame. In [12], a frame was coined *scalable*, if a scaling exists which leads to a Parseval frame. Notice that this notion is weakly related to the notion of signed frames, weighted frames as well as controlled frames (see, e.g., [13], [1], [14]).

Evidently, not every frame is scalable. For instance, a basis in $\mathbb{R}^2$ which is not an orthogonal basis is not scalable, since a frame with two elements in $\mathbb{R}^2$ is a Parseval frame if and only if it is an orthonormal basis. The relation to preconditioning is revealed by analyzing the finite-dimensional version of Proposition II.3 which shows that a frame Φ in $\mathbb{K}^N$ with analysis operator T_Φ is scalable if and only if there exists a diagonal matrix D such that DT_Φ is isometric. Since the condition number of such a matrix equals one, the scaling question is a particular instance of the problem of preconditioning of matrices.

The results in [12] were the leadoff results on this problem, which we present a survey about in this paper. The derived characterizations can be subdivided into the following two classes:

- Various characterizations of (strict) scalability of a frame for a general separable Hilbert space (see, e.g., Theorem II.5).
- Geometric characterization of scalability of finite frames (see Theorems III.1 and III.4).

We wish to note that recently, new results on this question from a slightly different angle have been derived in [6].

C. An Excursion to Numerical Linear Algebra

The problem of preconditioning is extensively studied in the numerical linear algebra community, see, e.g., [8], [10]. Preconditioners which are constructed by scaling appears in various forms in the numerical linear algebra literature. The most common approach is to minimize the condition number of the matrix multiplied by a preconditioning matrix – in our case of DT_Φ, where D runs through the set of diagonal matrices. It was for instance shown in [4], that this minimization problem can be reformulated as a convex problem. A major problem is however (see also [4]) that all algorithms solving this convex problem perform slowly, and, even worse, there exist situations in which the infimum is not attained. As additional references to this complex problem, we wish to mention [5], [2], [8], [11] and [15].

D. Outline

This paper is organized as follows. In Section II we focus on the situation of general separable Hilbert spaces and derive characterization of scalability and strict scalability. In Section III we then restrict to the situation of finite frames, and derive a yet different characterization of scalability as well as a geometric interpretation of scalable frames.

II. Strict Scalability of General Frames

This section is devoted to a very general characterization of (strictly) scalable frames.

A. Scalability and Frame Properties

The following definition makes the notion of scalability mathematically precise.

Definition II.1. A frame $\Phi = \{\varphi_j\}_{j\in J}$ for $\mathcal{H}$ is called *scalable* if there exist scalars $c_j \geq 0$, $j \in J$, such that $\{c_j\varphi_j\}_{j\in J}$ is a Parseval frame. If, in addition, $c_j > 0$ for all $j \in J$, then Φ is called *positively scalable*. If there exists $\delta > 0$, such that $c_j \geq \delta$ for all $j \in J$, then Φ is called *strictly scalable*.

For finite frames, it is immediate that positive and strict scalability coincide and that each scaling $\{c_j\varphi_j\}_{j\in J}$ of a finite frame $\{\varphi_j\}_{j\in J}$ with positive scalars c_j forms again a frame.

For infinite frames, the situation is significantly more involved. A partial answer was given in [1, Lemma 4.3], which proves that if there exist $K_1, K_2 > 0$ such that $K_1 \leq c_j \leq K_2$ holds for all $j \in J$, then also $\{c_j\varphi_j\}_{j\in J}$ is a frame. Our next result provides a complete characterization of when a scaling preserves the frame property. A crucial ingredient for this result is the *diagonal operator* D_c in $\ell^2(J)$ corresponding to a sequence $c = (c_j)_{j\in J} \subset \mathbb{K}$, which is defined by

$$D_c(v_j)_{j\in J} := \big(c_j v_j\big)_{j\in J}, \quad (v_j)_{j\in J} \in \operatorname{dom} D_c,$$

where

$$\operatorname{dom} D_c := \big\{(v_j)_{j\in J} \in \ell^2(J) : (c_j v_j)_{j\in J} \in \ell^2(J)\big\}.$$

It is a well-known fact that D_c is a (possibly unbounded) selfadjoint operator in $\ell^2(J)$ if and only if $c_j \in \mathbb{R}$ for all $j \in J$. If even $c_j \geq 0$ ($c_j > 0$, $c_j \geq \delta > 0$) for each $j \in J$, then the selfadjoint operator D_c is non-negative (positive, strictly positive, respectively).

The following result indeed provides a complete characterization of when a scaled frame constitutes again a frame. For stating this, as usual, we denote the domain, the kernel and the range of a linear operator T by $\operatorname{dom} T$, $\ker T$ and $\operatorname{ran} T$, respectively. Also, a closed linear operator T between two Hilbert spaces $\mathcal{H}$ and $\mathcal{K}$ will be called *ICR* (or an *ICR-operator*), if it is injective and has a closed range, i.e., if there exists $\delta > 0$ such that $\|Tx\| \geq \delta\|x\|$ for all $x \in \operatorname{dom} T$.

Proposition II.2 ([12])**.** *Let $\Phi = \{\varphi_j\}_{j\in J}$ be a frame for $\mathcal{H}$ with analysis operator T_Φ and let $c = (c_j)_{j\in J}$ be a sequence of non-negative scalars. Then the following conditions are equivalent.*

(i) *The scaled sequence of vectors* $\Psi := \{c_j\varphi_j\}_{j\in J}$ *is a frame for* $\mathcal{H}$.
(ii) *We have* $\operatorname{ran} T_\Phi \subset \operatorname{dom} D_c$ *and* $D_c|\operatorname{ran} T_\Phi$ *is ICR.*

Moreover, in this case, the frame operator of the frame Ψ *is given by*

$$S_\Psi = (D_cT_\Phi)^*(D_cT_\Phi) = \overline{T_\Phi^*D_c}D_cT_\Phi,$$

where $\overline{T_\Phi^*D_c}$ *denotes the closure of the operator* $T_\Phi^*D_c$.

B. General Equivalent Condition

The following result seems to be quite obvious. However, in the general setting of an arbitrary separable Hilbert space, it is not straightforward at all.

Proposition II.3 ([12])**.** *Let* $\Phi = \{\varphi_j\}_{j\in J}$ *be a frame for* $\mathcal{H}$. *Then the following conditions are equivalent.*

(i) Φ *is (positively, strictly) scalable.*
(ii) *There exists a non-negative (positive, strictly positive, respectively) diagonal operator* D *in* $\ell^2(J)$ *such that*

$$\overline{T_\Phi^*D}DT_\Phi = I_\mathcal{H}. \tag{2}$$

We can now easily draw the conclusion that scalability is invariant under unitary transformations.

Corollary II.4. *Let* U *be a unitary operator in* $\mathcal{H}$. *Then a frame* $\Phi = \{\varphi_j\}_{j\in J}$ *for* $\mathcal{H}$ *is scalable if and only if the frame* $U\Phi = \{U\varphi_j\}_{j\in J}$ *is scalable.*

C. Main Result

Our main result provides several equivalent conditions for a frame Φ to be strictly scalable. For this, recall that a sequence $\{v_k\}_k$ of non-zero vectors in a Hilbert space $\mathcal{K}$ is called an *orthogonal basis* of $\mathcal{K}$, if $\inf_k \|v_k\| > 0$ and $(v_k/\|v_k\|)_k$ is an orthonormal basis of $\mathcal{K}$.

Theorem II.5 ([12])**.** *Let* $\Phi = \{\varphi_j\}_{j\in J}$ *be a frame for* $\mathcal{H}$ *such that* $\liminf_{j\in J}\|\varphi_j\| > 0$, *and let* $T = T_\Phi$ *denote its analysis operator. Then the following statements are equivalent.*

(i) *The frame* Φ *is strictly scalable.*
(ii) *There exists a strictly positive bounded diagonal operator* D *in* $\ell^2(J)$ *such that* DT *is isometric (that is,* $T^*D^2T = I_\mathcal{H}$*).*
(iii) *There exist a Hilbert space* $\mathcal{K}$ *and a bounded ICR operator* $L : \mathcal{K} \to \ell^2(J)$ *such that* $TT^* + LL^*$ *is a strictly positive bounded diagonal operator.*
(iv) *There exist a Hilbert space* $\mathcal{K}$ *and a frame* $\Psi = \{\psi_j\}_{j\in J}$ *for* $\mathcal{K}$ *such that the vectors*

$$\varphi_j \oplus \psi_j \in \mathcal{H}\oplus\mathcal{K}, \quad j \in J,$$

form an orthogonal basis of $\mathcal{H}\oplus\mathcal{K}$.

If one of the above conditions holds, then the frame Ψ *from* (iv) *is strictly scalable, its analysis operator is given by an operator* L *from* (iii)*, and with a diagonal operator* D *from* (ii) *we have*

$$L^*D^2L = I_\mathcal{K}, \quad \text{and} \quad L^*D^2T = 0. \tag{3}$$

We next analyze this result in the special case of finite frames. Although this restriction seems trivial, in fact restricting conditions (iii) and (iv) in Theorem II.5 to the situation of finite frames is not immediate.

Corollary II.6. *Let* $\Phi = \{\varphi_j\}_{j=1}^M$ *be a frame for* $\mathbb{K}^N$ *and let* $T = T_\Phi \in \mathbb{K}^{M\times N}$ *denote the matrix representation of its analysis operator. Then the following statements are equivalent.*

(i) *The frame* Φ *is strictly scalable.*
(ii) *There exists a positive definite diagonal matrix* $D \in \mathbb{K}^{M\times M}$ *such that* DT *is isometric.*
(iii) *There exists* $L \in \mathbb{K}^{M\times(M-N)}$ *such that* $TT^* + LL^*$ *is a positive definite diagonal matrix.*
(iv) *There exists a frame* $\Psi = \{\psi_j\}_{j=1}^M$ *for* $\mathbb{K}^{M-N}$ *such that* $\{\varphi_j\oplus\psi_j\}_{j=1}^M \in \mathbb{K}^M$ *forms an orthogonal basis of* $\mathbb{K}^M$.

III. Scalability of Real Finite Frames

Finally, we take a geometric viewpoint with respect to scalability. For this, we will focus on frames for $\mathbb{R}^N$ due to the fact that the proof of Theorem III.1 requires the utilization of Farkas' Lemma which only exists for real vector spaces.

A. Characterization Result

The following theorem provides a characterization of non-scalability of a finite frame specifically tailored to the finite-dimensional case. Condition (iii) of this result will be reinterpreted in Subsection III-B as a geometric condition for non-scalability.

Theorem III.1 ([12])**.** *Let* $\Phi = \{\varphi_j\}_{j=1}^M \subset \mathbb{R}^N \setminus \{0\}$ *be a frame for* $\mathbb{R}^N$. *Then the following statements are equivalent.*

(i) Φ *is not scalable.*
(ii) *There exists a symmetric matrix* $Y \in \mathbb{R}^{N\times N}$ *with* $\operatorname{tr}(Y) < 0$ *such that* $\varphi_j^TY\varphi_j \geq 0$ *for all* $j = 1,\dots,M$.
(iii) *There exists a symmetric matrix* $Y \in \mathbb{R}^{N\times N}$ *with* $\operatorname{tr}(Y) = 0$ *such that* $\varphi_j^TY\varphi_j > 0$ *for all* $j = 1,\dots,M$.

The following corollary, for whose proof we refer to [12], can be easily drawn from the previous result, showing that the set of non-scalable frames for $\mathbb{R}^N$ is open in the following sense.

Corollary III.2. *Let* $\Phi = \{\varphi_j\}_{j=1}^M \subset \mathbb{R}^N \setminus \{0\}$ *be a frame for* $\mathbb{R}^N$ *which is not scalable. Then there exists* $\varepsilon > 0$ *such that each set of vectors* $\{\psi_j\}_{j=1}^M \subset \mathbb{R}^N$ *with*

$$\|\varphi_j - \psi_j\| < \varepsilon \quad \text{for all } j = 1,\dots,M \tag{4}$$

is a frame for $\mathbb{R}^N$ *which is not scalable.*

B. Geometric Interpretation

We now derive a geometric interpretation of the characterization result Theorem III.1, in particular of condition (iii). For this, first notice that each of the sets

$$C(Y) := \{x \in \mathbb{R}^N : x^TYx > 0\}, \quad Y \in \mathbb{R}^{N\times N} \text{ symmetric},$$

considered in Theorem III.1 (iii) forms an open cone with the additional property that $x \in C(Y)$ implies $-x \in C(Y)$.

Hence, from now on, we will analyze the impact of the condition $\mathrm{tr}(Y) = 0$ on the shape of these cones.

We will require the following particular class of conical surfaces. Their special relation to quadrics inspired us to coin those 'conical zero-trace quadrics'.

Definition III.3. Let the *class of conical zero-trace quadrics* $\mathcal{C}_N$ be defined as the family of sets

$$\left\{ x \in \mathbb{R}^N : \sum_{k=1}^{N-1} a_k \langle x, e_k \rangle^2 = \langle x, e_N \rangle^2 \right\}, \tag{5}$$

where $\{e_k\}_{k=1}^N$ runs through all orthonormal bases of $\mathbb{R}^N$ and $(a_k)_{k=1}^{N-1}$ runs through all tuples of elements in $\mathbb{R} \setminus \{0\}$ with $\sum_{k=1}^{N-1} a_k = 1$.

Utilizing this notion, we can state the following result on a geometric characterization of non-scalability.

Theorem III.4 ([12])**.** *Let $\Phi \subset \mathbb{R}^N \setminus \{0\}$ be a frame for $\mathbb{R}^N$. Then the following conditions are equivalent.*

(i) *Φ is not scalable.*
(ii) *All frame vectors of Φ are contained in the interior of a conical zero-trace quadric of $\mathcal{C}_N$.*
(iii) *All frame vectors of Φ are contained in the exterior of a conical zero-trace quadric of $\mathcal{C}_N$.*

By $\mathcal{C}_N^*$ we denote the subclass of $\mathcal{C}_N$ consisting of all conical zero-trace quadrics in which the orthonormal basis is the standard basis of $\mathbb{R}^N$. Thus, the elements of $\mathcal{C}_N^*$ are in fact quadrics of the form

$$\left\{ x \in \mathbb{R}^N : \sum_{k=1}^{N-1} a_k x_k^2 = x_N^2 \right\}.$$

with non-zero a_k's satisfying $\sum_{k=1}^{N-1} a_k = 1$.

This allows us to draw the following corollary from Theorem III.4 and Corollary II.4.

Corollary III.5. *Let $\Phi \subset \mathbb{R}^N \setminus \{0\}$ be a frame for $\mathbb{R}^N$. Then the following conditions are equivalent.*

(i) *Φ is not scalable.*
(ii) *There exists an orthogonal matrix $U \in \mathbb{R}^{N \times N}$ such that all vectors of $U\Phi$ are contained in the interior of a conical zero-trace quadric of $\mathcal{C}_N^*$.*
(iii) *There exists an orthogonal matrix $U \in \mathbb{R}^{N \times N}$ such that all vectors of $U\Phi$ are contained in the exterior of a conical zero-trace quadric of $\mathcal{C}_N^*$.*

Finally, in the 2- and 3-dimensional case Theorem III.4 reduces to the following results.

Corollary III.6. (i) *A frame $\Phi \subset \mathbb{R}^2 \setminus \{0\}$ for $\mathbb{R}^2$ is not scalable if and only if there exists an open quadrant cone which contains all frame vectors of Φ.*
(ii) *A frame $\Phi \subset \mathbb{R}^3 \setminus \{0\}$ for $\mathbb{R}^3$ is not scalable if and only if all frame vectors of Φ are contained in the interior of an elliptical conical surface with vertex 0 and intersecting the corners of a rotated unit cube.*

ACKNOWLEDGMENT

G. Kutyniok acknowledges support by the Einstein Foundation Berlin, by Deutsche Forschungsgemeinschaft (DFG) Grant SPP-1324 KU 1446/13 and DFG Grant KU 1446/14, by the DFG Collaborative Research Center TRR 109 "Discretization in Geometry and Dynamics", and by the DFG Research Center MATHEON "Mathematics for Key Technologies" in Berlin. F. Philipp is supported by the DFG Research Center MATHEON. K. A. Okoudjou was supported by ONR grants N000140910324 and N000140910144, by a RASA from the Graduate School of UMCP and by the Alexander von Humboldt foundation. He would also like to express his gratitude to the Institute for Mathematics at the University of Osnabrück for its hospitality while part of this work was completed.

REFERENCES

[1] P. Balazs, J.-P. Antoine, and A. Grybos, *Weighted and controlled frames: Mutual relationship and first numerical properties,* Int. J. Wavelets Multiresolut. Inf. Process. **8** (2010), 109–132.
[2] V. Balakrishnan, and S. Boyd, *Existence and uniqueness of optimal matrix scalings,* SIAM J. Matrix Anal. Appl. **16** (1995), 29–39.
[3] L. D. Berkovitz, *Convexity and optimization in $\mathbb{R}^n$,* John Wiley & Sons, Inc., New York, 2002.
[4] R. D. Braatz and M. Morari, *Minimizing the Euclidian condition number,* SIAM J. Control Optim. **32** (1994), 1763–1768.
[5] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear matrix inequalities in system and control theory,* SIAM Studies in Applied Mathematics. 15. Philadelphia, PA: SIAM, Society for Industrial and Applied Mathematics. xii, 1994.
[6] J. Cahill and X. Chen, *A note on scalable frames*, preprint.
[7] P. G. Casazza and G. Kutyniok, eds., *Finite frames: Theory and applications,* Birkhäuser Boston, Inc., Boston, MA, 2012.
[8] K. Chen, *Matrix preconditioning techniques and applications,* Cambridge Monographs on Applied and Computational Mathematics 19. Cambridge: Cambridge University Press xxiii, 2005.
[9] Y. C. Eldar and G. Kutyniok, *Compressed Sensing: Theory and applications,* Cambridge University Press, Cambridge, UK, 2012.
[10] I. Faragó and J. Karátson, *Numerical solution of nonlinear elliptic problems via preconditioning operators,* Nova Science Publishers, Inc., New York, 2002.
[11] L. Y. Kolotilina, *Solution of the problem of optimal diagonal scaling for quasi-real Hermitian positive-definite 3×3 matrices,* Zap. Nauchn. Semin. POMI **309** (2004), 84–126, 191–192; translation in J. Math. Sci. **132** (2006), 190–213.
[12] G. Kutyniok, K. A. Okoudjou, F. Philipp, and E. K. Tuley, *Scalable Frames,* Linear Algebra Appl. **438** (2013), 2225-2238.
[13] I. Peng and S. Waldron, *Signed frames and Hadamard products of Gram matrices,* Linear Algebra Appl. **347** (2002), 131–157.
[14] A. Rahimi and P. Balazs, *Multipliers for p-Bessel sequences in Banach spaces,* Integral Equations Oper. Theory **68** (2010), 193–205.
[15] A. Shapiro, *Upper bounds for nearly optimal diagonal scaling of matrices,* Linear Multilinear Algebra **29** (1991), 145–147.

Characterizing Completions of Finite Frames

Matthew Fickus
Department of Mathematics and Statistics
Air Force Institute of Technology
Wright-Patterson Air Force Base, Ohio 45433, USA
Email: Matthew.Fickus@gmail.com

Miriam Poteet
Department of Mathematics and Statistics
Air Force Institute of Technology
Wright-Patterson Air Force Base, Ohio 45433, USA

***Abstract*—Finite frames are possibly-overcomplete generalizations of orthonormal bases. We consider the "frame completion" problem, that is, the problem of how to add vectors to an existing frame in order to make it better conditioned. In particular, we discuss a new, complete characterization of the spectra of the frame operators that arise from those completions whose newly-added vectors have given prescribed lengths. To do this, we build on recent work involving a frame's eigensteps, namely the interlacing sequence of spectra of its partial frame operators. We discuss how such eigensteps exist if and only if our prescribed lengths are majorized by another sequence which is obtained by comparing our completed frame's spectrum to our initial one.**

I. Introduction

Let M and N be positive integers, and let $\{\varphi_n\}_{n=1}^N$ be a finite sequence of vectors in $\mathbb{C}^M$. The corresponding *synthesis operator* is the $M \times N$ matrix $\Phi = [\,\varphi_1 \cdots \varphi_N\,]$ obtained by stacking these vectors as columns. Multiplying this matrix by its adjoint (conjugate-transpose) Φ^* yields the $N \times N$ *Gram matrix* $\Phi^*\Phi$ as well as the $M \times M$ *frame operator* $\Phi\Phi^*$:

$$\Phi\Phi^* x = \left(\sum_{n=1}^N \varphi_n \varphi_n^*\right) x = \sum_{n=1}^N \langle x, \varphi_n\rangle \varphi_n.$$

Note that when $\{\varphi_n\}_{n=1}^N$ is an orthonormal basis for $\mathbb{C}^M$, we have $M = N$ and $\Phi^*\Phi = \mathrm{I} = \Phi\Phi^*$. In this case, the above expression gives the traditional orthonormal expansion of x.

Frame theory generalizes the notion of an orthonormal basis in order to provide possibly-overcomplete (nonorthogonal) expansions of x. It does this by relaxing Parseval's identity. To be precise, $\{\varphi_n\}_{n=1}^N$ is a *frame* for $\mathbb{C}^M$ if there exist *lower* and *upper frame bounds* $0 < A \leq B < \infty$ such that

$$A\|x\|^2 \leq \sum_{n=1}^N |\langle x, \varphi_n\rangle|^2 \leq B\|x\|^2, \quad \forall x \in \mathbb{C}^M. \qquad (1)$$

In this finite-dimensional setting, one can show that the optimal frame bounds A and B of any $\{\varphi_n\}_{n=1}^N$ are the least and greatest eigenvalues of $\Phi\Phi^*$, respectively. In particular, when $\{\varphi_n\}_{n=1}^N$ is a frame for $\mathbb{C}^M$, we have that $\Phi\Phi^*$ is invertible, having condition number at most B/A. This enables us to define the *canonical dual frame* $\{\tilde{\varphi}_n\}_{n=1}^N$, $\tilde{\varphi}_n := (\Phi\Phi^*)^{-1}\varphi_n$. Together, a frame and its dual provide the decompositions:

$$x = \sum_{n=1}^N \langle x, \tilde{\varphi}_n\rangle \varphi_n = \sum_{n=1}^N \langle x, \varphi_n\rangle \tilde{\varphi}_n, \quad \forall x \in \mathbb{C}^M.$$

In recent years, these "painless nonorthogonal expansions" have been exploited in a variety of finite-dimensional signal processing applications in which redundancy is useful [5].

Much of the recent research on finite frames has focused on constructing frames that satisfy a given list of desired, application-motivated constraints. Sometimes these constraints are nonlinear. For example, we often want our frames to be *tight*, namely have $A = B$ in (1), which happens precisely when $\Phi\Phi^* = A\mathrm{I}$. Tightness ensures that Φ is as well-conditioned as possible, and makes it easy to compute the canonical dual: $\tilde{\varphi}_n = \frac{1}{A}\varphi_n$. Moreover, finite tight frames are easy to construct: we simply need the rows of Φ to be orthogonal and have constant norm. In short, tight frames behave much more like orthonormal bases than frames do in general, while still permitting overcompleteness.

In order to find overcomplete frames which are even more faithful to the concept of an orthonormal basis, we can further restrict ourselves to *unit norm tight frames* (UNTFs), that is, tight frames $\{\varphi_n\}_{n=1}^N$ for $\mathbb{C}^M$ that have the additional property that $\|\varphi_n\| = 1$ for all n. Whereas the synthesis operator Φ of an orthonormal basis satisfies $\Phi\Phi^* = \mathrm{I} = \Phi^*\Phi$, a UNTF instead has that $\Phi\Phi^* = A\mathrm{I}$ and that the diagonal entries of $\Phi^*\Phi$ are 1; the fact that these two matrices have the same trace implies A is necessarily $\frac{N}{M}$.

UNTFs are known to exist for every $N \geq M$. For example, one may form Φ by extracting M rows from an $N \times N$ discrete Fourier transform matrix. However, the problem of constructing *every* UNTF was open for many years, due to the fact that the entries of Φ must satisfy a large system of intertwined quadratic equations. This problem was recently solved in [1] and [3]. In fact, as detailed in the next section, [1] and [3] give an explicit, closed-form algorithm for constructing every sequence of vectors $\{\varphi_n\}_{n=1}^N \subseteq \mathbb{C}^M$ whose frame operator $\Phi\Phi^*$ has a given spectrum $\{\lambda_m\}_{m=1}^M$ and whose Gram matrix $\Phi^*\Phi$ has diagonal entries $\{\mu_n\}_{n=1}^N$.

In this paper, we outline recent results from [4] and [7] that generalize the techniques of [1] and [3] to address the problem of *frame completions*. To be precise, given some initial sequence of vectors $\{\varphi_n\}_{n=1}^N \subseteq \mathbb{C}^M$ and some desired lengths $\{\mu_{N+p}\}_{p=1}^P$, we consider the problem of *completing* $\{\varphi_n\}_{n=1}^N$ by adding P new vectors $\{\varphi_{N+p}\}_{p=1}^P$ to this collection with the property that $\|\varphi_{N+p}\|^2 = \mu_{N+p}$ for all p.

We, like several other teams of researchers, are interested in the best (tightest) possible completions. Several cases of

the optimal frame completion problem have already been solved, such as the case where the lengths permit a tight completion [2], and the case where all the added vectors have equal length [6]. To our knowledge, the general case of this problem (arbitrary lengths, tightness unobtainable) remains open. Our work here serves to characterize *every* possible completion that can be formed using a given sequence of lengths. Our longer-term goal is to use this characterization in order to find the optimal completion in the general case.

II. Eigensteps

Let M and N be any positive integers and let $\{\lambda_m\}_{m=1}^M$ and $\{\mu_n\}_{n=1}^N$ be any nonnegative nonincreasing sequences. Recent work given in [1] and [3] provides a method for explicitly constructing every finite sequence of vectors $\{\varphi_n\}_{n=1}^N$ in $\mathbb{C}^M$ whose frame operator $\Phi\Phi^*$ has spectrum $\{\lambda_m\}_{m=1}^M$ and whose vectors have lengths $\|\varphi_n\|^2 = \mu_n$ for all n. This method is based on the concept of *eigensteps*. To be precise, given any such frame and any $k = 0, \ldots, N$, let $\{\lambda_{k;m}\}_{m=1}^M$ denote the spectrum of its kth *partial frame operator*

$$\Phi_k\Phi_k^* = \sum_{n=1}^k \varphi_n\varphi_n^*. \tag{2}$$

In practice, we arrange these values in an $M \times (N+1)$ table:

$$\begin{bmatrix} \lambda_{0;M} & \lambda_{1;1} & \cdots & \lambda_{N;M} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{0;1} & \lambda_{1;M} & \cdots & \lambda_{N;1} \end{bmatrix}. \tag{3}$$

In order to arise from a sequence $\{\varphi_n\}_{n=1}^N$ whose frame operator has spectrum $\{\lambda_m\}_{m=1}^M$ and whose elements have lengths $\|\varphi_n\|^2 = \mu_n$, the values in this table necessarily satisfy four rules. First, for $k = N$, we have $\Phi_N\Phi_N^* = \Phi\Phi^*$ and so $\lambda_{N;m} = \lambda_m$ for all m. This means the last column of (3) corresponds to our final desired spectrum. Our second rule comes from the fact that when $k = 0$, we regard the empty sum defining $\Phi_0\Phi_0^*$ to be a matrix of zeros, and so $\lambda_{0;m} = 0$ for all m. This means the first column of (3) is all zeros.

The third rule is that for any k, the sum of the entries in the kth column of (3) is necessarily the sum of the first k of our μ_n's; this follows from the fact that

$$\sum_{m=1}^M \lambda_{k;m} = \mathrm{Tr}(\Phi_k\Phi_k^*) = \mathrm{Tr}(\Phi_k^*\Phi_k) = \sum_{n=1}^k \mu_n.$$

The fourth rule is the least obvious. For any $k = 1, \ldots, N$, note that the kth partial frame operator is the sum of the previous one with an outer product: $\Phi_k\Phi_k^* = \Phi_{k-1}\Phi_{k-1}^* + \varphi_k\varphi_k^*$. As such, a classical result from matrix analysis tells us that the spectrum of $\Phi_k\Phi_k^*$ necessarily interlaces on that of $\Phi_{k-1}\Phi_{k-1}^*$. To be precise, we say that a finite sequence of real numbers $\{\gamma_m\}_{m=1}^M$ *interlaces* on another such sequence $\{\beta_m\}_{m=1}^M$, denoted $\{\beta_m\}_{m=1}^M \sqsubseteq \{\gamma_m\}_{m=1}^M$, provided

$$\beta_M \le \gamma_M \le \beta_{M-1} \le \gamma_{M-1} \le \cdots \le \beta_2 \le \gamma_2 \le \beta_1 \le \gamma_1.$$

That is, $\{\beta_m\}_{m=1}^M \sqsubseteq \{\gamma_m\}_{m=1}^M$ when $\gamma_{m+1} \le \beta_m \le \gamma_m$ for all $m = 1, \ldots, M$, provided we adopt the convention that $\gamma_{M+1} := 0$. As mentioned above, a classical result from matrix analysis gives that the spectra of the partial frame operators necessarily satisfy $\{\lambda_{k-1;m}\}_{m=1}^M \sqsubseteq \{\lambda_{k,m}\}_{m=1}^M$ for all $k = 1, \ldots, N$. This means that each pair of neighboring columns in (3) necessarily satisfy a zigzag of inequalities, each entry being no more than its neighbor to its right, which in turn is no more than its neighbor to its lower left. Gathering these four rules together, we arrive at the definition of a sequence of eigensteps:

Definition 1: A sequence $\{\lambda_{k;m}\}_{k=1,\ m=1}^{N\ \ M}$ is a sequence of *eigensteps* for given nonnegative nonincreasing sequences $\{\lambda_m\}_{m=1}^M$ and $\{\mu_n\}_{n=1}^N$ if:

(i) $\lambda_{0;m} = 0$ for all $m = 1, \ldots, M$,
(ii) $\lambda_{N;m} = \lambda_m$ for all $m = 1, \ldots, M$,
(iii) $\sum_{m=1}^M \lambda_{k;m} = \sum_{n=1}^k \mu_n$ for all $k = 1, \ldots, N$
(iv) $\{\lambda_{k-1;m}\}_{m=1}^M \sqsubseteq \{\lambda_{k;m}\}_{m=1}^M$ for all $k = 1, \ldots, N$.

To summarize, if $\{\varphi_n\}_{n=1}^N$ is any sequence of vectors in $\mathbb{C}^M$ whose frame operator has spectrum $\{\lambda_m\}_{m=1}^M$ and for which $\|\varphi_n\| = \mu_n$ for all n, then the spectra of its partial frame operators (2) necessarily form a corresponding sequence of eigensteps.

Remarkably, these relatively-simple necessary conditions on the existence of such frames are also sufficient. Indeed, as shown in [1], given a valid sequence of eigensteps, one can explicitly construct a sequence of vectors $\{\varphi_n\}_{n=1}^N$ with the desired spectrum and lengths. The approach is iterative: given $\{\varphi_n\}_{n=1}^k$ such that $\Phi_k\Phi_k^*$ has the desired spectrum $\{\lambda_{k;m}\}_{m=1}^M$, such that $\|\varphi_n\| = \mu_n$ for all $n = 1, \ldots, k$, and such that the eigenvectors of $\Phi_k\Phi_k^*$ are explicitly known, the algorithm shows how to choose φ_{k+1} as a linear combination of these eigenvectors so that $\Phi_{k+1}\Phi_{k+1}^*$ has spectrum $\{\lambda_{k+1;m}\}_{m=1}^M$ and such that $\|\varphi_{k+1}\|^2 = \mu_{k+1}$; the algorithm then goes on to explicitly update the eigenvectors of $\Phi_k\Phi_k^*$ into those of $\Phi_{k+1}\Phi_{k+1}^*$, as needed for the next iteration. Apart from possible rotations and reflections during each step of the process, the vectors constructed by the algorithm are unique. As such, eigensteps corresponding to a given $\{\lambda_m\}_{m=1}^M$ and $\{\mu_n\}_{n=1}^N$ can be viewed as the truly meaningful "parameters" of all vector sequences whose frame operator has that spectrum and whose elements have those lengths.

For example, in order to use these ideas to construct a UNTF of $N = 5$ elements for $M = 3$-dimensional space, we want a 3×6 table of eigensteps whose last column has the desired spectrum $\lambda_1 = \lambda_2 = \lambda_3 = \frac{N}{M} = \frac{5}{3}$ and whose zeroth column is zero; we also want the entries in the kth column to sum to $k = \sum_{n=1}^k \mu_n$, and for the values in any column to interlace on those in the preceding one. An example of such a table is

$$\begin{bmatrix} 0 & 0 & 0 & 0 & \frac{2}{3} & \frac{5}{3} \\ 0 & 0 & \frac{1}{3} & \frac{4}{3} & \frac{5}{3} & \frac{5}{3} \\ 0 & 1 & \frac{5}{3} & \frac{5}{3} & \frac{5}{3} & \frac{5}{3} \end{bmatrix}. \tag{4}$$

We emphasize that this table does not contain the frame vectors themselves, but rather the spectra of the partial frame operators. The process of transforming this table into the actual frame elements is nontrivial [1]. For example, to define φ_2,

we need to find a vector that makes the correct angle with φ_1 in order for $\varphi_1\varphi_1^* + \varphi_2\varphi_2^*$ to have spectrum $\{\frac{5}{3}, \frac{1}{3}, 0\}$.

Moreover, we also note that the above table corresponds to just one way of constructing a 3×5 UNTF. There are infinitely many others, meaning there are infinitely many UNTFs of five elements in three-dimensional space, even modulo rotations. In fact, one can show in this case that every sequence of eigensteps is of the form

$$\begin{bmatrix} 0 & 0 & 0 & x & \frac{2}{3} & \frac{5}{3} \\ 0 & 0 & y & \frac{4}{3}-x & \frac{5}{3} & \frac{5}{3} \\ 0 & 1 & 2-y & \frac{5}{3} & \frac{5}{3} & \frac{5}{3} \end{bmatrix}, \qquad (5)$$

where, in order to satisfy the interlacing requirements, we need to take our parameters (x, y) from the convex set

$$0 \le x \le \tfrac{2}{3}, \quad \max\{x, \tfrac{1}{3}\} \le y \le \min\{\tfrac{4}{3} - x, \tfrac{2}{3} + x\}.$$

This problem of constructing *every* sequence of eigensteps for a given $\{\lambda_m\}_{m=1}^M$ and $\{\mu_n\}_{n=1}^N$—thereby in effect constructing every sequence of vectors with this spectrum and set of lengths—is addressed in [3]. Here, it is important to note that there does not exist a set of eigensteps for every possible choice of $\{\lambda_m\}_{m=1}^M$ and $\{\mu_n\}_{n=1}^N$. Indeed, at a bare minimum, the second and third conditions of eigensteps require the λ_m's and the μ_n's to have the same sum; this corresponds to our final frame operator and Gram matrix having the same trace. Moreover, as evidenced in (4), the first and fourth conditions of eigensteps require us to have a "triangle of zeros" at the beginning of our table (3). That is, we necessarily have that $\lambda_{k;m} = 0$ for all $m > k$. When combined with our third and fourth conditions of eigensteps, this implies that the partial sums of our μ_n's are less than those of our λ_m's; for any $k = 1, \ldots, \min\{M, N\}$, we necessarily have

$$\sum_{n=1}^{k} \mu_n = \sum_{m=1}^{M} \lambda_{k;m} = \sum_{m=1}^{k} \lambda_{k;m} \le \sum_{m=1}^{k} \lambda_m.$$

Together, these facts state that in order for eigensteps to exist, our desired spectrum $\{\lambda_m\}_{m=1}^M$ must necessarily majorize our desired lengths $\{\mu_n\}_{n=1}^N$.

To be precise, we say that a nonnegative nonincreasing sequence $\{\lambda_m\}_{m=1}^M$ *majorizes* another such sequence $\{\mu_n\}_{n=1}^N$, denoted $\{\mu_n\}_{n=1}^N \preceq \{\lambda_m\}_{m=1}^M$, if

$$\sum_{n=1}^{N} \mu_n = \sum_{m=1}^{M} \lambda_m,$$
$$\sum_{n=1}^{k} \mu_n \le \sum_{m=1}^{k} \lambda_m, \ \forall k = 1, \ldots, \min\{M, N\}.$$

As we have just discussed, in order for a table of eigensteps to exist for a given $\{\lambda_m\}_{m=1}^M$ and $\{\mu_n\}_{n=1}^N$—that is, in order for there to exist a sequence of vectors whose frame operator has spectrum $\{\lambda_m\}_{m=1}^M$ and whose elements have lengths $\{\mu_n\}_{n=1}^N$—we necessarily have $\{\mu_n\}_{n=1}^N \preceq \{\lambda_m\}_{m=1}^M$. Remarkably, the converse of this statement is also true; this fact has been known for a long time, being a straightforward application of the classical Schur-Horn Theorem to the Gram matrix $\Phi^*\Phi$. However, the traditional proof of the converse is nonconstructive. The main contribution of [3] is to give a constructive proof of this converse and moreover, generalize the idea behind that construction so as to explicitly parameterize the convex polytope of *every* possible sequence of eigensteps.

To elaborate, the main idea of [3] is a new algorithm, dubbed *Top Kill*, for producing a valid sequence of eigensteps from a given $\{\lambda_m\}_{m=1}^M$ and $\{\mu_n\}_{n=1}^N$. The algorithm is iterative, starting with the final desired spectrum $\{\lambda_{N;m}\}_{m=1}^M = \{\lambda_m\}_{m=1}^M$ and working backwards from it to produce $\{\lambda_{N-1;m}\}_{m=1}^M$, then $\{\lambda_{N-2;m}\}_{m=1}^M$, etc., until finally arriving at $\{\lambda_{1;m}\}_{m=1}^M$. This algorithm has an intuitive, geometric motivation behind it (see [3]) that we do not have the space to discuss here.

In brief, however, note that looking at eigenstep tables such as (4), one quickly realizes that it is harder to get positive numbers in higher rows than it is in lower ones. This is because interlacing requires us to first build a suitable "foundation." That is, a number in a given column can only be a big as the one to its lower left, which, in turn, can only be as big as the one to its lower left, and so on. This can make it difficult to build the upper levels of our spectrum, especially if we do not plan ahead.

The Top Kill algorithm handles this issue by (i) working backwards from right to left, so that you are always explicitly using the final spectrum $\{\lambda_m\}_{m=1}^M$ you are trying to build and (ii) recognizing the higher rows are the most difficult to fill, and as such, making them the first thing we want to "kill" off. Indeed, the example table given in (4) is the result of applying Top Kill for $\lambda_m = \frac{5}{3}$ for all m and $\mu_n = 1$ for all n: to build each column from its neighbor to the right, we remove $\mu_n = 1$ units of "area," removing as much as possible from the highest row before removing from the second highest, and so on. Put another way, Top Kill's goal is to take x and y in (5) to be as small as possible, namely $(x, y) = (0, \frac{1}{3})$.

As detailed in [3], it turns out that the Top Kill algorithm will produce a valid sequence of eigensteps if and only our λ_m's majorize our μ_n's. In particular, if, for a given $\{\lambda_m\}_{m=1}^M$ and $\{\mu_n\}_{n=1}^N$, there is any way to produce a sequence of eigensteps, then Top Kill will produce such a sequence. That is, if anything works, then Top Kill does. This surprising fact led us to look for ways of generalizing Top Kill so that it could be applied to the frame completion problem.

III. Frame Completions

Recall the frame completion problem: given $\{\varphi_n\}_{n=1}^N$ in $\mathbb{C}^M$ and a set of desired lengths $\{\mu_{N+p}\}_{p=1}^P$, we want to add P new measurement vectors to $\{\varphi_n\}_{n=1}^N$ so that the frame operator of $\{\varphi_n\}_{n=1}^{N+P}$ has spectrum $\{\lambda_m\}_{m=1}^M$ and such that $\|\varphi_{N+p}\|^2 = \mu_{N+p}$ for all $p = 1, \ldots, P$. Letting $\mu_n := \|\varphi_n\|^2$ for all n, note that in accordance with the theory of eigensteps in [1], any such frame completion necessarily corresponds to an $M \times (N + P + 1)$ table of eigensteps. Moreover, letting $\{\alpha_m\}_{m=1}^M$ denote the spectrum of the frame operator of the "initial frame" $\{\varphi_n\}_{n=1}^N$, we necessarily have that the values $\{\alpha_m\}_{m=1}^M$ lie in the $k = N$ column of this table. We thus see that each of our desired frame completions corresponds to a

way of extending an existing $M \times (N+1)$ table of eigensteps by adding P new interlacing columns with the appropriate column sums.

As detailed in [4] and [7] it turns out that the existence of such "continued eigensteps" can be characterized in terms of majorization. However, it is not as simple as requiring that the final λ_m's majorize the μ_n's. Indeed, any such characterization must take into account the initial spectrum $\{\alpha_m\}_{m=1}^M$. At this point, we recall the motivation behind the Top Kill algorithm: when building eigensteps from a spectrum of zeros, we are faced with an "upper triangle" of zeros that makes it difficult to get large numbers in the higher rows of our table. However, this may not be the case when building eigensteps on top of an initial spectrum $\{\alpha_m\}_{m=1}^M$. Rather, it turns out that in this setting, what truly matters is how high the desired spectrum is *relative to the initial spectrum*.

A nonobvious concept such as this is best explained in pictures. In Figure 1(a), we see a given initial spectrum $\{\alpha_1, \alpha_2, \alpha_3\} = \{\frac{7}{4}, \frac{3}{4}, \frac{1}{2}\}$. In (b), this spectrum is overlaid with a desired completion $\{\lambda_1, \lambda_2, \lambda_3\} = \{\frac{13}{4}, \frac{9}{4}, 1\}$. Suppose we want to know whether or not our initial frame can be completed to one with spectrum $\{\lambda_m\}_{m=1}^3$ by adding four new frame vectors having lengths $\{\mu_{N+1}, \mu_{N+2}, \mu_{N+3}, \mu_{N+4}\} = \{2, 1, \frac{1}{4}, \frac{1}{4}\}$. To answer this question, we "chop" up the λ_m's according to m and the α_m's; see (c). In (d), we then label the area in each chopped region according to its height above the initial spectrum. The total "amount" of $\{\lambda_m\}_{m=1}^3$ that lies one unit above the existing spectrum is our first "diagonal sum":

$$\mathrm{DS}_1 := (\tfrac{13}{4} - \tfrac{7}{4}) + (\tfrac{7}{4} - \tfrac{3}{4}) + (\tfrac{3}{4} - \tfrac{1}{2}) = \tfrac{11}{4}.$$

Meanwhile, our second diagonal sum represents the total amount that lies two units above the existing spectrum:

$$\mathrm{DS}_2 := (\tfrac{9}{4} - \tfrac{7}{4}) + (\tfrac{5}{4} - 1) = \tfrac{3}{4}.$$

Finally, as there is no component of the λ_m's that lies three units above the existing spectrum, we take $\mathrm{DS}_3 = 0$.

The main result of our forthcoming paper [4] states that a given sequence $\{\lambda_m\}_{m=1}^M$ is realizable as the spectrum of a completion of a frame with initial spectrum $\{\alpha_m\}_{m=1}^M$ via the addition of P new measurements of lengths $\{\mu_{N+p}\}_{p=1}^P$ if and only if $\{\mathrm{DS}_m\}_{m=1}^M$ majorizes $\{\mu_{N+p}\}_{p=1}^P$. In particular, our example is constructible since $\mathrm{DS}_1 \geq \mu_{N+1}$, $\mathrm{DS}_1 + \mathrm{DS}_2 \geq \mu_{N+1} + \mu_{N+2}$, $\mathrm{DS}_1 + \mathrm{DS}_2 + \mathrm{DS}_3 \geq \mu_{N+1} + \mu_{N+2} + \mu_{N+3}$, and $\mathrm{DS}_1 + \mathrm{DS}_2 + \mathrm{DS}_3 = \frac{14}{4} = \mu_{N+1} + \mu_{N+2} + \mu_{N+3} + \mu_{N+4}$.

The necessity of this majorization follows from the fact that for any $k \leq \min\{M, P\}$, interlacing forces all of the "area" of $\{\lambda_{N+k;m}\}_{m=1}^M$ to lie at most k units above the initial spectrum $\{\alpha_m\}_{m=1}^M$. The part of the $\lambda_{N+k;m}$'s that lies outside of the α_m's envelope has a total area of $\sum_{n=N+1}^{N+k} \mu_n$. The amount of area in these diagonals will only grow as more frame vectors are added, meaning

$$\sum_{n=N+1}^{N+k} \mu_n = \sum_{m=1}^{k} \mathrm{DS}_{m;k} \leq \sum_{m=1}^{k} \mathrm{DS}_{m;P} = \sum_{m=1}^{k} \mathrm{DS}_m.$$

When all P vectors are added, the above inequalities become equalities.

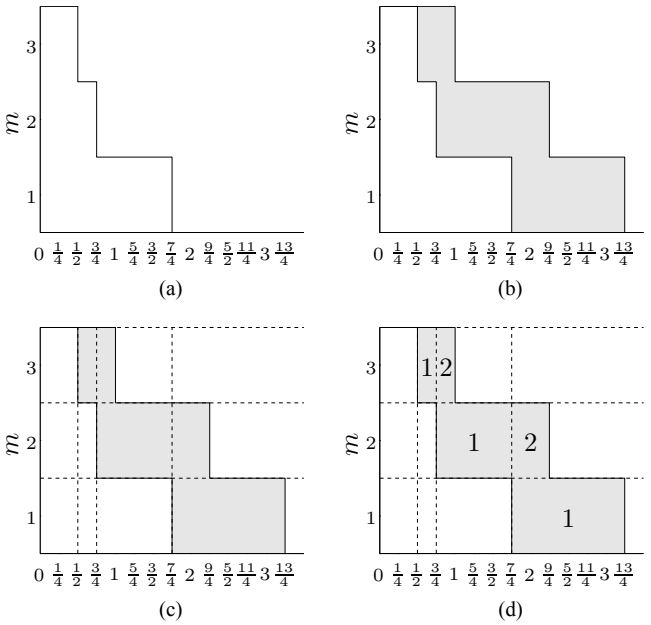


Fig. 1. Determining the relative height of a desired completion's spectrum above an existing one.

The sufficiency of this majorization condition follows from a variation of the Top Kill algorithm, called "Chop Kill." Here, we can build a valid sequence of eigensteps, provided we once again start with the desired spectrum and work backwards. However, rather than removing as much of the "top" of the spectrum as quickly as possible, we instead remove as much as possible from the outermost *diagonals*. This is consistent with the original motivation behind Top Kill: once we identify the hardest parts of our spectrum to construct, we work backwards, taking care of those parts as soon as possible.

ACKNOWLEDGMENT

This work was supported by NSF DMS 1042701 and NSF CCF 1017278. The views expressed in this article are those of the authors and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the U.S. Government.

REFERENCES

[1] J. Cahill, M. Fickus, D. G. Mixon, M. J. Poteet, N. Strawn, Constructing finite frames of a given spectrum and set of lengths, to appear in: Appl. Comput. Harmon. Anal.

[2] D. J. Feng, L. Wang, Y. Wang, Generation of finite tight frames by Householder transformations, Adv. Comput. Math. 24 (2006) 297-309.

[3] M. Fickus, D. G. Mixon, M. J. Poteet, N. Strawn, Constructing all self-adjoint matrices with prescribed spectrum and diagonal, to appear in: Adv. Comput. Math.

[4] M. Fickus, M. J. Poteet, A generalized Schur-Horn Theorem for frame completions, in preparation.

[5] J. Kovačević, A. Chebira, Life beyond bases: The advent of frames (Part I), IEEE Signal Process. Mag. 24 (2007) 86–104.

[6] P. G. Massey, M. A. Ruiz, D. Stojanoff, Optimal dual frames and frame completions for majorization, Appl. Comput. Harmon. Anal. 34 (2013) 201–223.

[7] M. J. Poteet, Parametrizing finite frames and optimal frame completions, Ph.D. dissertation, Air Force Institute of Technology, 2012.

A note on scalable frames

Jameson Cahill
Department of Mathematics
University of Missouri
jameson.cahill@gmail.com

Xuemei Chen
Department of Mathematics
University of Maryland
xuemeic@math.umd.edu

***Abstract*—We study the problem of determining whether a given frame is scalable, and when it is, understanding the set of all possible scalings. We show that for most frames this is a relatively simple task in that the frame is either not scalable or is scalable in a unique way, and to find this scaling we just have to solve a linear system. We also provide some insight into the set of all scalings when there is not a unique scaling. In particular, we show that this set is a convex polytope whose vertices correspond to minimal scalings.**

I. Introduction

A collection of vectors $\{\varphi_i\}_{i=1}^n \subseteq \mathbb{C}^d$ is called a *frame* if there are positive numbers $A \leq B < \infty$ such that

$$A\|x\|^2 \leq \sum_{i=1}^n |\langle x, \varphi_i\rangle|^2 \leq B\|x\|^2$$

for every x in $\mathbb{C}^d$. If we have $A = B$ we say the frame is *tight*, and if $A = B = 1$ we say it is a *Parseval frame*. Given a frame $\{\varphi_i\}_{i=1}^n$ we define the *frame operator* $S : \mathbb{C}^d \to \mathbb{C}^d$ by

$$Sx = \sum_{i=1}^n \langle x, \varphi_i\rangle \varphi_i. \tag{1}$$

It is easy to see that S is always positive, invertible, and Hermitian. Furthermore, $\{\varphi_i\}_{i=1}^n$ is a Parseval frame if and only if $S = I_d$ (the identity operator on $\mathbb{C}^d$). By a slight abuse of notation, given any set of vectors $\{\varphi_i\}_{i=1}^n \subseteq \mathbb{C}^d$ we will refer to the operator defined in (1) as their frame operator, even if they do not form a frame (in this case S will not be invertible, but it will still be positive). If we have that $\|\varphi_i\| = 1$ for every $i = 1, ..., n$ we say it is a *unit norm* frame. For more background on finite frames we refer to the book [4].

A frame $\{\varphi_i\}_{i=1}^n$ is said to be *scalable* if there exists a collection of scalars $\{v_i\}_{i=1}^n \subseteq \mathbb{C}$ so that $\{v_i\varphi_i\}_{i=1}^n$ is a Parseval frame. In this case, we call the vector $(|v_1|^2, ..., |v_n|^2) \in \mathbb{R}_+^n$ a *scaling* of $\{\varphi_i\}_{i=1}^n$. Scalable frames have been studied previously in [5].

We will work in the space $\mathbb{H}_{d\times d}$ of all $d \times d$ Hermitian matrices. Note that this is a **real** vector space of dimension d^2 (it is not a space over the complex numbers since a Hermitian matrix multiplied by a complex scalar is no longer Hermitian). The inner product on this space is given by $\langle S, T\rangle = \text{Trace}(ST)$ and the norm induced by this inner product is the Froebenius norm, *i.e.*, $\langle S, S\rangle = \|S\|_F^2$.

In what follows we will always consider frames in the complex space $\mathbb{C}^d$, however all of our results hold in the real space $\mathbb{R}^d$ as well. The only difference is in this case we must replace the space $\mathbb{H}_{d\times d}$ with its subspace $\mathbb{S}_{d\times d}$ consisting of all $d \times d$ real symmetric matrices, which is a real vector space of dimension $d(d+1)/2$. Thus, if one replaces $\mathbb{H}_{d\times d}$ with $\mathbb{S}_{d\times d}$ and d^2 with $d(d+1)/2$ all of our results will hold for frames $\{\varphi_i\}_{i=1}^n \subseteq \mathbb{R}^d$ and the same proofs will work.

II. Scaling generic frames

Consider the mapping from $\mathbb{C}^d$ to $\mathbb{H}_{d\times d}$ given by

$$x \mapsto xx^*.$$

Note that xx^* is the rank one projection onto $\text{span}\{x\}$ scaled by $\|x\|^2$. xx^* is called the *outer product* of x with itself. Also note that if $x = \lambda y$ for $\lambda \in \mathbb{C}$ then $xx^* = (\lambda y)(\lambda y)^* = |\lambda|^2 yy^*$.

Given a frame $\{\varphi_i\}_{i=1}^n$, in this setting we have that the frame operator is given by

$$S = \sum_{i=1}^n \varphi_i\varphi_i^*,$$

so $\{\varphi_i\}_{i=1}^n$ is scalable if and only if there exists a collection of **positive** scalars $\{w_i\}_{i=1}^n$ so that

$$\sum_{i=1}^n w_i\varphi_i\varphi_i^* = I_d,$$

in this case $\{\sqrt{w_i}\varphi_i\}_{i=1}^n$ is a Parseval frame, and the vector $(w_1, ..., w_n) \in \mathbb{R}_+^n$ is the scaling.

Before stating our first theorem we need one more definition. A subset $Q \subseteq \mathbb{R}^n$ is called *generic* if there exists a polynomial $p(x_1, ..., x_n)$ such that $Q^c = \{(x_1, ..., x_n) \in \mathbb{R}^n : p(x_1, ..., x_n) = 0\}$. It is a standard fact that generic sets are open, dense, and full measure. When we talk about a generic set in $\mathbb{C}^d$ we mean that it is generic when we identify $\mathbb{C}^d$ with $\mathbb{R}^{2d}$.

Theorem 1. *For a generic choice of vectors* $\{\varphi_i\}_{i=1}^{d^2} \subseteq \mathbb{C}^d$ *we have that* $\text{span}\{\varphi_i\varphi_i^*\}_{i=1}^{d^2} = \mathbb{H}_{d\times d}$.

Proof: First let $\{T_i\}_{i=1}^{d^2}$ be any basis for $\mathbb{H}_{d\times d}$. Since each T_i is Hermitian we can use the spectral theorem to get a decomposition $T_i = \sum_{j=1}^n \lambda_{ij}P_{ij}$ where each P_{ij} is rank 1. So it follows that $\text{span}\{P_{ij}\} = \mathbb{H}_{d\times d}$ and therefore this set contains a basis of $\mathbb{H}_{d\times d}$. Thus, we have constructed a basis of $\mathbb{H}_{d\times d}$ consisting only of rank 1 matrices.

Now observe that for a given choice of vectors $\{\varphi_i\}_{i=1}^{d^2}$ we have that $\text{span}\{\varphi_i\varphi_i^*\} = \mathbb{H}_{d\times d}$ if and only if the determinant of the frame operator is nonzero (note that we are refering to the frame operator of $\{\varphi_i\varphi_i^*\}_{i=1}^{d^2}$ as an operator on $\mathbb{H}_{d\times d}$, not the frame operator of $\{\varphi_i\}_{i=1}^n$ as an operator on $\mathbb{C}^d$). But the determinant of the frame operator is a polynomial in the (real and imaginary parts) of the entries of the φ_i's, and by the first paragraph we know that there is at least one choice for which this does not vanish, so we can conclude that for a generic choice it does not vanish. ∎

Corollary 1. *If $n \leq d^2$ then for a generic choice of vectors $\{\varphi_i\}_{i=1}^n \subseteq \mathbb{C}^d$ we have that $\{\varphi_i\varphi_i^*\}_{i=1}^n$ is linearly independent.*

Given a frame $\{\varphi_i\}_{i=1}^n \subseteq \mathbb{C}^d$ define the operator $\mathcal{A} : \mathbb{R}^n \to \mathbb{H}_{d\times d}$ by

$$\mathcal{A}w = \sum_{i=1}^n w_i\varphi_i\varphi_i^*$$

where $w = (w_1, ..., w_n)^T$. To determine whether $\{\varphi_i\}_{i=1}^n$ is scalable boils down to finding a nonnegative solution to

$$\mathcal{A}w = I_d.$$

In the generic case when $\{\varphi_i\varphi_i^*\}_{i=1}^n$ is linearly independent, this system is guaranteed to have either no solution, or one unique solution. So if it either has a solution with a negative entry or has no solution we can conclude that this frame is not scalable, and if it has a nonnegative solution then it is scalable and this solution tells us the unique scalars to use. We summarize this in the following corollary:

Corollary 2. *Given frame $\{\varphi_i\}_{i=1}^n \subseteq \mathbb{C}^d$ such that $\{\varphi_i\varphi_i^*\}_{i=1}^n$ is linearly independent in $\mathbb{H}_{d\times d}$, we can determine its scalability by solving the linear system*

$$\mathcal{A}w = I_d. \tag{2}$$

Furthermore, in this case if it is scalable then it is scalable in a unique way.

In particular, if $n \leq d^2$ then with probability 1, determining the scalability of $\{\varphi_i\}_{i=1}^n$ is equivalent to solving the linear system given in (2)*.*

III. Linearly dependent outer products

In this section we will address the situation when $\{\varphi_i\varphi_i^*\}_{i=1}^n$ is linearly dependent. The main problem here is that the system $\mathcal{A}w = I_d$ may have many solutions, and possibly none of them are nonnegative. In this section we will find it convenient to assume that $\|\varphi_i\| = 1$ for every $i = 1, ..., n$, note that we lose no generality by making this assumption.

Given a collection of vectors $\{x_i\}_{i=1}^n \subseteq \mathbb{R}^d$ we define their *affine span* as

$$\text{aff}\{x_i\}_{i=1}^n := \{\sum_{i=1}^n c_i x_i : \sum_{i=1}^n c_i = 1\}$$

and we say that $\{x_i\}_{i=1}^n$ is *affinely independent* if

$$x_j \notin \text{aff}\{x_i\}_{i\neq j}$$

for every $j = 1, ..., n$. We also define their *convex hull* as

$$\text{conv}\{x_i\}_{i=1}^n := \{\sum_{i=1}^n c_i x_i : c_i \geq 0, \sum_{i=1}^n c_i = 1\}.$$

We say a set $\mathcal{P} \subseteq \mathbb{R}^d$ is called a *polytope* if it is the convex hull of finitely many points.

Proposition 1. *Given a collection of unit norm vectors $\{\varphi_i\}_{i=1}^n \subseteq \mathbb{C}^d$ we have that $\{\varphi_i\varphi_i^*\}_{i=1}^n$ is linearly independent if and only if it is affinely independent.*

Proof: Clearly linear independence always implies affine independence. So suppose that $\{\varphi_i\varphi_i^*\}_{i=1}^n$ is not linearly independent. Then we have an equation of the form

$$\varphi_j\varphi_j^* = \sum_{i\neq j} c_i\varphi_i\varphi_i^*$$

for some j. Also note that since $\|\varphi_i\| = 1$ it follows that $\langle\varphi_i\varphi_i^*, I_d\rangle = 1$ for every $i = 1, ..., n$. Therefore, we have

$$\begin{aligned} 1 &= \langle\varphi_j\varphi_j^*, I_d\rangle = \langle\sum_{i\neq j} c_i\varphi_i\varphi_i^*, I_d\rangle \\ &= \sum_{i\neq j} c_i\langle\varphi_i\varphi_i^*, I_d\rangle = \sum_{i\neq j} c_i. \end{aligned}$$

Therefore $\{\varphi_i\varphi_i^*\}_{i=1}^n$ is not affinity independent. ∎

Proposition 2. *A unit norm frame $\{\varphi_i\}_{i=1}^n \subseteq \mathbb{C}^d$ is scalable if and only if $\frac{1}{d}I_d \in \text{conv}\{\varphi_i\varphi_i^*\}_{i=1}^n$. Furthermore, if $\lambda I_d \in \text{conv}\{\varphi_i\varphi_i^*\}_{i=1}^n$ then $\lambda = \frac{1}{d}$ and if $\sum_{i=1}^n w_i\varphi_i\varphi_i^* = \frac{1}{d}I_d$ then $\sum_{i=1}^n w_i = 1$.*

Proof: Suppose we have a scaling w so that

$$I_d = \sum_{i=1}^n w_i\varphi_i\varphi_i^*.$$

Then

$$\begin{aligned} d &= \langle I_d, I_d\rangle = \langle\sum_{i=1}^n w_i\varphi_i\varphi_i^*, I_d\rangle \\ &= \sum_{i=1}^n w_i\langle\varphi_i\varphi_i^*, I_d\rangle = \sum_{i=1}^n w_i. \end{aligned}$$

Thus, $\sum_{i=1}^n \frac{w_i}{d} = 1$ and since $w_i \geq 0$ for every $i = 1, ..., n$ it follows that $\frac{1}{d}I_d = \sum_{i=1}^n \frac{w_i}{d}\varphi_i\varphi_i^* \in \text{conv}\{\varphi_i\varphi_i^*\}_{i=1}^n$. The converse is obvious.

The furthermore part follows from a similar argument. Suppose $\lambda I_d = \sum_{i=1}^n w_i\varphi_i\varphi_i^*$ with $\sum_{i=1}^n w_i = 1$. Then

$$d\lambda = \langle\lambda I_d, I_d\rangle = \sum_{i=1}^n w_i = 1.$$

Now suppose $\frac{1}{d}I_d = \sum_{i=1}^n w_i\varphi_i\varphi_i^*$. Then

$$1 = \langle\sum_{i=1}^n w_i\varphi_i\varphi_i^*, I_d\rangle = \sum_{i=1}^n w_i.$$

∎

The following theorem is known as Carathéodory's theorem:

Theorem 2. *Given a set of points $\{x_i\}_{i=1}^n \subseteq \mathbb{R}^d$ suppose $y \in \text{conv}\{x_i\}_{i=1}^n$. Then there exists a subset $I \subseteq \{1,...,n\}$ such that $y \in \text{conv}\{x_i\}_{i\in I}$ and $\{x_i\}_{i\in I}$ is affinely independent.*

Corollary 3. *Suppose $\{\varphi_i\}_{i=1}^n \subseteq \mathbb{C}^d$ is a scalable frame. Then there is a subset $\{\varphi_i\}_{i\in I}$ which is also scalable and $\{\varphi_i\varphi_i^*\}_{i\in I}$ is linearly independent.*

Given a unit norm frame $\{\varphi_i\}_{i=1}^n \subseteq \mathbb{C}^d$ we define the set

$$\mathcal{P}(\{\varphi_i\}_{i=1}^n) := \{(w_1,...,w_n) : w_i \geq 0, \sum_{i=1}^n w_i\varphi_i\varphi_i^* = \frac{1}{d}I_d\}.$$

Proposition 2 tells us two things about this set: first we have that $w \in \mathcal{P}(\{\varphi_i\}_{i=1}^n)$ if and only if $d \cdot w$ is a scaling of $\{\varphi_i\}_{i=1}^n$, and second, that $\mathcal{P}(\{\varphi_i\}_{i=1}^n)$ is a (possibly empty) polytope (see, for example, Theorem 1.1 in [6]).

Suppose $\{\varphi_i\}_{i=1}^n \subseteq \mathbb{C}^d$ is a scalable frame, and we are given a scaling $w = (w_1,...,w_n)$. We say the scaling is *minimal* if $\{\varphi_i : w_i > 0\}$ has no proper subset which is scalable.

Theorem 3. *Suppose $\{\varphi_i\}_{i=1}^n \subseteq \mathbb{C}^d$ is a scalable, unit norm frame. If $w = (w_1,...,w_n)$ is a minmal scaling then $\{\varphi_i\varphi_i^* : w_i > 0\}$ is linearly independent. Furthermore, $\mathcal{P}(\{\varphi_i\}_{i=1}^n)$ is the convex hull of the minimal scalings, i.e., every scaling is a convex combination of minimal scalings.*

Proof: The first statement follows directly from Corollary 3.

We now show that every vertex of $\mathcal{P}(\{\varphi_i\}_{i=1}^n)$ is indeed a minimal scaling. Let $u \in \mathcal{P}(\{\varphi_i\}_{i=1}^n)$ be a vertex and assume to the contrary that u is not minimal, then there exists a $v \in P$ such that $\text{supp}(v) \subsetneq \text{supp}(u)$. Let $w(t) = v + t(u-v)$, and $t_0 = \min\{\frac{v_i}{v_i-u_i} : v_i > u_i\}$. We observe that $t_0 > 1$ and $w(t_0)_i \geq 0$ since $\text{supp(v)} \subsetneq \text{supp(u)}$. This means $w(t_0) \in P$, and u lies on the line segment connecting v and $w(t_0)$ which contradicts the fact that u is a vertex.

Finally we show that every minimal scaling is a vertex of $\mathcal{P}(\{\varphi_i\}_{i=1}^n)$. Suppose we are given a minimal scaling w which is not a vertex of $\mathcal{P}(\{\varphi_i\}_{i=1}^n)$. Then we can write w as a convex combination of vertices, say $w = \sum t_i v_i$, where we know at least two t_i's are nonzero, without loss of generality say t_1 and t_2. Since both t_1 and t_2 are positive and all the entries of v_1 and v_2 are nonnegative, it follows that $\text{supp}(v_1) \cup \text{supp}(v_2) \subseteq \text{supp}(w)$, which contradicts the fact the w is a minimal scaling. ∎

Theorem 3 reduces the problem of understanding the scalings of the frame $\{\varphi_i\}_{i=1}^n$ to that of finding the vertices of the polytope $\mathcal{P}(\{\varphi_i\}_{i=1}^n)$. Relatvely fast algorithms for doing this are known, see [2].

IV. When are outer products linearly independent?

Since most of the results in this paper deal with linear independence of the outer products of subsets of our frame vectors we will address this issue in this section. It would be nice if there were conditions on a frame $\{\varphi_i\}_{i=1}^n$ which could guarantee that the set of outer products $\{\varphi_i\varphi_i^*\}_{i=1}^n$ is linearly independent, or conversely if knowing that $\{\varphi_i\varphi_i^*\}_{i=1}^n$ is linearly independent tells anything about the frame $\{\varphi_i\}_{i=1}^n$. One obvious condition is that in order for $\{\varphi_i\varphi_i^*\}_{i=1}^n$ to be linearly independent we must have $n \leq d^2$, and when this is satisfied Theorem 1 tells us that this will usually be the case.

Another condition which is easy to prove is that if $\{\varphi_i\}_{i=1}^n$ is linearly independent then so is $\{\varphi_i\varphi_i^*\}_{i=1}^n$. The converse of this is certainly not true, and since we are usually interested in frames for which $n > d$ this condition is not very useful. The main idea here is that while the frame vectors live in a d-dimensional space the outer products live in a d^2-dimensional space, so there is much more "room" for them to be linearly independent.

Given a frame $\{\varphi_i\}_{i=1}^n$ we define its *spark* to be the size of its smallest linearly dependent subset, more precisely

$$\text{spark}(\{\varphi_i\}_{i=1}^n) := \min\{|I| : \{\varphi_i\}_{i\in I} \text{ is linearly dependent}\}.$$

Clearly for a frame $\{\varphi_i\}_{i=1}^n \subseteq \mathbb{C}^d$ we must have that $\text{spark}(\{\varphi_i\}_{i=1}^n) \leq d+1$, if its spark is equal to $d+1$ we say it is *full spark*. For more background on full spark frames see [1].

Proposition 3. *Suppose $\{\varphi_i\}_{i=1}^n \subseteq \mathbb{C}^d$ is a frame with $n \leq 2d-1$. If $\{\varphi_i\}_{i=1}^n$ is full spark then $\{\varphi_i\varphi_i^*\}_{i=1}^n$ is linearly independent.*

Proof: Suppose by way of contradiction that $\{\varphi_i\}_{i=1}^n$ is full spark but $\{\varphi_i\varphi_i^*\}_{i=1}^n$ is linearly dependent. Then we can write an equation of the form

$$\sum_{i\in I} a_i\varphi_i\varphi_i^* = \sum_{j\in J} b_j\varphi_j\varphi_j^*$$

with $a_i > 0$ for every $i \in I$, $b_j > 0$ for every $j \in J$, and $I \cap J = \emptyset$. This implies that

$$\begin{aligned}\text{span}(\{\varphi_i\}_{i\in I}) &= \text{Im}(\sum_{i\in I} a_i\varphi_i\varphi_i^*) \\ &= \text{Im}(\sum_{j\in J} b_j\varphi_j\varphi_j^*) = \text{span}(\{\varphi_j\}_{j\in J}).\end{aligned}$$

But since $n \leq 2d-1$ we have either $|I| \leq d-1$ or $|J| \leq d-1$, so this contradicts the fact the $\{\varphi_i\}_{i=1}^n$ is full spark. ∎

We first remark that the converse of Proposition 3 is not true:

Example 1. *Let $\{e_1,e_2,e_3\}$ be an orthonormal basis for $\mathbb{C}^3$ and consider the frame $\{e_1,e_2,e_3,e_1+e_2,e_2+e_3\}$. Clearly this frame is not full spark and yet it is easy to verify that $\{e_1e_1^*, e_2e_2^*, e_3e_3^*, (e_1+e_2)(e_1+e_2)^*, (e_2+e_3)(e_2+e_3)^*\}$ is linearly independent.*

Next we remark that the assumption $n \leq 2d-1$ is necessary:

Example 2. *Let $\{e_1,e_2\}$ be an orthonormal basis for $\mathbb{C}^2$ and consider the frame $\{e_1,e_2,e_1+e_2,e_1-e_2\}$. Clearly this frame is full spark but*

$$e_1e_1^*+e_2e_2^* = I_2 = \frac{1}{2}((e_1+e_2)(e_1+e_2)^*+(e_1-e_2)(e_1-e_2)^*).$$

Finally we remark that with only slight modifications the proof of Propostion 3 can be used to prove the following more general result:

Proposition 4. *If* $\operatorname{spark}(\{\varphi_i\}_{i=1}^n) \geq s$ *then* $\operatorname{spark}(\{\varphi_i\varphi_i^*\}_{i=1}^n) \geq 2s-2$.

Unfortunately, the converse of Proposition 4 is still not true. The main problem here is that given any three vectors such that no one of them is a scalar multiple of another, the corresponding outer products will be linearly independent (we leave the proof of this as an exercise). Therefore it is easy to make examples (such as Example 1 above) of frames that have tiny spark, but the corresponding outer products are linearly independent.

We conclude our discussion of spark by remarking that in [1] it is shown that computing the spark of a general frame is NP-hard. Thus, the small amount of insight we gain from Proposition 4 is of little practical use.

Another property worth mentioning in this section is known as the *complement property*. A frame $\{\varphi_i\}_{i=1}^n \subseteq \mathbb{C}^d$ has the complement property if for every $I \subseteq \{1,...,n\}$ we have either $\operatorname{span}(\{\varphi_i\}_{i\in I}) = \mathbb{C}^d$ or $\operatorname{span}(\{\varphi_i\}_{i\in I^c}) = \mathbb{C}^d$. We remark that the complement property is usually discussed for frames in a real vector space, but for our purposes it is fine to discuss it for frames in a complex space. In [3] the complement property was shown to be necessary and sufficient to do phaseless reconstruction in the real case.

If a frame $\{\varphi_i\}_{i=1}^n \subseteq \mathbb{C}^d$ has the complement property then clearly we must have $n \geq 2d-1$ (if not we could partition the frame into two sets each of size at most $d-1$) and that in this case full spark implies the complement property. If $n = 2d-1$ then the complement property is equivalent to full spark, but for $n > 2d-1$ the complement property is (slightly) weaker. One might ask if the complement property tells us anything about the linear independence of the outer products, or vice versa. Example 1 above is an example of a frame which does not have the complement property but the outer products are linearly independent, and Example 2 is an example of a frame that does have the complement property but the outer products are linearly dependent. So it seems like the complement property has nothing to do with the linear independence of the outer products.

Given a frame with the complement property we can add any set of vectors to it without losing the complement property. Thus it seems natural to ask whether every frame with the complement property has a subset of size $2d-1$ which is full spark. This also turns out to be not true as the following example shows:

Example 3. *Consider the frame in Example 1 with the vector* e_1+e_3 *added to it. It is not difficult to verify that this frame does have the complement property, but no subset of size 5 is full spark.*

We conclude by noting that as in the proof of Proposition 3, a set of outer products $\{\varphi_i\varphi_i^*\}_{i=1}^n$ is linearly dependent if and only if we have an equation of the form

$$\sum_{i\in I} a_i\varphi_i\varphi_i^* = \sum_{j\in J} b_j\varphi_j\varphi_j^*$$

with $a_i > 0$ for every $i \in I$, $b_j > 0$ for every $j \in J$, and $I \cap J = \emptyset$. This is equivalent to $\{\varphi_i\}_{i=1}^n$ having two disjoint subsets, namely $\{\varphi_i\}_{i\in I}$ and $\{\varphi_j\}_{j\in J}$, which can be scaled to have the same frame operator. Thus, determining whether $\{\varphi_i\varphi_i^*\}_{i=1}^n$ is linearly independent is equivalent to solving a more difficult scaling problem than the one presented in this paper.

ACKNOWLEDGMENT

The authors would like to thank Peter Casazza and Dustin Mixon for insightful conversations during the writing of this paper. The first author was supported by NSF 1008183; NSF ATD 1042701; and AFOSR DGE51: FA9550-11-1-0245

REFERENCES

[1] B. Alexeev, J. Cahill, D.G. Mixon, *Full spark frames*, J. Fourier Anal. Appl. **18** No. 6 (2012), 1167-1194.
[2] D. Avis and K. Fukuda, *A pivoting algorithm for convex hulls and vertex enumeration of arrangements and polyhedra*, Discrete Comput. Geom., **8** No. 1 (1992) 295-313.
[3] R. Balan, P.G. Casazza, D. Edidin, *On signal reconstruction without phase*, Appl. Comput. Harmon. A. **20** No. 3 (2006), 345-356.
[4] P.G. Casazza, G. Kutyniok, (eds.), *Finite Frames: Theory and Applications*, Birkäuser, 2012.
[5] G. Kutyniok, K. Okoudjou, F. Philipp, and E.K. Tuley, *Scalable frames*, arXiv:1204.1880.
[6] G. Ziegler, *Lectures on polytopes*, Springer, 1995.

Measurement Structures and Constraints in Compressive RF Systems

Nathan A. Goodman
School of Electrical and Computer Engineering
Advanced Radar Research Center
The University of Oklahoma, Norman, Oklahoma 73019
goodman@ou.edu

***Abstract*—Compressive sensing (CS) is a powerful technique for sub-sampling of signals combined with reconstruction based on sparsity. Many papers have been published on the topic; however, they often fail to consider practical hardware factors that may prevent or alter the implementation of desired CS measurement kernels. In particular, different compressive architectures in the RF domain either sacrifice collected signal energy or create noise folding, both of which cause SNR reduction. In this paper, we consider valid signal models and other system aspects of RF compressive systems.**

I. Introduction

Interest in the application of compressive sensing (CS) to radar and other radio frequency (RF) applications has grown rapidly. This interest is fueled by the potential to implement RF systems that perform well while reducing the burden on data collection hardware. For example, the idea of a sparsely populated or thinned array has been used for a long time as a way of obtaining high resolution from a long array baseline without the cost and weight of a fully populated array. The difference in recent years is that CS principles are now being used to design the array, to analyze its performance, and to process its data via sparse reconstruction methods. Similar statements can be made regarding other examples of compressive RF systems (not just antenna arrays).

The current literature on compressive RF systems is skewed toward demonstrating the ability to recover signals from such systems using sparse reconstruction methods. While this approach is interesting, there is still a shortage of analysis on the system impacts of RF compression and on the ultimate performance of compressive RF systems in useful exploitation tasks such as signal detection and parameter estimation. Unfortunately, some published papers also fail to consider the architecture of the compressive system and resulting constraints that this architecture imposes on the structure of the compression kernels and the relevant signal model.

In this paper, we address some of these structural and system considerations in compressive RF systems. We consider fundamental models of measurement as linear projections implemented in time and space, and map these models to appropriately structured sensing matrices for several types of compressive RF sensing. We then focus on compression via sub-Nyquist analog-to-digital conversion (ADC) and consider a compressive version of the traditional quadrature receiver.

II. Models for RF Compression

In this section, we start with a model for conventional sampling and map that model to the matrix-vector notation typically used in CS. We then use the model to represent two types of compression that RF systems might employ, namely measurement "thinning" and measurement "mixing". We also discuss the inclusion of additive receiver noise in the models.

Let a compressive RF receiver observe a signal, $s(\mathbf{r},t)$, that varies over space and time where $\mathbf{r}$ is a three-dimensional vector of spatial coordinates. Let the compressive system comprise a P-element antenna array with element coordinates $\mathbf{r}_p$ and each element having its own receiver. We can express a "traditional" sample as the projection of the signal onto a measurement kernel that is localized in time and space. In ideal sampling, this kernel is an impulse-like space-time function located at the element position and sample time where the sample is to occur, such that

$$\begin{aligned} s(\mathbf{r}_p,t_n) &= \left\langle s(\mathbf{r},t), \delta(\mathbf{r}-\mathbf{r}_p)\delta(t-t_n)\right\rangle \\ &= \iint s(\mathbf{r},t)\delta(\mathbf{r}-\mathbf{r}_p)\delta(t-t_n)\,d\mathbf{r}dt \end{aligned} . \quad (1)$$

We can then express data collected by the array and sampled over time by a set of impulsive measurement kernels located at every element location and sampling time instant.

In order to represent (1) with a discrete model suitable for computer simulations or manipulation via linear algebra, we can approximate the integral in (1) with a summation over small, finite-sized *bins* in space and time, such that

$$\iint s(\mathbf{r},t)\delta(\mathbf{r}-\mathbf{r}_p)\delta(t-t_n)\,d\mathbf{r}dt \approx \Delta\mathbf{r}\Delta t\sum_{\mathbf{r}(i)}\sum_{t(j)} s(\mathbf{r}(i),t(j))\delta[\mathbf{r}(i)-\mathbf{r}_p]\delta[t(j)-t_n] \quad (2)$$

where $\mathbf{r}(i)$ denotes the ith spatial bin in the approximation, $t(j)$ denotes the jth temporal bin, and the delta function with brackets, $\delta[\cdot]$ is used to denote the Kronecker delta function that is equal to one when the argument is zero (to within the quantization error of the bins) and zero elsewhere. If the bins are chosen smaller than or equal to the Nyquist sampling interval, then the discrete approximation will be accurate.

Next, we form a signal vector by taking all signal values over the discrete bins and organizing them into a vector according to a specific ordering; for example,

$$\mathbf{s} = \begin{bmatrix} s(\mathbf{r}(1), t(1)) \\ s(\mathbf{r}(1), t(2)) \\ s(\mathbf{r}(1), t(3)) \\ \vdots \\ s(\mathbf{r}(N_s), t(N_t)) \end{bmatrix} \tag{3}$$

where N_s is the number of discrete bins covering the signal's spatial volume and N_t is the number of discrete bins covering the signal's temporal duration. The length of $\mathbf{s}$ is $N = N_t N_s$. The expression in (2) can then be expressed as

$$\begin{aligned} &\iint s(\mathbf{r},t)\delta(\mathbf{r}-\mathbf{r}_p)\delta(t-t_n)\,d\mathbf{r}dt \\ &\quad \approx \Delta\mathbf{r}\Delta t\begin{bmatrix}0 & \cdots & 0 & 1 & 0 & 0 & \cdots & 0\end{bmatrix}\mathbf{s} \\ &\quad = \Delta\mathbf{r}\Delta t\boldsymbol{\delta}_m\mathbf{s} \end{aligned} \tag{4}$$

where $\boldsymbol{\delta}_m$ is defined (as shown) as a row vector with zeros everywhere except in the entry corresponding to the discretized bin where the mth data sample is collected. A sampling matrix can then be represented as a collection of $\boldsymbol{\delta}_m$'s, with each row having the '1' in a different location. For example, if only three data samples are collected, the sensing matrix might look like

$$\boldsymbol{\Phi} = \begin{bmatrix}\boldsymbol{\delta}_1 \\ \boldsymbol{\delta}_2 \\ \boldsymbol{\delta}_3\end{bmatrix} = \begin{bmatrix}1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & \cdots & 1\end{bmatrix}. \tag{5}$$

The matrix in (5) represents a "thinning" type of compression where the sampling kernels are still localized in space and time, but not all signal elements are sampled. Some radar-specific examples of where this type of compression might occur include 1) some pulses in a coherent pulse train are not transmitted, causing gaps in the slow-time data collection, 2) the thinned or sparse antenna array mentioned above, and 3) stepped-frequency waveforms where frequency steps can be skipped in the data collection process [1]. In these structures, there will be groups of nearby samples taken at the Nyquist rate, followed by gaps in the sampling. Full Nyquist sampling can also be represented by using $\boldsymbol{\Phi} = \mathbf{I}_N$.

On the other hand, the "thinning" type compression depicted in (5) is usually not a suitable representation for compression in the ADC process (i.e., in fast time). For the examples above, it is easy to envision how some samples will be closely spaced (for example, two pulses in a row), but

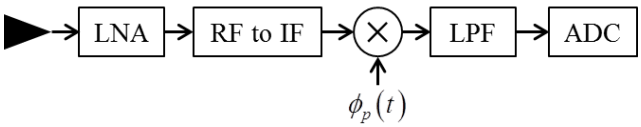


Figure 1. Block diagram of sub-Nyquist, fast-time compression implemented at an intermediate frequency (IF).

for fast-time compression, the thinning approach means that the ADC must occasionally collect samples at the full bandwidth of the signal. If the ADC must have the capability to sample at the full bandwidth, then the hardware advantages of compressive sampling disappear. Therefore, fast-time compression will typically be implemented with an ADC operating at a uniform sampling rate lower than the Nyquist rate. In order to avoid aliasing, the signal must be mixed with a non-localized measurement kernel before being sampled; therefore, this second form of compression is a "mixing" type compression that requires an analog multiplication. Hardware structures for sub-Nyquist sampling, including the random demodulator [2] and the modulated wideband converter [3] fall into this category of compression. A block diagram of an example RF compressive receiver is shown in Figure 1 where we can see the required elements including low-noise amplifier (LNA), downconversion from RF frequency to an intermediate frequency (IF) where analog multiplication with another wideband kernel can be performed, a lowpass filter to complete the projection, and finally sampling at a sub-Nyquist rate.

Mixing type compression can be represented in the projection notation above by replacing the localized delta sampling functions with an arbitrary measurement kernel according to

$$\begin{aligned} &\iint s(\mathbf{r},t)\phi(\mathbf{r},t)\,d\mathbf{r}dt \approx \\ &\qquad \Delta\mathbf{r}\Delta t\sum_{\mathbf{r}(i)}\sum_{t(j)} s(\mathbf{r}(i),t(j))\phi(\mathbf{r}(i),t(j)). \end{aligned} \tag{6}$$

If the compression is being performed via sub-Nyquist sampling of the signal captured by a particular antenna element, then the spatial component of the measurement kernel can be localized such that

$$\begin{aligned} &\iint s(\mathbf{r},t)\delta(\mathbf{r}-\mathbf{r}_p)\phi_p(t)\,d\mathbf{r}dt \approx \\ &\qquad \Delta\mathbf{r}\Delta t\sum_{\mathbf{r}(i)}\sum_{t(j)} s(\mathbf{r}(i),t(j))\delta[\mathbf{r}(i)-\mathbf{r}_p]\phi_p(t(j)) \end{aligned} \tag{7}$$

where $\phi_p(t)$ is the temporal mixing kernel applied to the receiving chain of the pth antenna element. For sub-Nyquist sampling on multiple antenna elements, the sensing matrix representation will be a composite of the thinning structure (due to elements that may or may not be present) with a structure that implements the non-localized temporal kernels.

Until a technology exists to implement the temporal modulation component of the measurement kernel directly at the antenna (using, for example, current distributions on an antenna element varying at the bandwidth of the incoming signal), the fast-time compression must be implemented with hardware such as analog multipliers, mixers, and filters as

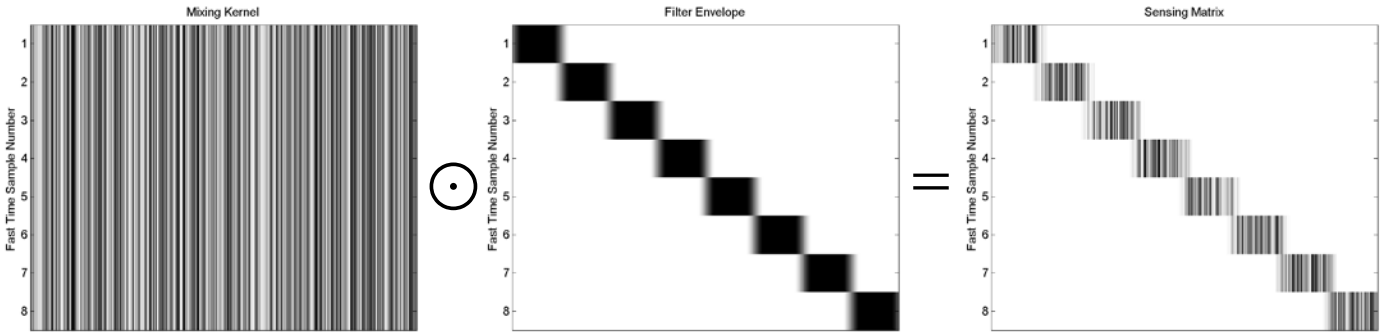


Figure 2. Structure of a sensing matrix for fast-time compression using analog multiplication followed by lowpass filtering.

depicted in Figure 1. There are two implications of the architecture in Figure 1. First, because the compression is performed after the signal has entered the receiver, additive receiver noise must be applied to the signal prior to the compression operation, which leads to noise folding [4], or more generally a loss in signal-to-noise ratio (SNR). Second, the time duration over which the signal is integrated (and, therefore, the length of any single projection) is determined by the time support of the lowpass filter's impulse response. Each successive sample taken by the ADC will be the result of integrating approximately T_c seconds of multiplier output where T_c is the approximate duration of the filter's impulse response. If the ADC sampling interval, T_{ADC}, is less than T_c, then successive samples will be partially correlated due to overlapping integration periods. Typically, we will set $T_{ADC} = T_c$ such that each sample is a result of an adjacent, and approximately non-overlapping, integration period.

Considering the above statements, (7) can be modified as

$$\begin{aligned}&\int\left[\left(s(\mathbf{r},t)\delta(\mathbf{r}-\mathbf{r}_p)\phi_p(t)+n_p(t)\right)*h(t)\right]d\mathbf{r}\\&=\int\left(s(\mathbf{r}_p,\tau)\phi_p(\tau)+n_p(\tau)\right)h(t-\tau)\,d\tau\\&\approx\Delta t\sum_{\tau(j)}\left(s(\mathbf{r}_p,\tau(j))\phi_p(\tau(j))+n_p(\tau(j))\right)h(t-\tau(j))\end{aligned}\quad,\quad(8)$$

which introduces a precise structure to the sensing matrix for each antenna element. This structure, which is depicted in Figure 2 (right panel) for compression down to eight ADC samples on a single antenna element, is a combination of the mixing kernel (left panel) and the LPF's impulse response shifted to implement the correct convolution output at the ADC sample times (middle panel). The discrete sensing model is then

$$\mathbf{y}_p=\mathbf{\Phi}_p\left(\mathbf{s}_p+\mathbf{n}_p\right)\quad(9)$$

where $\mathbf{s}_p$ is the temporally varying signal incident on the pth antenna, $\mathbf{n}_p$ is the additive noise on the pth receiving channel, and $\mathbf{\Phi}_p$ is the fast-time sensing matrix for the pth channel in the structure shown in Figure 2 according to the ADC rate, filter impulse response, and pth channel's mixing kernel $\phi_p(t)$. An overall space-time sensing matrix can then be expressed by concatenating the sensing matrices for individual channels according to a pattern along the lines of

$$\mathbf{\Phi}=\begin{bmatrix}\mathbf{\Phi}_1&\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\cdots&\mathbf{0}\\\mathbf{0}&\mathbf{0}&\mathbf{\Phi}_2&\mathbf{0}&\mathbf{0}&&\mathbf{0}\\\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{\Phi}_3&\mathbf{0}&&\mathbf{0}\\\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&&\mathbf{0}\\\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\ddots&\vdots\\\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\ddots&\mathbf{0}\\\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\cdots&\mathbf{\Phi}_P\end{bmatrix}\quad(10)$$

where blocks of columns containing all zeros are due to missing antenna elements that have been thinned from the system. The resulting measurements are

$$\mathbf{y}=\mathbf{\Phi}(\mathbf{s}+\mathbf{n})\quad(11)$$

where $\mathbf{s}$ and $\mathbf{n}$ have been formed by concatenating the signal and noise vectors for all antenna elements, including elements not sampled by the sensing matrix.

The model in (11) shows noise added prior to application of the sensing matrix (pre-projection noise model) for many reasons. First, even though the zero columns of the sensing matrix may result in a larger representation than necessary, the full representation reinforces the true input dimensionality of the space-time, Nyquist-sampled signal. Second, as described earlier, pre-projection noise is the correct representation for mixing-type compression implemented in analog hardware. Third, while faithfully representing mixing-type compression, the full representation also encompasses thinning-type compression as a special case. The sensing model in (11) can be expressed as

$$\mathbf{y}=\mathbf{\Phi s}+\mathbf{\Phi n}=\mathbf{\Phi s}+\hat{\mathbf{n}}\quad(12)$$

where the post-projection noise covariance matrix can be easily calculated from the pre-projection covariance and the sensing matrix. Therefore, the post-projection additive noise model used in some of the RF CS literature is valid in certain situations, but (12) explicitly shows that care must be taken in considering the post-projection noise statistics. If the input noise is uncorrelated and the rows of $\mathbf{\Phi}$ are orthogonal, then it is valid to go directly to a post-projection uncorrelated additive noise model, but in general, post-projection noise skips over a more fundamental starting point that may be helpful for proper treatment of system constraints and noise statistics. Finally, the full representation admits interpretation

of SNR loss due to compression as a loss in collected signal energy, as a noise folding behavior, or both. From the zero columns in (10), it is easy to see that for every measurement that is removed, collected signal energy is lowered. Radar systems are typically limited in transmit power and can't arbitrarily transmit additional power to make up for fewer samples. It is easy to ignore this loss or model it improperly when starting from a post-projection additive noise model.

III. Quadrature Compression

Many RF receivers implement quadrature reception where the down-conversion from RF (or IF) to baseband results in an in-phase (I) branch and a quadrature (Q) branch. These branches are 90 degrees out of phase with respect to each other such that signals with a random phase component are captured by one of the branches or a combination of the two. In this section we consider I/Q compression and its impact on the relationship between the I and Q signals.

A narrowband bandpass signal can be represented as

$$s(t) = a(t)\cos\left(2\pi F_0 t + \theta(t) + \theta_0\right) \tag{13}$$

where $a(t)$ and $\theta(t)$ are amplitude and phase modulations, respectively, with modulation bandwidths, B, much smaller than the carrier frequency F_0. The signal is assumed to have an unknown global phase θ_0. Without knowledge of θ_0, if we demodulate with only the cosine of the carrier, we risk demodulating with a carrier term that is out of phase with the received carrier, and the signal will be lost. Therefore, quadrature receivers demodulate against quadrature components of the carrier, guaranteeing signal capture regardless of global phase. However, because the receiver has I and Q branches, compression should be performed in each.

Figure 3 shows a potential architecture for a compressive quadrature receiver (hardware considerations may mean that the best design is a two-stage downconversion, but the format in Figure 3 is sufficient for the sampling analysis considered here). The incoming signal is split into two branches. The in-phase branch is demodulated with a cosine of the carrier and the quadrature branch is demodulated with a sine of the carrier. The first LPF in each branch has cutoff frequency at or above $B/2$ where B is the signal's bandpass bandwidth and is meant to reject all but the baseband copy of the signal. Next, the signal is mixed with a compression kernel, which might be a different kernel for the I and Q brances, followed by a LPF that completes the projection. The second LPF's in each branch have the same cutoff frequency, which is related to the sub-Nyquist sampling rate.

After passing the signal in (13) through the first mixer/LPF pair (downconversion step), the resulting signals are

$$\tilde{s}_I(t) = \frac{1}{2}a(t)\cos\left(\theta(t) + \theta_0\right) \tag{14}$$

and

$$\tilde{s}_Q(t) = \frac{1}{2}a(t)\sin\left(\theta(t) + \theta_0\right) \tag{15}$$

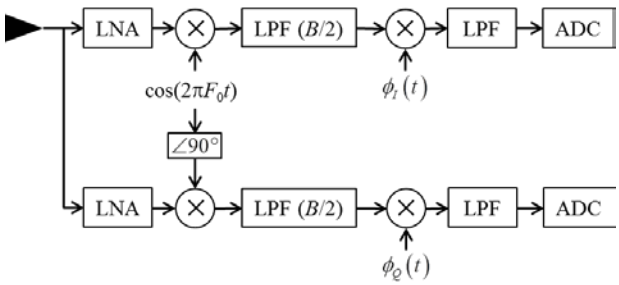


Figure 3. Compressive I/Q receiver architecture.

in the I and Q branches, respectively. These signals are then separately compressed by passing them through the second multiplier/LPF pair corresponding to the compression kernels in each branch, $\phi_I(t)$ and $\phi_Q(t)$. Letting the matrix-vector representations of (14) and (15) be $\tilde{\mathbf{s}}_I$ and $\tilde{\mathbf{s}}_Q$, respectively, the compressed data samples for the two branches are

$$\mathbf{y}_I = \mathbf{\Phi}_I\left(\tilde{\mathbf{s}}_I + \tilde{\mathbf{n}}_I\right) \tag{16}$$

and

$$\mathbf{y}_Q = \mathbf{\Phi}_Q\left(\tilde{\mathbf{s}}_Q + \tilde{\mathbf{n}}_Q\right). \tag{17}$$

In the absence of compression, the outputs of the two branches are typically treated as orthogonal and placed as the real and imaginary components of complex-valued data samples. Compression, however, decreases the distance between the two components. Therefore, the fundamental I/Q relationship may be altered, which may impact traditional processing steps such as envelope detection. Appropriate detectors and estimators for compressive quadrature receivers have yet to be fully developed in the literature.

IV. Conclusions

We have consider several aspects of compressive RF sensing, including appropriate signal and noise models for compression in different dimensions (slow time, fast time, spatial) and corresponding constraints on sensing matrices. We have also begun to consider a compressive version of the traditional quadrature receiver and the impacts that compression of in-phase and quadrature components may have on subsequent processing algorithms.

Acknowledgment

We acknowledge support from the Defense Advanced Research Projects Agency via grant #N66001-10-1-4079.

References

[1] L. Anitori, et al., "Design and analysis of compressed sensing radar detectors," *IEEE Trans. Sig. Proc.*, vol. 61, no. 4, pp. 813-827, Feb. 2013.

[2] J.A. Tropp, et al., "Beyond Nyquist: efficient sampling of sparse bandlimited signals," *IEEE Trans. Info.Theory*, vol. 56, no. 1, pp. 520-544 (2010).

[3] M. Mishali and Y.C. Eldar, "From theory to practice: sub-Nyquist sampling of sparse wideband analog signals," *IEEE J. Sel. Topics Sig. Proc.*, vol. 4, no. 2, pp. 375 – 391 (2010).

[4] E. Arias-Castro and Y. Eldar, "Noise folding in compressed sensing," *IEEE Sig. Proc. Letters*, vol. 18, no. 8, pp. 478-481 (2011).

Calibration—An open challenge in creating practical computational- and compressive-sensing systems

M.E. Gehm
Department of Electrical and Computer Engineering
University of Arizona
Tucson, AZ 85721
Email: gehm@ece.arizona.edu

***Abstract*—The goal of this manuscript (and associated talk) is not to present any recent experimental results from my laboratory. Rather, the purpose is to elucidate why I believe that *calibration* is one of the few remaining significant challenges in the struggle to create a wide range of practical computational sensing and compressive sensing (CS) systems. Toward this end, I briefly describe the fundamental and implementation difficulties associated with calibration as well as the existing calibration approaches and their associated limitations before sketching the theoretical question that must be addressed in order to solve the calibration challenge.**

I. Introduction

Computational sensing is the general term for a sensing approach in which estimation of the input signal $\mathbf{x}$ proceeds from a set of measurements $\mathbf{y}$ that result from the action of a linear measurement operator $\mathbf{H}$ (including the possibility of potential noise corruption). The specific form of the measurements depends on the physical nature of the system and the noise. For example $\mathbf{y} = \mathbf{H}\mathbf{x}+\mathbf{n}$ is the appropriate form for an optical system with post-measurement additive noise $\mathbf{n}$, while $\mathbf{y} = \mathbf{H}(\mathbf{x}+\mathbf{n})$ is the appropriate form for an RF system with pre-measurement additive noise (such as that which arises at the input to the first-stage amplifier).

Regardless of the specific form, for traditional isomorphic sensor systems operating with impulse-like sampling, the measurement operator (matrix) $\mathbf{H}$ is the identity matrix $\mathbf{I}$. Computational sensing then generalizes this to consider sensor systems that implement measurement matrices $\mathbf{H}$ that have *non-zero off-diagonal elements*. In this manner, the measurements $\mathbf{y}$ become *multiplexed* and estimation of $\mathbf{x}$ becomes a non-trivial inverse problem. In this picture, *compressive sensing* can then be described as a subset of computational sensing where the sensing matrix $\mathbf{H}$ not only has off-diagonal elements, but is also rectangular with fewer rows than columns. Thus, the number of acquired measurements in $\mathbf{y}$ is less than the number of native signal elements in $\mathbf{x}$.

A. The Importance of Calibration

Solution of the inverse problem—that is, estimation of the input signal $\mathbf{x}$ from the measurements $\mathbf{y}$ requires knowledge of the measurement matrix $\mathbf{H}$. While the measurement system will have been designed to implement a specific measurement matrix $\mathbf{H}_{\text{des}}$, experimental reality dictates that the implemented matrix $\mathbf{H}_{\text{imp}}$ will deviate from the design to some extent. An example of the possible deviation between $\mathbf{H}_{\text{des}}$ and $\mathbf{H}_{\text{imp}}$ for an experimental system is shown in Fig. 1. For reasons outlined below, high-quality recovery of $\mathbf{x}$ is frequently sensitive to these variations. Determining the actual form of $\mathbf{H}_{\text{imp}}$ is then the role of *calibration*.

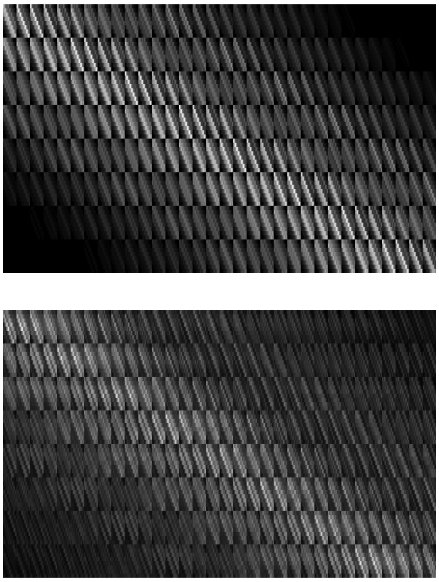


Fig. 1. The designed (top) and as-implemented (bottom) measurement matrices for an experimental compressive tracking system [1]. The implemented version of the matrix is an estimate created via a calibration process.

The sensitivity of system performance with respect to small deviations between $\mathbf{H}_{\text{des}}$ and $\mathbf{H}_{\text{imp}}$ can be understood by considering the multiplex nature of the measurement matrix. As mentioned above, the distinguishing feature of computational and compressive sensing approaches is that their measurement matrices contain non-zero off-diagonal elements. As a result, multiple signal elements are multiplexed together in each of the measurements. In cases where the input signal is dense in the native basis, this has the effect of encoding information about the input signal $\mathbf{x}$ into the variations of $\mathbf{y}$ about its mean.

This mean value (or *baseline*) frequently utilizes a significant fraction of the available system dynamic range, limiting the dynamic range available for the variations—where the information about $\mathbf{x}$ is encoded. An example is shown schematically in Fig. 2. This issue is analogous to the *interfero-*

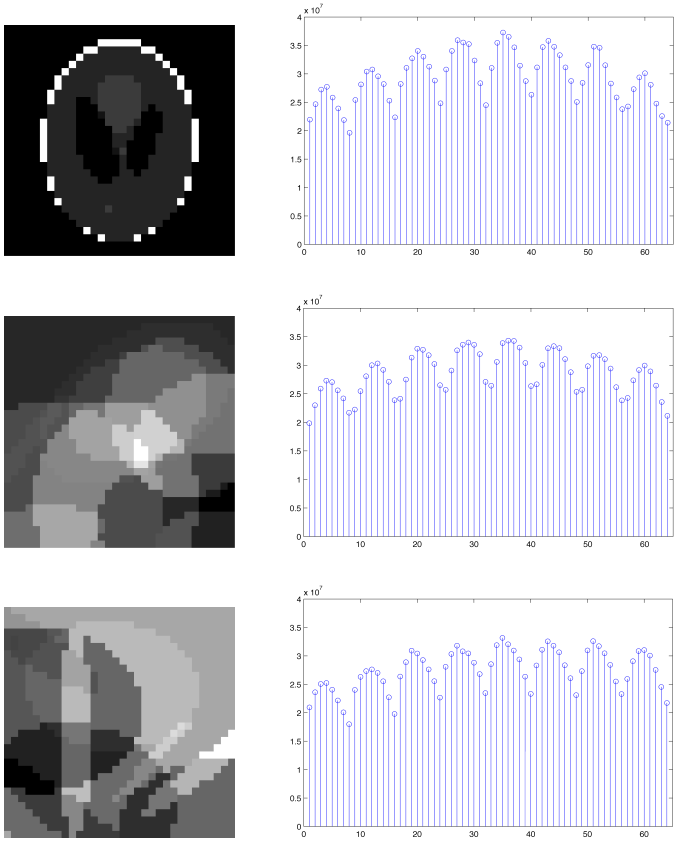


Fig. 2. An example of the multiplexing baseline. (Left) Three phantoms of size 32×32 composed of superposed partially-transparent ellipses. The mean value of the three phantoms is the same. (Right) Plots of the measurements produced by the phantoms when sampled with the measurement matrix in Fig. 1 (bottom). Information regarding the structure of the phantoms undergoes dynamic range compression and is encoded in the variations about the measurement mean.

metric baseline problem that arises in interferometric systems and is the primary manifestation of the so-called *multiplex disadvantage*.

It is true that the effect is mitigated somewhat in systems which are sparse in the native basis. However, practical situations that are natively sparse are rare (hence the need for the sparsifying transform in the majority of compressive sensing implementations).

The net effect of the dynamic range compression that results from the multiplex measurement that converts $\mathbf{x}$ into $\mathbf{y}$ is that accurate reconstruction becomes more sensitive to the specific form of $\mathbf{H}$ that encoded the measurements—thus driving the preference for $\mathbf{H}_{\text{imp}}$ (which is determined via calibration) over $\mathbf{H}_{\text{des}}$ (which is known from the intended system design).

B. Difficulties of Direct Calibration

The most direct calibration approach is what we might term *point-by-point*, and is effectively a Green's function (shift-variant impulse response) approach. The experimenter sequentially energizes each of the individual signal elements with unit amplitude. For each, the system response is recorded and then placed sequentially as the columns of a matrix. Once every signal element has been probed in this manner and the results integrated into the matrix, the result is an estimate of true measurement matrix $\mathbf{H}_{\text{imp}}$.

Although the point-by-point calibration approach is admirably direct, there are a number of potential difficulties that limit its practicality:

1) **Signal response too weak:** Energizing a single signal element at a time may produce a system response that is swamped by noise. Here we directly encounter the fact that *calibration is itself a measurement process*. Specifically, direct calibration estimates the measurement matrix $\mathbf{H}_{\text{imp}}$ via a traditional isomorphic approach where the estimate of each column of $\mathbf{H}_{\text{imp}}$ is taken as the measured system response for the corresponding single-element excitation. Any measurement noise is directly imposed on the estimate and may be non-negligible.
2) **Too many signal elements:** As computational and compressive sensing is applied to broader ranges of systems, the dimensionality (number of native signal elements) continues to grow. In the most advanced systems, the number of signal elements is of such a size that direct calibration is no longer practical—the time required is either beyond the patience of the experimentalist, or is on a timescale that is comparable with the timescales over which the measurement matrix $\mathbf{H}_{\text{imp}}$ varies (e.g. via thermal drift). For example, a compressive spectral imaging system under construction in my laboratory [2] has $\approx 8.4 \times 10^6$ native signal elements. Making the (optimistic) assumption that the apparatus will allow direct calibration at a rate of 10 Hz, we see that direct calibration would complete in just under 10 days of continual operation!
3) **Lack of desired control:** The direct approach requires the ability to isolate single signal elements and to control their amplitude. It is frequently the case that the experimentalist does not have a source that provides this level of control. For example, in a computational spectroscopic application, a tunable, narrowband spectral source may not be available. Instead, the experimenter may have access to a discrete set of spectral sources, each with a unique spectral profile that is a linear combination of the individual spectral channels.

II. Existing Non-Direct Methods and Their Limitations

The realization (mentioned above) that calibration is itself a measurement process potentially provides the key to overcoming the pitfalls inherent in direct calibration. Direct calibration represents a traditional, isomorphic measurement approach to estimating the measurement matrix $\mathbf{H}_{\text{imp}}$. That is, if we imagine lexicographically reordering the elements of $\mathbf{H}_{\text{imp}}$ into a vector $\mathbf{H}_{\text{imp,vec}}$, the calibration process utilizes a measurement matrix $\mathbf{\Phi}_{\text{cal}}$ to capture the calibration measurements. For an optical system with post-measurement additive noise, this would have the form $\mathbf{y} = \mathbf{\Phi}_{\text{cal}}\mathbf{H}_{\text{imp,vec}} + \mathbf{n}$, with the

obvious extensions to other measurement models. For direct calibration, $\mathbf{\Phi}_{\text{cal}}$ is the identity matrix $\mathbf{I}$.

As with the original measurement problem, however, we can apply computational or compressive sensing ideas to the calibration process and consider more general forms of $\mathbf{\Phi}_{\text{cal}}$. There are a number of existing calibration approaches that make this generalizing step. The following subsections describe these approaches, their benefits with respect to direct calibration, and their limitations.

A. Multiplexed Calibration

Multiplexed calibration simultaneously energizes multiple signal elements for each measurement in the calibration process, resulting in a $\mathbf{\Phi}_{\text{cal}}$ that contains non-zero off-diagonal elements [3]–[5]. Estimation of $\mathbf{H}_{\text{imp,vec}}$ then proceeds through the solution of an inverse problem. If the column rank of the resulting $\mathbf{\Phi}_{\text{cal}}$ is equal to the number of native signal elements, then traditional algorithms can be brought to bear to yield the estimate. If the column rank is smaller than the number of native signal elements, compressive sensing techniques are more appropriate.

The plausibility of compressive methods for determining $\mathbf{H}_{\text{imp,vec}}$ can be understood by examining the structure of measurement matrices such as Fig. 1 (bottom) and noting the large degree of structure present. Obviously, this structure is a form of redundancy that indicates that $\mathbf{H}$ is fundamentally a lower-dimensional object than the native number of matrix elements would suggest. Note that this argument would not hold for *random* measurement matrices (although structure imposed as a result of implementation deviations would provide some reduction in the dimensionality)—a severe downside to random measurement in extremely high-dimensional systems.

1) Advantages:

- Multiplex calibration combines signal elements in every measurement. For systems dominated by *additive noise*, this reduces the impact of the noise, increasing the measurement SNR.
- *Compressive* multiplexed calibration—where the column rank of $\mathbf{\Phi}_{\text{cal}}$ is less than the number of native signal elements—reduces the number of measurement acquisitions and hence the required calibration time. This may prove helpful in situations where direct calibration is unfeasible as a result of the number of signal elements.

2) Drawbacks:

- Multiplex-based improvement in measurement SNR does not occur in systems that are *Poisson (shot) noise* limited. The mean SNR of such measurements remains constant upon multiplexing.
- Estimation via solution of the inverse problem results in *transform noise*—The total noise in the measurements is redistributed among the estimated signal elements in ways that can radically modify the noise statistics. For example in Poisson noise limited systems, noise is preferentially redistributed from strong to weak areas of the signal. This produces sub-Poisson noise statistics in the strong signal areas and super-Poisson statistics in the weak signal areas. Transform noise also frequently introduces correlations between the noise present at different signal elements, creating the appearance of structure when none is truly present. This redistribution results in errors in the estimated $\mathbf{H}_{\text{imp,vec}}$ that can potentially impact system performance.

B. Matrix Completion

A closely-related approach applies the techniques of *matrix completion* [6]–[8] to the problem. In this approach, assumptions regarding the low-rank nature of the measurement matrix $\mathbf{H}_{\text{imp}}$ allow its full structure to be estimated given knowledge of only a subset of its entries. In a recent paper, Vetterli et al. explore the use of matrix completion methods to the calibration problem in ultrasound imaging [9] and obtain promising results.

1) Advantages:

- The method is well-matched to the central task at hand—estimating a low-rank (structured) matrix from a set of possibly incomplete calibration measurements. In some cases this would allow the experimenter to achieve an accurate estimation of $\mathbf{H}_{\text{imp}}$ from a reduced number of calibration measurements and hence shorten the required calibration time.

2) Drawbacks:

- Matrix completion is generally posed in the context of *missing entries* that are distributed randomly throughout the matrix (see [9] for an example). This is suitable for systems where the output state of the system must be *sequentially acquired* in order to determine the full system response to a given calibration input. For systems where the output measurements are acquired in parallel, however, *skipping a calibration step* (to shorten the calibration time) would result in missing entries that are not arranged randomly throughout $\mathbf{H}_{\text{imp}}$, but rather are organized in *columns*, and existing algorithms perform poorly in this situation. Performing matrix completion after a basis change to redistribute the missing entries may possibly restore performance, but I am not aware of any work in that area.
- In their current form, matrix completion methods assume elements of the matrix are known in the native basis—implying single element excitation. This suffers from the same potential SNR issues as direct calibration. The previously mentioned idea of performing matrix completion after a basis change, should such an approach prove viable, would allow (require) multiplexed excitation. This would increase measurement SNR for cases which are limited by additive noise.

C. Parameterized Forward Model

The final (and most common) method of calibration is to create a *parameterized forward model*. In essence, it seeks to estimate $\mathbf{H}_{\text{imp}}$ through the creation of a more sophisticated $\mathbf{H}_{\text{des}}$. The system model is extended to include possible errors that could arise during implementation and the magnitude of

these errors are incorporated as *adjustable parameters*. Calibrating such a system is then a matter of determining certain experiments (excitation patterns) which reveal the appropriate magnitudes for these parameters.

1) Advantages:

- A well-developed parameterized forward model is extremely powerful. It incorporates a significant amount of prior knowledge regarding the intended structure of $\mathbf{H}_{\text{des}}$ as well as the physics of the likely effects that transform $\mathbf{H}_{\text{des}}$ into $\mathbf{H}_{\text{imp}}$. The resulting number of parameters captures the underlying dimensionality of $\mathbf{H}_{\text{imp}}$ with admirable efficiency.

2) Drawbacks:

- The parameterized forward model approach *trades* calibration acquisition time for model development time. The net benefit of this trade-off (if any) depends on the skill and insight of the person developing the model.
- Model mismatch is a serious concern; the model only incorporates terms that are explicitly included. Deviations between $\mathbf{H}_{\text{des}}$ and $\mathbf{H}_{\text{imp}}$ that arise from implementation errors that are not included in the model will not be captured during the calibration process.
- In advanced, high-dimensional systems, the number of necessary parameters can proliferate quickly (in correspondence with the increasing dimensionality of the underlying measurement structure of the instrument). The resulting models can become unwieldy and design of experiments to isolate the values of individual parameters may become difficult or impossible.

III. What is Needed

Although there are a number of non-direct calibration methods now in use, each has its own unique balance of advantages and drawbacks and none of the methods is ideal. In this section, I attempt to describe the properties that an ideal calibration approach would have and the theoretical questions that must be addressed in order to develop such an approach.

Over the past several years, there has been an evolution in the field of compressive sensing that emphasizes a move from random to designed sensing matrices. The various design strategies incorporate prior information regarding the statistical distributions of likely input signals, the nature of the sensing task, and the reconstruction/estimation algorithms that will be brought to bear. This design is then performed subject to the constraints of both physics and system architecture. The ideal calibration framework would provide a similar level of design by identifying the sequence of calibration measurements to be made subject to a variety of priors and constraints. Fundamental questions related to this goal include:

- What is the appropriate mathematical framework for the design of calibration sequences? Is there a mathematical reason to prefer the matrix, $\mathbf{H}_{\text{imp}}$ (matrix completion) or vector, $\mathbf{H}_{\text{imp,vec}}$ (computational/compressive sensing) form to the problem?
- Can constraints on the available input signals be incorporated in the design process? What if only a fixed set of inputs are possible?
- The experimenter will have general knowledge regarding the approximate form of $\mathbf{H}_{\text{imp}}$ (via knowledge of $\mathbf{H}_{\text{des}}$). How can this prior knowledge be incorporated into the design process?
- Clearly not all errors in estimating $\mathbf{H}_{\text{imp}}$ will be equivalent. The effect of specific errors is likely to depend on the ultimate *sensing task* of the system. How can prior information regarding this task be incorporated into the design of the calibration sequence? How can prior information regarding the likely input signals (in the course of the sensor task, not calibration) be incorporated into the design of the calibration sequence?
- Can the framework be made adaptive? Can the results of early stages in the calibration sequence influence the design of subsequent calibration measurements?
- Are there fundamental limits or guarantees that can be stated about designed calibration?

IV. Conclusion

Calibration is currently an open challenge with regards to developing advanced compressive and computational sensing systems. The fact that calibration is itself a measurement process provides a key opening through which to attack this problem. It is my hope that the rough ideas presented here can spark an engagement between the theoretical and experimental communities on this crucial issue.

Acknowledgment

The author gratefully acknowledges the support of the Defense Advanced Research Projects Agency (DARPA) through the Knowledge-Enhanced Compressive Measurement (KECoM) program (Grant #N66001-10-1-4079).

References

[1] D. Townsend, P. Poon, S. Wehrwein, T. Osman, A. Mariano, E. Vera, M. Stenner, and M. Gehm, "Static compressive tracking," *Optics Express*, vol. 20, no. 19, pp. 21 160–21 172, 2012.

[2] M. Dunlop, P. Jansen, D. R. Golish, and M. E. Gehm, "Afssi-c: the adaptive feature-specific spectral imaging classifier," in *Computational Optical Sensing and Imaging*. Optical Society of America, 2012.

[3] S. D. Silverstein, "Application of orthogonal codes to the calibration of active phased array antennas for communication satellites," *Signal Processing, IEEE Transactions on*, vol. 45, no. 1, pp. 206–218, 1997.

[4] R. D. Batten, A. Eshraghi, and T. S. Fiez, "Calibration of parallel $\delta\sigma$ adcs," *Circuits and Systems II: Analog and Digital Signal Processing, IEEE Transactions on*, vol. 49, no. 6, pp. 390–399, 2002.

[5] M. Kasper, E. Fedrigo, D. P. Looze, H. Bonnet, L. Ivanescu, and S. Oberti, "Fast calibration of high-order adaptive optics systems," *JOSA A*, vol. 21, no. 6, pp. 1004–1008, 2004.

[6] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational mathematics*, vol. 9, no. 6, pp. 717–772, 2009.

[7] E. J. Candès and T. Tao, "The power of convex relaxation: Near-optimal matrix completion," *Information Theory, IEEE Transactions on*, vol. 56, no. 5, pp. 2053–2080, 2010.

[8] E. J. Candes and Y. Plan, "Matrix completion with noise," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 925–936, 2010.

[9] A. Karbasi, S. Oh, R. Parhizkar, and M. Vetterli, "Ultrasound tomography calibration using structured matrix completion," in *International Congress on Acoustics (ICA2010)*, 2010.

Compressive CFAR Radar Processing

Laura Anitori, Wim van Rossum and Matern Otten
TNO, The Netherlands
Email: laura.anitori@tno.nl

Arian Maleki
Columbia Univeristy
New York City, USA

Richard Baraniuk
Rice University
Houston, USA

***Abstract*—In this paper we investigate the performance of a combined Compressive Sensing (CS) Constant False Alarm Rate (CFAR) radar processor under different interference scenarios using both the Cell Averaging (CA) and Order Statistic (OS) CFAR detectors. Using the properties of the Complex Approximate Message Passing (CAMP) algorithm, we demonstrate that the behavior of the CFAR processor is independent of the combination with the non-linear recovery and therefore its performance can be predicted using standard radar tools. We also compare the performance of the CS CFAR processor to that of an ℓ_1-norm detector using an experimental data set.**

I. Introduction

Compressive Sensing (CS) is a novel data acquisition scheme that enables reconstruction of sparse signals from highly undersampled measurements. In many radar applications, such as air traffic control, obstacle avoidance, and wide area surveillance, it is reasonable to assume that the scene is sparse, since the number of targets is much smaller than the number of resolution cells in the illuminated area. Examples of CS applied to radar can be found in [1]–[5].

However, while classical radar architectures use well-established processing algorithms and detection schemes, such as Matched Filtering (MF) and Constant False Alarm Rate (CFAR) detectors, the reconstruction of the target scene from the CS measurements involves the use of highly nonlinear algorithms such as ℓ_1-norm minimization. These algorithms have a number of parameters that must be tuned properly in order to achieve good performance. The optimal value of the parameters depend on both the underlying noise power and the number of non-zero coefficients. Hence, in a practical scenario, where neither the disturbance variance nor the number of targets are known a priori, it is not clear how to tune these parameters to achieve the desired performance.

In most operational radars, to deal with the uncertainties about the background and the interference scenario, CFAR processors are widely used for adaptive target detection. Several CFAR schemes have been designed to attain good performance in the presence of different types of clutter and target scenarios [6]–[8]. The modeling and prediction of False Alarm Probability (FAP) is essential for the design of CFAR schemes. This in turn requires some level of knowledge of the underlying noise (or clutter) distribution that is input to the detector. Designing CFAR schemes seem to be out of reach for CS radar systems, due to the so far unknown relations between FAP/noise statistics and the parameters involved in the ℓ_1-norm reconstruction.

In [5], [9] we show that, using the properties of the Complex Approximate Message Passing (CAMP) [10], CFAR processing can be combined with ℓ_1-minimization to obtain fully adaptive detection schemes. In this paper, we further investigate the performance of the joint CS CFAR detector in combination with both the Cell Averaging (CA) and the Order Statistic (OS) CFAR processors under different interference scenarios using a set of CS radar measurements.

II. Complex Approximate Message Passing (CAMP)

In CS, we are concerned with the problem of recovering a k-sparse signal $\mathbf{x_0} \in \mathbb{C}^N$ from an undersampled set of linear measurements $\mathbf{y} \in \mathbb{C}^n$ of the form

$$\mathbf{y} = \mathbf{A}\mathbf{x_0} + \mathbf{n}, \tag{1}$$

where $\mathbf{A} \in \mathbb{C}^{n \times N}$ is the sensing matrix, and $\mathbf{n}$ is complex white Gaussian noise with variance σ_{in}^2. Let $n < N$ and define $\delta = n/N$ and $\rho = k/n$.

Since the number of measurements n is smaller than the number of signal samples N, the problem of recovering $\mathbf{x_0}$ is ill-posed. However, under certain conditions on $\mathbf{A}$, n, and k the following convex optimization problem, known in the literature as the LASSO [11] or Basis Pursuit Denoising (BPDN) [12], recovers a close approximation of $\mathbf{x_0}$ [13], [14]:

$$\widehat{\mathbf{x}} = \min_{\mathbf{x}} \frac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda \|\mathbf{x}\|_1, \tag{2}$$

where λ is a regularization parameter that controls the trade off between the sparsity of the solution and the ℓ_2-norm of the residual. Finding the "optimal" value of λ is a major practical problem when dealing with CS reconstruction algorithms. In particular, for radar applications the relations between the parameter λ and the detection and false alarm rates are unknown.

The Complex Approximate Message Passing (CAMP) is an iterative algorithm for solving (2) for signals in the complex domain.[1] Interestingly, the CAMP algorithm has a number properties that enable us to solve both the problem of optimal tuning and adaptive target detection. These properties are summarized in P1–P3 [10], [15], [16]:

P1: Under an appropriate tuning of the regularization parameter used in CAMP and the parameter λ in (2), CAMP solves LASSO exactly. See Section 3.4 in [10].

[1] A detailed description of the algorithm and its properties can be found in [5], [9], [10].

P2: At every iteration, $\tilde{\mathbf{x}}^t$ can be considered as $\mathbf{x_0}+\mathbf{w}^t$, where the distribution of $\mathbf{w}^t$ converges to complex Gaussian with zero mean and variance σ_t^2. See Section 3.4 in [10].

P3: The performance of CAMP can be predicted theoretically by the so-called state evolution equation. See Section 3.1 in [10].

An important relation derived from the analytical framework used in CAMP is that the variance of the total noise σ_t^2 present in the signal $\tilde{\mathbf{x}}$ at each iteration t is expressed as a linear combination of the input noise variance and the MSE of the solution: $\text{MSE}_t=\frac{\|\hat{\mathbf{x}}^t-\mathbf{x_0}\|^2}{N}$.[2] In CAMP an estimate $\hat{\sigma}_t^2$ of the noise variance is computed at each iteration by means of median filtering.

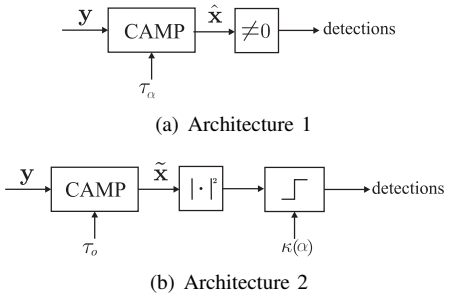
Also, using the signal-plus-noise model described in P2, the problem of tuning the regularization parameter in CAMP, which we refer to as τ, can be easily solved. Amongst all τ, the optimal threshold τ_o in CAMP is the one that achieves the minimum MSE or, equivalently, the minimum σ_∞^2. For the practical case of unknown signal and noise statistics, we can use the *Adaptive CAMP* algorithm described in [9] to obtain a good estimate $\hat{\tau}_o$ of the optimal threshold multiplier τ_o. The optimum estimated threshold $\hat{\tau}_o$ is the one that minimizes the estimated CAMP output noise variance. This choice, in turn, also maximizes the recovery SNR of CAMP.

III. CS Target Detection using CAMP

In radar, the detection problem is to determine the presence or absence of a target in a given range/Doppler bin when the received signal is corrupted by noise and clutter. In practice, both the noise and clutter power are unknown a priori, and therefore an adaptive detection scheme must be designed. Also, it is desirable that the detector has the CFAR property. We consider here two different CS CAMP based architectures, whose block diagrams are shown in Figure 1.

In the first system, the CS reconstruction is considered as the detector itself. This implies that in CAMP we should set the threshold, say τ_α, such that the desired FAP α is achieved. It is shown in [5] that for complex signals, if $\mathbf{x_0}=0$, then setting the CAMP threshold $\tau_\alpha=\sqrt{-\ln\alpha}$ results in a FAP equal to α. We will refer to this detection strategy as Architecture 1; its block diagram is shown in Figure 1(a).

However, theoretical and empirical results [9] show that better performance can be achieved in terms of detection (P_d) and false alarm probability (P_{fa} or FAP) if the CS recovery is followed by a second detector. This means that, just as in conventional radar processing, we can use the recovery stage (a Matched Filter (MF) in classical architectures) to maximize the recovery SNR (i.e., maximize detection for a given FAP), and later use the detector to obtain the desired FAP. In this case the CAMP threshold is selected to achieve the minimum MSE at the output of CAMP, i.e., $\tau=\tau_o$, and τ_o is estimated using Adaptive CAMP. We refer to this scheme as Architecture 2. In this architecture, the input to the detector is the signal $\tilde{\mathbf{x}}$. According to P2, this signal can be modeled as the sum of the true observable $\mathbf{x_0}$ plus Gaussian noise.



(a) Architecture 1

(b) Architecture 2

Fig. 1. Detection schemes based on CAMP. Note that in Architecture 2 the output of CAMP is the noisy version of the estimated signal $\tilde{\mathbf{x}}$.

Thanks to the statistical properties of CAMP summarized in P1–P3, the CAMP thresholds can now be set, adaptively in Architecture 2 and as a function of the FAP for Architecture 1.

Ideally, if the noise statistics were homogeneous, stationary and known, the detector threshold in Architecture 2 could be set once and remain fixed. This represents the ideal case of a fixed threshold (FT) detector. In practice, however, these conditions are never satisfied and CFAR processors are employed to adaptively estimate the detector threshold $\kappa(\alpha)$ when the noise statistics are not known in advance. In CFAR schemes the cell under test (CUT) is tested for the presence of a target against a threshold that is derived based on an estimated clutter plus noise power. The cells surrounding the CUT (CFAR window) are used to derive an estimate of the background and they are assumed to be target free. The great advantage of CFAR schemes is that they are able to maintain a constant false alarm rate via adaptation of the threshold to a changing environment. It is known that for the case of homogeneous Gaussian background, the optimum CFAR processor is the well-known Cell Average CFAR (CA-CFAR) detector [6]. However, in situations in which the clutter changes rapidly or in the presence of interfering targets in the CFAR window, or when the clutter and noise distribution are not Gaussian, the CA-CFAR detector performance degrades severely. For this reason many alternative CFAR schemes have been developed in the past, such as the Order Statistic (OS) CFAR detector [7], [8]. In OS-CFAR processing, the power received from the cells in the CFAR window are rearranged in increasing order and the kth ordered cell (order statistic) is used as an estimate of the environment. OS-CFAR processing has the advantage of being robust against interfering targets in the CFAR window and clutter power transitions, while preserving reasonably good performance in homogenous background.

[2]Specifically, for the case of Gaussian sensing matrices, at the fixed point solution ($t\to\infty$), the relation $\sigma_\infty^2=\sigma_{in}^2+\frac{1}{\delta}\text{MSE}_\infty$ holds; see [10], [15] for a more detailed analysis on the (C)AMP input/output relations. From the previous equation, it is clear that the CAMP total output noise power is the sum of the effective system noise plus noise introduced by the recovery itself. Consequently, for a given input SNR, minimizing MSE also minimizes the output noise variance and therefore maximizes the reconstruction (or recovery) SNR.

IV. Experimental Data

In the this section, we compare the performance of the proposed detection schemes under different interference scenarios using a set of experimental CS radar measurements.

A. Experimental Set-up

In our experiments, we consider the case of a one dimensional radar operating in the range domain. We use as targets five stationary corner reflectors with different Radar Cross Sections (RCS). For each transmitted waveform 300 measurements (with the same set-up) were performed.

The measurements were carried out at Fraunhofer FHR, in Germany, using the LabRadOr experimental radar system described in [5]. We used a stepped frequency (SF) waveform and the TX signal consists of a number of discrete frequencies f_m. In the Nyquist case (that represents unambiguous mapping of ranges to phases over the whole bandwidth) we transmit $N = 200$ frequencies over a bandwidth of 800 MHz. The achievable range resolution is therefore $\delta_R = 18.75$ cm. Each frequency is transmitted during 0.512 μs, corresponding to a bandwidth of $B_f = 1.95$ MHz, and sequential frequencies are separated by $\Delta f = 4$ MHz, resulting in an unambiguous range of $\Delta R = 37.5$ m.

In the CS case, the number of TX frequencies is reduced from N to n ($n < N$). The subset of transmitted frequencies is chosen uniformly at random within the total transmitted bandwidth, with the constraints that we always use the first and last frequencies in the bandwidth (to span the same total bandwidth to preserve range resolution), and we also force at least two of the transmitted frequencies to be separated by the nominal frequency separation Δf, to guarantee that the unambiguous range is preserved.

After reception and demodulation each range bin maps to n phases proportional to the n transmitted frequencies, and the n samples y_m, $m = 1, \cdots, n$, of the compressed measurement vector $\mathbf{y}$ are given by

$$y_m = \frac{1}{\sqrt{n}} \sum_{i=1}^{N} e^{-j4\pi f_m r_i / c} x_{0,i}, \tag{3}$$

where $r_i = r_0 + i\,\Delta R/N$, and $i = 1, \ldots, N$ is the range bin index.

B. Results

In this section, we use ROC curves to analyze the performance of the the two CAMP based detection schemes for both interfering and non-interfering target scenarios, which we obtain by changing the CFAR window size. For Architecture 2, we combine the CAMP recovery with both the CA and OS CFAR processors.

Figure 2 exhibits the signals reconstructed by using the two CAMP based architectures introduced in Section III in addition to the MF, which represent the reference case. We use $\delta = 0.5$ for the CS measurements and $N = 200$ measurements for the MF. There are five corner reflectors (T1–T5) at ranges from 20m to 36m. For Architecture 1, τ_α was set using $\alpha = 10^{-4}$.

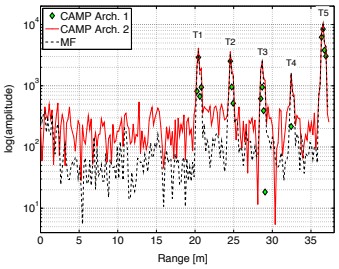


Fig. 2. Reconstructed range profile using CAMP Architectures 1 and 2, and the MF. For the MF, $N = 200$ (i.e., no subsampling); for all other schemes $n = 100$ and $\delta = 0.5$. The y-axis is in log scale and arbitrary units [au].

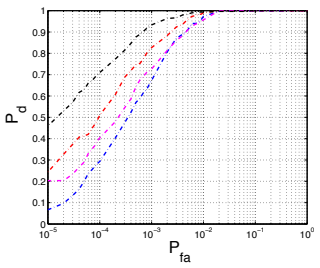
Notice that the signal from Architecture 2 is a noisy version of the estimated sparse signal before soft thresholding is applied, whereas the signal estimated from Architecture 1 is the sparse signal, where each non-zero coefficient represents a detection.

In the interest of space, we report only the ROC curves for target T3. For the other targets, the behavior of the detectors is the same, although the actual values of P_d are different due to the different SNRs of both the desired target and the interferers. For estimating P_d, we used the detection at the location of the highest target peak. For Architecture 2 we use both the CA and OS CFAR processors, preceded by a Square Law (SL) detector. For the CFAR processors, we use 4 guard cells and 3 different CFAR windows of length 20, 40 and 90 respectively. For the OS-CFAR, the selected order statistic is chosen as $k_{OS} = 0.6\%$.

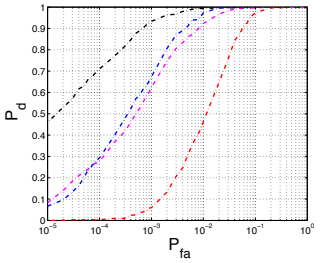
Note that for all detector cases (adaptive and non-adaptive), the CAMP reconstruction threshold τ_o of Architecture 2 is always adaptive, whereas in Architecture 1 the threshold τ_α is non-adaptive and fixed. Furthermore, the performance of the two architectures are upper bounded by the performance of Architecture 2 that uses an ideal (non-adaptive) fixed threshold (FT) detector instead of a CFAR one.

Figure 3(a) shows the ROC curve for T3 with a CFAR window of length $M = 20$. For this choice of M, none of the other targets fall in the CFAR window of T3, and therefore the CA-CFAR processor performs better than the OS one. Furthermore, we can see that it also outperforms Architecture 1, where the noise variance is estimated inside the CAMP algorithm using the median estimator. Therefore, Architecture 1 is similar to an OS processor that uses the entire range as the CFAR window and $k_{OS} = 0.5$. Clearly, in this case the CA-CFAR performs better than both the OS and Architecture 1, since it excludes the other targets from the (local) estimation of the noise level, therefore resulting in an unbiased estimate. For this window size, CA is the best choice since there are no noise/clutter power transition, and the targets are never in the reference window of one another.

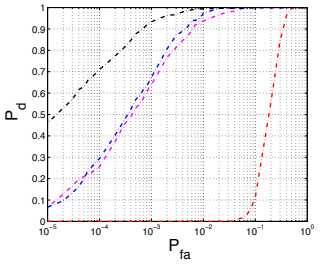
Figures 3(b) and 3(c) show the results for the same data set but for CFAR windows of sizes 40 and 90, which result in 2 and 3 interferers in the CFAR window of the target of interest. In both cases we observe that, in accordance with conventional CFAR processing, Architecture 2 with OS-CFAR outperforms



(a) CFAR window length $M = 20$



(b) CFAR window length $M = 40$



(c) CFAR window length $M = 90$

Fig. 3. ROC curves for T3 using Architecture 1 (blue), and Architecture 2 in combination with FT detector (black), CA-CFAR (red), and OS-CFAR (magenta) processors. For the OS-CFAR processor, $k_{OS} = 0.6M$. $\delta = 0.5$.

Architecture 2 with CA-CFAR but performs very similarly to Architecture 1. Furthermore, the performance of the CA-CFAR processor degrades as the number of interfering targets in the reference window increases. Note that the ROC curve of Architecture 1 (and also Architecture 2 with FT detector) is unchanged for different CFAR window sizes. In fact, for Architecture 1 we do not use a CFAR processor and the CAMP reconstruction is independent of the locations of the targets, as it uses the whole range response. It is clear that, in cases where there might be multiple interfering targets either an OS-CFAR processor should be used after Architecture 2 or otherwise the theoretically suboptimum Architecture 1 can represent a simple, effective alternative to CFAR processing. However, the disadvantage of Architecture 1 is that it lacks the local adaptivity provided by CFAR processing. Clearly, there is a trade off between the number of range bins used for the noise power estimation and the bias in the estimate that can be caused by including in the reference window interfering targets and /or noise and clutter power transitions.

V. Conclusions

In this paper we compare the results of different CS based radar detection architectures. From the experimental results we conclude that:

- the combination of CS with standard CFAR processing does not alter the behavior of the CFAR processor compared to the case when this is used in combination with a standard MF;
- in the presence of interfering targets in the CFAR window, as expected, OS is better than CA-CFAR processing;
- although the performance of Architectures 1 and Architecture 2 plus OS-CFAR are similar, Architecture 2 seems to be preferable as it leaves the user the freedom to chose the most appropriate processing parameters and it allows to perform a local adaptation of the threshold. With the combined architecture, the CFAR loss can be controlled by changing both the type of CFAR processor and the window length.

Acknowledgment

Thanks to Prof. J. Ender and T. Mathy from Fraunhofer FHR, Wachtberg, Germany, for making available the radar system and for technical support during the experiments.

References

[1] R. G. Baraniuk and T. P. H. Steeghs. Compressive radar imaging. In *Proc. IEEE Radar Conf.*, 2007.
[2] M. A. Herman and T. Strohmer. High-resolution radar via compressed sensing. *IEEE Trans. Signal Process.*, 57(6):2275–2284, 2009.
[3] J. H. G. Ender. On compressive sensing applied to radar. *Elsevier J. Signal Process.*, 90(5):1402–1414, May 2010.
[4] L. C. Potter, E. Ertin, J. T. Parker, and M. Cetin. Sparsity and compressed sensing in radar imaging. *Proc. IEEE*, 98(6):1006–1020, Jun. 2010.
[5] L. Anitori, A. Maleki, M. Otten, R. G. Baraniuk, and P. Hoogeboom. Design and analysis of compressive sensing radar detectors. Feb. 2013.
[6] P. P. Gandhi and S.A. Kassam. Analysis of CFAR processors in homogeneous background. *IEEE Trans. Aerosp. Electron. Syst.*, 24(4):427–445, Jul. 1988.
[7] H. Rohling. Radar CFAR thresholding in clutter and multiple target situations. *IEEE Trans. Aerosp. Electron. Syst.*, 19(4):608–621, Jul. 1983.
[8] S. Blake. Os-cfar theory for multiple targets and nonuniform clutter. *IEEE Trans. Aerosp. Electron. Syst.*, 24(6):785–790, Nov. 1988.
[9] L. Anitori, A. Maleki, W. van Rossum, R. Baraniuk, and M. Otten. Compressive CFAR radar detection. In *Proc. IEEE Radar Conf.*, 2012.
[10] A. Maleki, L. Anitori, Y. Zai, and R. G. Baraniuk. Asymptotic analysis of complex LASSO via complex approximate message passing (CAMP). submitted to *IEEE Trans. Inf. Theory*, 2011.
[11] Robert Tibshirani. Regression shrinkage and selection via the LASSO. *J. Roy. Stat. Soc., Series B*, 58(1):pp. 267–288, 1996.
[12] S.S. Chen, D.L. Donoho, and M.A. Saunders. Atomic decomposition by basis pursuit. *SIAM J. on Sci. Computing*, 20:33–61, 1998.
[13] E. Candès and T. Tao. Near optimal signal recovery from random projections: Universal encoding strategies? *IEEE Trans. Inf. Theory*, 52(12):5406–5425, Dec. 2006.
[14] D. L. Donoho. Compressed sensing. *IEEE Trans. Inf. Theory*, 52(4):1289 – 1306, Apr. 2006.
[15] D. L. Donoho, A. Maleki, and A. Montanari. Message passing algorithms for compressed sensing. *Proc. Natl. Acad. Sci.*, 106(45):18914–18919, 2009.
[16] D. L. Donoho, A. Maleki, and A. Montanari. The noise-sensitivity phase transition in compressed sensing. *IEEE Trans. Inf. Theory*, 57(10):6920–6941, Oct. 2011.
[17] D. P. Meyer and H. A. Mayer. *Radar target detection: handbook of theory and practice*. Academic Press Inc., New York, NY, 1973.

Sampling Techniques for Improved Algorithmic Efficiency in Electromagnetic Sensing

Kyle R. Krueger, James H. McClellan, Waymond R. Scott Jr.
School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332–0250

***Abstract*—Ground-penetrating radar (GPR) and electromagnetic induction (EMI) sensors are used to image and detect subterranean objects; for example, in landmine detection. Compressive sampling at the sensors is important for reducing the complexity of the acquisition process. However, there is a second form of sampling done in the imaging-detection algorithms where a parametric forward model of the EM wavefield is used to invert the measurements. This parametric model includes all the features that need to be extracted from the object; for subterranean targets this includes but is not limited to type, 3D location, and 3D orientation. As parameters are added to the model, the dimensionality increases. Current sparse recovery algorithms employ a dictionary created by sampling the entire parameter space of the model. If uniform sampling is done over the high-dimensional parameter space, the size of the dictionary and the complexity of the inversion algorithms grow rapidly, exceeding the capability of real-time processors. This paper shows that strategic sampling practices can be exploited in both the parameter space, and the acquisition process to dramatically improve the efficiency and scalability of the these EM sensor systems.**

I. Introduction

Parameter estimation of unknown objects through the use of wavefield sensors is a well researched area. An increasingly popular solution to these types of problems comes from the advancements in compressive sensing (CS) and sparse recovery [1]. These inversion algorithms rely on the fact that a highly accurate forward model of the data could be created to describe the dependence of the physical sensor data (i.e., the measurements) on the interesting parameters of the objects being imaged. This approach highlights an issue with CS. The inherent need for a random sensing matrix does not always lend itself easily to practical data acquisition from sensors. On the other hand, creating a comprehensive target model, oftentimes called a dictionary, and referred to in the CS world as a sparsifying transform, can quickly become too large and too computationally intensive for real-time computers. The data collection and imaging flow is shown in Fig. 1.

The key sampling issue is creating a dictionary of manageable size, even when it is desirable to add more parameters to the model. A d-parameter, m-measurment model leads to a dictionary of size $\mathcal{O}(N^{d+m})$, assuming equal sampling (N) of each variable. This paper will show, through the use of strategic parameter-space sampling, that the dimensionality of the dictionary can be reduced in two different acquisition environments. Thus the computational complexity of these parameter detection problems can be drastically reduced for these sensor systems.

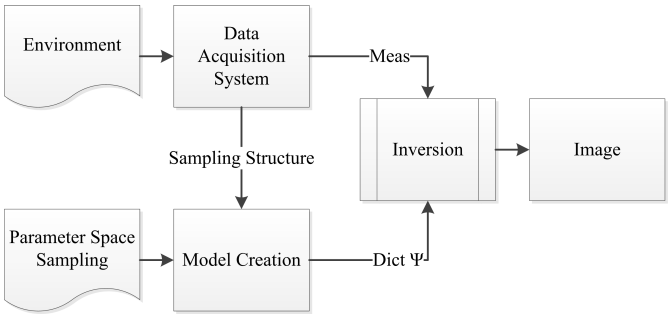


Fig. 1. Imaging algorithm flow.

There are two different acquisition systems that are discussed in this paper. The first system is three-dimensional (3D) imaging of subterranean targets using a ground-penetrating radar (GPR). The typical acquisition sampling pattern in GPR allows for reduced data acquisition time through the use of a random sensing pattern. The simplification of sampling the parametric forward model comes from exploiting the translationally-invariant nature of the physical model. The second system uses electromagnetic induction (EMI) sensors to detect and classify underground metallic targets. In this case, the strategic sampling of the parameter space comes from adopting an efficient tensor model to describe the orientation and magnetic polarizability of the target. This can be extracted using a rank-minimization detection algorithm. This model represents orientation space continuously with very few samples, instead of requiring the entire 3D angle space to be enumerated. This dramatically reduces computational complexity and also increases accuracy by eliminating the off-grid parameter sampling problem with regards to orientation.

II. Ground-penetrating Radar

The GPR system considered here is a stepped-frequency system that has been previously described in detail [2]. The forward model is a point-target model, and the detection algorithm is based on sparse recovery (CS). The remainder of this section explains how acquisition sampling and model-parameter sampling together lead to a translational-invariance property that can achieve the computational complexity reductions in the detection algorithm that were shown in [3].

A. Model

The point-target model used for the detection algorithm is

$$\psi(f, \boldsymbol{l}_s, \boldsymbol{l}_t) = \frac{g(f)e^{-2j\pi f\tau(\boldsymbol{l}_s, \boldsymbol{l}_t)}}{S(\boldsymbol{l}_s, \boldsymbol{l}_t)}, \tag{1}$$

which is a function of the stepped frequencies, f; the sensor positions, $\boldsymbol{l}_s = (l_s(x), l_s(y), 0)$; the target locations, $\boldsymbol{l}_t = (l_t(x), l_t(y), l_t(z))$; and the spreading parameter, $S(\boldsymbol{l}_s, \boldsymbol{l}_t)$. This forward model, $\psi(f, \boldsymbol{l}_s, \boldsymbol{l}_t)$, is used as the dictionary, or sparsifying transform. If ψ is discretized and enumerated for all possible frequencies, sensor locations and target locations, the resulting model $\boldsymbol{\Psi}$ is 6D. The storage requirements are $\mathcal{O}(N^6)$ for equal discretization of all parameters, [3]. Our objective is to use properties inherent in the model and the acquisition system to reduce this storage and computational burden.

B. Special Properties

A special property that can be used for increased efficiency is the fact that the model above can be translationally invariant. A translationally-invariant model can be applied using the Fast Fourier Transform (FFT), which eliminates the storage requirements for each dimension having this property. The translational-invariance property is true when the parameter space and the measurement space are evenly sampled in the same direction. In other words, when the target and the sensor are moved an equal distance in a horizontal dimension, x or y, the radar response will remain the same. Also, to use the FFT to garner the complexity reduction, the stepped frequencies at each sensor position, $\boldsymbol{l}_s$, must be the same. This runs counter to the usual random sampling approach in CS, but it is a very important constraint when trying to exploit this special property even in a CS environment.

C. Compressive Sensing Detection Algorithm

Now that the special properties in the model are identified, the detection algorithm itself can reduce the time needed for computation and data acquisition, if the sampling is done properly. The idea behind CS is that if the model parameters can be sparse, then projecting the model onto a known random subspace with much lower dimensionality than the original can still enable an accurate inversion [1]. Often the projections are done with a random sampling matrix $\boldsymbol{\Phi}$ applied to the model, $\boldsymbol{\Psi}$. For good results, $\boldsymbol{\Phi}$ should be independent and identically distributed (IID) random. There are a few techniques that will reduce the computational complexity of this general matrix multiplication, but they do not allow for any reduction in acquisition time for this particular GPR acquisition system [4], [5].

To get a mix of computational complexity reduction and data acquisition time reduction while staying within the CS framework, a strategic $\boldsymbol{\Phi}$ should be designed. An in-depth analysis of the trade-offs in designing $\boldsymbol{\Phi}$ for this GPR acquisition system were studied by Gurbuz et al. [6]. The basic trade-off is that the more structured $\boldsymbol{\Phi}$ becomes, the higher the coherence of the dictionary, and thus the higher the number of samples required for reconstruction, but this can also allow for reduction in data acquisition time. For this problem, some additional structure is needed in $\boldsymbol{\Phi}$ to exploit the translational invariance. To use the FFT across x and y dimensions, the full x and y must be sampled for a given f. This means that $\boldsymbol{\Phi}$ can be built to randomly select a small number of f to get both a reduction in complexity for the CS algorithm, but also reduce the data acquisition time, and amount of data that needs to be collected.

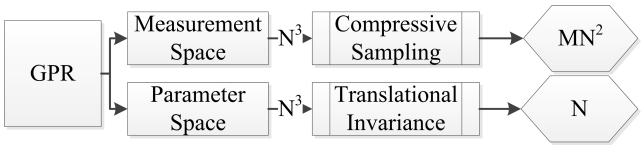


Fig. 2. Flow chart of GPR acquisition system, models, and complexity reducing properties

D. Complexity Reduction

The complexity reduction for exploiting these model properties in this particular system are quite significant. In terms of storage space, the original fully discretized parameter-space model is 6D having a storage requirement of $\mathcal{O}(N^6)$. By using the FFT and CS, the storage requirement for the dictionary was reduced to $\mathcal{O}(MN^3)$, where $M < N$ is the number of random frequency measurements. In practical application of this method to laboratory measurements, the frequency requirements are reduced from 401 to 10 [6]. For an actual system, the data acquisition would take a fortieth of the time, as well as saving a factor of 40 in the amount of storage needed. The flow of the GPR acquisition system, the special properties, and their effect on complexity are summarized in Fig. 2.

There is also a rather significant reduction in algorithm time in using the translationally-invariant model over using a direct approach. Direct matrix multiplication for the 6D problem has a complexity of $\mathcal{O}(N^6)$, but the translationally-invariant model can be applied in $\mathcal{O}(N^4 \log_2(N))$ because the FFT can be applied along two of the parameter dimensions. A semilog plot of computation time versus problem size (N) for both of these models in theory, and the FFT-based method in practice, can be seen in Fig. 3. For $N = 70$, the FFT-based method is more than 400 times faster. In fact, the direct method was not measured since it cannot be applied for $N = 70$ because the storage requirements are about 950 Gbytes, while the FFT-based method requires around 200 Mbytes.

III. Electromagnetic Induction Sensor

A different acquisition system used for collecting target data is a multi-frequency (wideband) EMI sensor system. Multiple sensors are scanned in a down-track pattern, acquiring a sampled frequency response at uniformly spaced locations along the scan path. The forward model is a frequency domain model with many more parameters than the point target model [7]–[10]. The sparse recovery algorithm used is formulated as a combined least-squares and low-rank approximation problem.

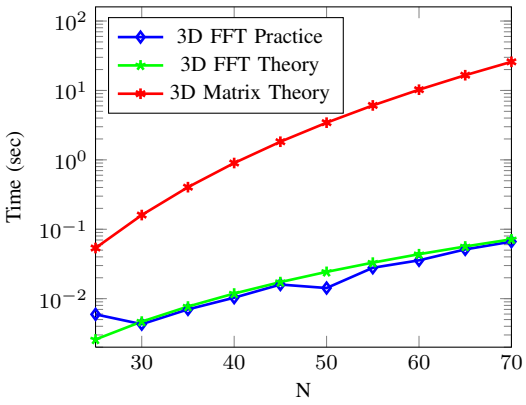


Fig. 3. Run time for different model applications

A. Model

The basic model used for this system is one that is written in the frequency, ω, domain, for a single target type, μ,

$$r(\omega, \boldsymbol{l}_s, \boldsymbol{l}_t, \boldsymbol{o}_t, \mu) = \boldsymbol{g}^T(\boldsymbol{l})\boldsymbol{R}(\boldsymbol{o}_t)\boldsymbol{A}(\omega, \mu)\boldsymbol{R}^T(\boldsymbol{o}_t)\boldsymbol{f}(\boldsymbol{l}). \quad (2)$$

$\boldsymbol{l}_s = (l_s(x), 0, 0)$ is sensor position, $\boldsymbol{l}_t = (l_t(x), l_t(y), l_t(z))$ is 3D target location, $\boldsymbol{o}_t = (o_t(\alpha), o_t(\beta), o_t(\gamma))$ is target orientation, $\boldsymbol{A}(\omega, \mu)$ is a 3×3 matrix that defines the magnetic polarizability of the target, and $\boldsymbol{l} = \boldsymbol{l}_t - \boldsymbol{l}_s$ is the relative location vector for the target if the sensor was the origin. $\boldsymbol{g}(\boldsymbol{l})$ and $\boldsymbol{f}(\boldsymbol{l})$ are vectors that contain the spatial components of the magnetic field on the receive coil and the transmit coil respectively based on the relative location vector $\boldsymbol{l}$. $\boldsymbol{R}(\boldsymbol{o}_t)$ is a simple rotation matrix that rotates by angle $\boldsymbol{o}_t$. When all the measurements, $\{\omega, l_s(x)\}$, and parameters, $\{l_t(x), l_t(y), l_t(z), o_t(\alpha), o_t(\beta), o_t(\gamma), \mu\}$, are enumerated; the result is a data hyper-cube of 9D. The storage requirement is enormous when the parameter space is sampled finely enough.

An important change is to model the response as an expansion of magnetic dipoles, each with a frequency relaxation [9]. The coefficients of the expansion can be computed from the experimental data [11]. The coefficient for the term in the expansion with a relaxation frequency, ζ, is

$$r^{\zeta}(\boldsymbol{l_s}, \boldsymbol{l_t}, \boldsymbol{o}_t, \boldsymbol{\Lambda}) = \boldsymbol{g}^T(\boldsymbol{l})\boldsymbol{R}(\boldsymbol{o}_t)\boldsymbol{\Lambda}\boldsymbol{R}^T(\boldsymbol{o}_t)\boldsymbol{f}(\boldsymbol{l}). \quad (3)$$

Each individual ζ can be imaged separately using the same model (3), regardless of type. Typically, the number of relaxation frequencies, N_ζ, is between one and six. There are two significant benefits of the expansion. First, a specific frequency response for each target type is no longer needed. Second, $\boldsymbol{A}(\omega, \mu)$ changes to $\boldsymbol{\Lambda}$, which is a 3×3 diagonal, positive semidefinite, real matrix that does not depend on ω or μ. These benefits greatly reduce the storage requirements. However, since each ζ must be imaged independently, the number of imaging steps increases from one to N_ζ, even though the model itself does not depend on ζ.

B. Special Properties

This model (3) has two special properties. First, the model is separable into a product of functions. There are separate functions for location, orientation, and magnetic polarizibility that contribute to the product. This means that individual parameters can be isolated from one another. The second property comes from thinking of $\boldsymbol{R}(\boldsymbol{o}_t)\boldsymbol{\Lambda}\boldsymbol{R}^T(\boldsymbol{o}_t)$ as a "generalized amplitude" of the target. Usually, the response of a point target is a scalar that represents the strength of the target, but the matrix $\boldsymbol{R}(\boldsymbol{o}_t)\boldsymbol{\Lambda}\boldsymbol{R}^T(\boldsymbol{o}_t)$ encodes additional information about how the target strength depends on symmetry and orientation.

To build a dictionary, there needs to be an enumeration for every possible sample in the interesting parameter space, but it is undesirable to enumerate all possible entries of the matrix Λ along with all possible orientation angles $\boldsymbol{\alpha}$. To avoid storing a large number of samples, a change can be made to the fundamentals of sampling this model. Instead of thinking about a point target response as having a scalar amplitude, it can be thought of as having a tensor amplitude by rewriting the model in (3) as

$$r^{\zeta}(\boldsymbol{l}_s, \boldsymbol{l}_t, \boldsymbol{o}_t, \boldsymbol{\Lambda}) = \boldsymbol{g}^T(\boldsymbol{l})\boldsymbol{T}(\boldsymbol{o}_t, \boldsymbol{\Lambda})\boldsymbol{f}(\boldsymbol{l}), \quad (4)$$

where $\boldsymbol{T}$ is a symmetric, positive semidefinite matrix that is only 3×3. This will be referred to as a "tensor amplitude." It has a great advantage over just the scalar amplitude. It contains the continuous orientation and the magnetic polarizability of the target in its eigenvectors and eigenvalues respectively. This gives a more accurate model, because it does not require sampling of the orientation parameter, so there is no modeling error associated with having targets whose orientations do not lie exactly on the sampled orientation space. Also, this reformulation reduces a 3D grid of angle samples to just six independent values in $\boldsymbol{T}(\boldsymbol{o}_t, \boldsymbol{\Lambda})$ which provides a large computational savings. Once $\boldsymbol{T}(\boldsymbol{o}_t, \boldsymbol{\Lambda})$ is found, an eigen-decomposition will yield $\boldsymbol{o}_t$ and $\boldsymbol{\Lambda}$.

C. Detection Algorithm

The detection algorithm for the EMI acquisition system is a combination of least squares and a semidefinite programming (SDP) technique used to get a low-rank approximation. The solution to this problem is sparse in 3D, in just the same way the GPR system is sparse. In fact, in most cases it should be even more sparse, because the model (4) is much more sophisticated and is looking for magnetic dipoles, and not just a sum of point reflections.

The full problem can be solved using a block-tensor representation to simultaneously find the target location and the tensor amplitude through a convex relaxation to the rank-minimization algorithm [12],

$$\begin{aligned} \min \quad & \mathbf{tr}(\hat{\boldsymbol{T}}) \\ \text{s.t.} \quad & \hat{\boldsymbol{T}} \succeq 0, \| \mathbf{b} - \boldsymbol{\Psi}\mathbf{s} \| < \epsilon. \end{aligned} \quad (5)$$

$\hat{\boldsymbol{T}}$ is a block-diagonal tensor made up of 3×3 tensors $\boldsymbol{T}_t$, one for each possible target location. $\mathbf{b}$ is the collected measurement vector, $\boldsymbol{\Psi}$ is the dictionary enumerated from (4), and $\mathbf{s}$ is a sparse parameter vector that makes up the nonzero values in $\boldsymbol{T}$. This exploits the fact that the block-tensor structure will be extremely low rank. This is the case

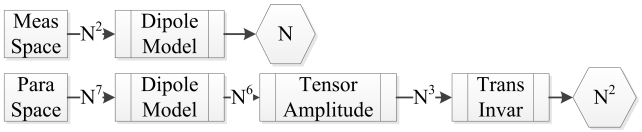


Fig. 4. Storage size for different versions of the sampled EMI acquisition and processing system. Complexity is reduced by exploiting special sampling properties of parametric models for EMI.

because the rank of the large block tensor structure, $\hat{\boldsymbol{T}}$, is just the sum of the rank of all tensor amplitudes, $\boldsymbol{T}_t$, of each present target. This work is in its initial stage of development because the full problem is computationally intense.

A shortcut can be used to break the algorithm into two steps to address the complexity of the full problem. Using an orthogonal matching pursuit type technique, a least-squares problem is solved to approximate the location of the target first [13]. Then (5) can be recast as a very small SDP,

$$\begin{aligned} \min \quad & \mathbf{tr}(\boldsymbol{T}_t) \\ \text{s.t.} \quad & \boldsymbol{T}_t \succeq 0, \| \mathbf{b} - \boldsymbol{\Psi}_t \mathbf{s}_t \| < \epsilon, \end{aligned} \tag{6}$$

to get the tensor amplitude of the target at location, $\boldsymbol{l}_t$. The target response is then subtracted from the measurement and the process is repeated until the stopping criteria, a small enough residual, is met.

The EMI system also has the translationally-invariant property in the scanning dimension, x, just like the GPR acquisition system. However, the detection algorithm for this model setup is more complicated than direct matrix multiplication, so it will be more difficult to take advantage simultaneously of both the tensor representation property and the translational invariance in the large problem. Such a combined algorithm would be desirable, but it has not been implemented yet for a practical application.

D. Complexity Reduction

The flow chart of measurement and parameter space simplifications of the EMI system in Fig. 4 summarizes the special properties exploited, and their resulting complexity reductions. Using the dipole model is a very important computation saving step, eliminating N^2 storage, going from a data hyper-cube of N^9 to N^7. Using the tensor amplitude representation, which changes the fundamentals of how the forward model is sampled, both increases the accuracy of the solution and garners an N^3 savings to drop the overall storage requirements to N^4. This is the result of tensor sampling (which needs six values) eliminating the need to finely sample the entire 3D orientation parameter. The EMI system also has the same translational invariance as the GPR system, and if it were exploited, there is another dimension of savings. Ultimately, the result of taking advantage of these special properties could obtain a savings of N^6.

IV. Conclusion

This paper emphasizes the importance of the model representation. How the measurements are acquired, how the parameter space of the forward model is sampled, and how these two sampling operations can be adjusted to take advantage of special properties can all contribute to reducing the computational complexity. The tensor representation is also a different way to think about modeling data, and has been exploited in other applications such as seismic [14]. Using discrete values to provide continuous responses can allow for more accurate models while still harnessing the power of computers. A variation of this idea has been done in modeling continuous signals with Taylor series and cosine representations which allow for discrete values to be acquired [15]. The advantages of these sampling structures have been shown to drastically reduce computational complexity, increase accuracy, and reduce data acquisition times when combined with dictionary based detection algorithms.

Acknowledgment

This work is supported in part by the US Army REDCOM CERDEC Night Vision and Electronic Sensors Directorate, Science and Technology Division, Countermine Branch and in part by the U. S. Army Research Office under grant number W911NF-11-1-0153.

References

[1] D. L. Donoho, "Compressed sensing," *IEEE Trans. Information Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[2] T. Counts, A. C. Gurbuz, W. R. Scott, Jr., J. H. McClellan, and K. Kim, "Multistatic Ground-Penetrating Radar Experiments," *IEEE Trans. Geoscience and Remote Sensing*, pp. 2544–2553, 2007.

[3] K. Krueger, J. H. McClellan, and W. R. Scott, Jr., "3-D imaging for ground penetrating radar using compressive sensing with block-toeplitz structures," in *Proc. IEEE SAM*, 2012, pp. 229 –232.

[4] W. U. Bajwa, J. D. Haupt, G. M. Raz, S. J. Wright, and R. D. Nowak, "Toeplitz-Structured Compressed Sensing Matrices," *Proc. IEEE Statistical Signal Processing Workshop*, pp. 294 –298, Aug. 2007.

[5] J. Romberg, "Compressive Sensing by Random Convolution," *SIAM Journal on Imaging Science*, Dec. 2009.

[6] A. C. Gurbuz, J. H. McClellan, and W. R. Scott, Jr., "Compressive sensing for subsurface imaging using ground penetrating radar," *Signal Processing*, vol. 89, no. 10, pp. 1959 – 1972, 2009.

[7] A. C. Gurbuz, W. R. Scott, Jr., and J. H. McClellan, "Location estimation using a broadband electromagnetic induction array," *Proc. SPIE*, 2009.

[8] M. Özdemir, E. L. Miller, and S. Norton, "Localization and characterization of buried objects from multifrequency array inductive data," *Proc. SPIE*, Apr. 1999.

[9] W. R. Scott, Jr. and G. D. Larson, "Measured dipole expansion of discrete relaxations to represent the electromagnetic induction response of buried metal targets," *Proc. SPIE*, 2010.

[10] G. D. Larson and W. R. Scott, Jr., "Automated, non-metallic measurement facility for testing and development of electromagnetic induction sensors for landmine detection," *Proc. SPIE*, 2009.

[11] M. Wei, W. R. Scott, Jr., and J. H. McClellan, "Robust Estimation of the Discrete Spectrum of Relaxations for Electromagnetic Induction Responses," *IEEE Trans. Geoscience and Remote Sensing*, pp. 1–11, 2009.

[12] C. Beck and R. D'Andrea, "Computational study and comparisons of LFT reducibility methods," in *Proc. American Control Conference*, vol. 2, Jun. 1998, pp. 1013 –1017 vol.2.

[13] J. Tropp and A. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Information Theory*, vol. 53, no. 12, pp. 4655–4666, 2007.

[14] C. H. Chapman and W. S. Leaney, "A new moment-tensor decomposition for seismic events in anisotropic media," *Geophysical Journal International*, vol. 188, pp. 343–370, 2012.

[15] C. Ekanadham, D. Tranchina, and E. P. Simoncelli, "Recovery of Sparse Translation-Invariant Signals With Continuous Basis Pursuit," *IEEE Trans. Signal Processing*, pp. 4735–4744, Oct. 2011.

Coding and sampling for compressive tomography

David J. Brady
Department of Electrical and
Computer Engineering
Duke University
Durham, North Carolina
Email: dbrady@duke.edu

***Abstract*—This paper discusses sampling system design for estimation of multidimensional objects from lower dimensional measurements. We consider examples in geometric, diffractive, coherence, spectral and temporal tomography. Compressive tomography reduces or eliminates conventional tradeoffs between temporal and spatial resolution.**

I. Introduction

Compressive measurement is generally defined as the estimation of N signal values from M measurements for $M < N$. While this definition has been highly useful and successful in many sensing and imaging applications, an alternative definition is of equal utlity in tomographic imaging. Tomography most commonly consists of imaging 3D objects from measurements distributed over 1D or 2D sensor arrays. Typical tomographic systems may be described by integral equations of the form

$$g(y) = \int f(x)h(x,y)dx \tag{1}$$

where $x \in \mathbb{R}^N$ and $y \in \mathbb{R}^M$. One may define "compressive tomography" as estimation of $f(x)$ from $g(y)$ in the case that $M < N$.

Tomographic systems typically use sensor arrays embedded on the boundary or surface of a volume under observation. In fan beam tomography, for example, a linear detector array measures attenuation of rays through a 2D object space. In cone beam tomography, a planar detector arrray measures rays projected through a 3D volume. Conventional tomography overcomes the dimensional mismatch between the object and measurement spaces by varying illumination and sensor geometry as a function of time, thereby increasing the dimension of the measurement space by 1. Thus in conventional systems $M = N - 1$ for measurements taken at a fixed time, but $M = N$ when time is taken into account.

The most unfortunate aspect of the conventional approach is that it requires that the object remain static as measurements are collected over time. Over the past several years, my group has applied compressive sampling theory to implement snapshot compressive tomography. For example, we have shown that 3D hyperspectral [1], diffraction [2] and x-ray scatter [3] images may be reconstructed from 2D data. We have also analyzed compressive sampling for reconstruction of 3D objects with conventional optics [4]. Most recently, we have shown that 3D video data cubes may be constructed from 2D frames [5], thus using compressive tomography to reconstruct time itself.

While the object distribution $f(x)$ is by definition distributed over continous space, measurements ultimately consist of discrete digital data. There is no fundamental requirement that discrete measurements be indexed by a continous variable. Standard compressive sampling models assume independent kernels for each measurement. Unfortunately, completely independent kernels are difficult or impossible to implement on measurements embedded in continuous physical space. Due in part to this challenge, Candes' early analysis of compressive tomography focused on discrete subsampling of the temporal portion of Radon space with continuous sampling in each snapshot [6].

My group's initial theoretical studies of compressive tomography focused on the use of multidimensional reference structures to enable random or decorrelated measurement over a continuous space [7]. However, most subsequent efforts to implement practical compressive tomography may be described in the context of three basic coding strategies

1) **Measurement space coding.** The standard model for increasing measurement dimensionality with time involves varying the the measurement kernel to obtain

$$g(y,t) = \int f(x)h(x,y,t)dx \tag{2}$$

Measurement space coding multiplexes diverse kernels in a snapshot to obtain

$$\tilde{g}(y) = \int g(y,t)C(y,t)dt = \int C(y,t)f(x)h(x,y,t)dxdt \tag{3}$$

where $C(y,t)$ is a code applied to each time slice in measurement space. $C(y,t)$ is designed to allow "code division multiple access" (CDMA) such that $g(y,t)$ can be isolated from $\tilde{g}(y)$.

2) **Object space coding** modulates the object density prior to measurement to obtain the forward model

$$g(y) = \int f(x)C(x)h(x,y)dx \tag{4}$$

Again, $C(x)$ enables the use of CDMA to increase the effective dimensionality of the measurements.

3) **Transform subsampling** expands the subsampling strategy of [6] to optimize which portions of the transform space measured.

CDMA is, of course, most commonly understood in the context of multiuser communications. CDMA considers the case that a set of relatively low frequency signals $f_i(t)$ must communicate over the same channel. Multiplication of each signal with an independent high frequency code $C_i(t)$ enables one to isolate each signal even when the overall transmitted data is $g(t) = \sum_i f_i(t)C_i(t)$. This is achieved by assuming that the codes are orthogonal over short time windows such that

$$\int_{t-T}^{t} g(t')C_j(t')dt' = \sum_i f_i(t) \int C_i(t')C_j(t')dt' = f_j(t) \tag{5}$$

In effect, coding turns the 1D measurement over time into a 2D measurement over time and transmitter index. In the same way, coding in tomography systems effectively increases the dimensionality of the measurements. The question "What is the maximum bandwidth of $f(t)$ relative to the bandwidth of $C(t)$ such that this dimensionality increase can be achieved is a central issue in compressive sampling theory.

The goal of the remainder of this paper is to relate these abstract coding strategies to practical tomographic imagers. Tomographic system design is inherently an integrated sensing and processing challenge by which physical and geometric constraints must be matched to mathematical conditioning and algorithms. The next section reviews the basic physical structure of tomographic imagers and dicusses how coding strategies 1-3 are implemented in these systems.

II. Field Models and Coding

While Eqn. (1) might describe many different measurement systems, the underlying concept that measurements and objects are distributed over continous spaces linked by a continous kernel uniquely describes remote sensing systems. The transformation from object to measurement is mediated by radiation fields propagating between the two spaces. While "tomographic imaging" in its most general sense refers to systems as diverse as MRI and electron microscopy, most analyses of computed tomographic imaging focus specifically on imaging using radiating fields [8].

Radiation fields are commonly described by (1) geometric models, under which the fields propagate as nondiffracting rays, (2) diffraction models, under which the fields propagate as waves and (3) coherence models, which generalize wave models to account for quantum noise and measurement characteristics [9]. Each field model is most applicable in specific contexts, corresponds to specific measureable features and is amenable to specific coding strategies.

For geometric tomography, attenuation or scatter of rays is the basic measureable quantity. Specifically, one measures

$$g(y, \theta) = \int f(y + \alpha\theta)d\alpha \tag{6}$$

where $y \in \mathbb{S}^{N-1}$ is a point on a boundary enclosing the object and $\theta \in \mathbb{S}^{N-1}$ is the direction vector for a ray passsing through y. For $N > 2$, the dimensionality of the potential ray measurement space, $M = 2N - 2$, is greater than N and inversion is over constrained. The challenge of geometric tomography is that it is not possible to simultaneously discriminate all rays passing through y. Typical detectors have no mechanism for discriminating rays and simple intetrate the total irradiance over all rays passing through the detector point. Conventional tomographic imagers overcome this problem by ensuring that only 1 ray passes through each measurement point in each measurement time. This is most often achieved by illuminating with a collimated pencil, fan or cone beam source. Under this scenario, θ is a single valued function of y and the measurement is

$$g(y, t) = \int f\left(y + \alpha\theta(y, t)\right) d\alpha \tag{7}$$

for $y \in \mathbb{S}^{N-1}$ and $t \in \mathbb{R}$. A dimensional match between measurements and the object is achieved by changing $\theta(y, t)$ as a function of time.

Each of coding strategies 1-3 may be implemented in geometric tomography. Measurement space coding is applied in x-ray scatter imaging by placing a coded aperture between the scattering target and the measurement plane. Where conventional scatter imaging scans a colimator as function of time, coding allows distinct range, cross range and momentum slices to be multiplexed and reconstructed from a single time step [3], [10]. Measurement space coding may also be applied using a coded aperture with multiple illumination sources. Illumination angle-based code shifts allow disambiguation of the sources and simultaneous aquisition of multiple source data [11]. Multisource coding in combination with scatter imaging may also be understood as object space coding. Rather than using coded aperture shadows to disambiguate scatter sources, one may use structured illumination to code scatter position of distributed targets. Finally, as noted above, subsampling of multiple source data is an example of transform subsampling. While in [6], this subsampling takes the form of discontinous selection of continuous subspaces of the Radon transformation, more effective compression is obtained by combining mulitsource illumination with coded apertures, reference structures or collimation filters to more randomly sample Radon space. As suggested by this survey of practical strategies, detailed analysis of the coding strategy depends both on physical feasibility, object priors and mathematical structure.

Despite all the complexity of wave mechanics, the most immediate difference between geometric tomography and diffraction tomography is that the diffraction sample surface integrals rather than via line integrals. More substantive differences arise from object field-interaction models, typical object priors and the use of time to measure space. Diffraction tomography, including radar, millimeter wave and terahertz imaging, ultrasound and optical holography, most often considers scattered radiation rather than attenuation or primary sources. Under the Born approximation, the scattered field for single plane wave illumination samples a spherical shell in the Fourier space of the object density [12]. A single measurement corresponds to a point in the Fourier space in this case.

From a practical perspective, the use of phase delay or time of flight to measure range is the most unique and powerful aspect of diffraction tomography. This technique enables optical coherence tomography (OCT), which measures spatial range with resolution proportional to spectral bandwidth rather than aperture size. Compressive OCT has been considered in several studies using transform subsampling [13]. Time of flight from a monostatic transceiver integrates the object density on a sphere surrounding the transceiver with a range proportional to the observation time. Measurement of a family of spheres obtained by translating the transceiver obtains a Randon-like transformation of the volume. Bistatic or multistatic systems sample integrals over hyperbolic surfaces between emitter and receiver positions.

As with geometric tomography, strategies 1-3 may be applied in compressive diffraction tomography. While I am not aware of any examples of measurement space coding with coherent waves, the use of a metamaterial transceiver to create structured illumination [14] is an example of compressive tomography using object space coding. 3D object estimation from Fourier space manifolds in [15], [16] is an example of tranform subsampling, although disjoint or randomized subsampling as described for 2D images in [17] may be considered more sample efficient. Accounting for the unique physical priors arising from diffuse and specular reflection of coherent radiation is the most challenging aspect of diffraction tomography, however.

The scattered field on the surface of a diffuse reflector is a complex Gaussian random variable. Since the mean of the field is 0, estimation of the mean is an ineffective imaging strategy. The magnitude of the field is exponentially distributed, estimation of the magnitude lead to speckled images. Given that the field is random and uncorrelated in each pixel, the field over a 2D image is not generally compressible. Compressive tomography is therefore best implemented by building a forward model on the nonnegative object scattering density, which corresponds to the variance of the Gaussian random process [18]. Whether the scatter is diffuse or specular, however, one notes that diffraction tomography tends to be most useful in imaging interfaces and surfaces rather than continuous volumes. While the reason for this may be simply that volume imaging is too noisy and random to allow imaging to occur, design of compressive diffraction tomography systems would most effectively build on the assumption that the object consists exclusively of surfaces. This prior should enable highly compressive and super-resolved estimation of even diffuse scatters and is thus a worthy area for ongoing research.

Optical coherence functions, most typically consisting of the cross spectral density, describe fields radiated by random natural sources. While one may apply interferometric methods to directly sample the cross spectral density for transform subsampling based compressive tomography [19], such methods are ill-conditioned for complex sources. Focal imaging is the only mathematically well conditioned strategy for measurement of random sources but is incapable of mapping volume distributions onto measurement planes [9]. Object space modulation [4] and focal stacking (sweeping focal parameters during exposure) may be used to overcome this limitation. Compressive tomography of random volume sources is much more challenging that geometric or diffraction tomography, however, and remains an active research challenge.

III. Conclusion

The reader may be surprised to complete an entire article on tomographic imaging without encountering a single image. To my knowledge, however, this is the first print article to explicitly consider compressive tomography as defined in Eqn. (1). As such I hope that the reader will find the intellectual exercise of mapping this definition onto essentially the complete gamut of remote sensing systems sufficiently fascinating as to agree that a few simple images of traditional phantoms would only be a distraction. The conventional concept of an image as a 2D object that can be captured on a focal plane and displayed in an article is an artifact of analog image processing. In the modern world of computational and compressive imaging, all images are multidimensional and all imaging systems are tomographic.

References

[1] A. A. Wagadarikar, N. P. Pitsianis, X. Sun, D. J. Brady *et al.*, "Video rate spectral imaging using a coded aperture snapshot spectral imager," *Opt. Express*, vol. 17, no. 8, pp. 6368–6388, 2009.

[2] J. Hahn, S. Lim, K. Choi, R. Horisaki, and D. J. Brady, "Video-rate compressive holographic microscopic tomography," *Optics Express*, vol. 19, no. 8, pp. 7289–7298, 2011.

[3] M. T. K. MacCabe, A. Holmgren and D. J. Brady, "Snapshot 2d tomography via coded aperture x-ray scatter imaging," to appear in Applied Optics.

[4] D. J. Brady and D. L. Marks, "Coding for compressive focal tomography," *Applied optics*, vol. 50, no. 22, pp. 4436–4449, 2011.

[5] P. Llull, X. Liao, X. Yuan, J. Yang, D. S. Kittle, L. Carin, G. Sapiro, and D. J. Brady, "Coded aperture compressive temporal imaging," *CoRR*, vol. abs/1302.2575, 2013.

[6] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *Information Theory, IEEE Transactions on*, vol. 52, no. 2, pp. 489–509, 2006.

[7] D. J. Brady, N. Pitsianis, and X. Sun, "Reference structure tomography," *J. Opt. Soc. Am. A*, vol. 21, no. 7, pp. 1140–1147, 2004.

[8] M. Slaney and A. Kak, "Principles of computerized tomographic imaging," *SIAM, Philadelphia*, 1988.

[9] D. Brady, *Optical imaging and spectroscopy*. Wiley-OSA, 2009.

[10] K. MacCabe, K. Krishnamurthy, A. Chawla, D. Marks, E. Samei, and D. Brady, "Pencil beam coded aperture x-ray scatter imaging," *Optics Express*, vol. 20, no. 15, pp. 16 310–16 320, 2012.

[11] K. Choi and D. J. Brady, "Coded aperture computed tomography," in *Adaptive Coded Aperture Imaging, Non-Imaging, and Unconventional Imaging Sensor Systems*, vol. 7468. SPIE, 2009, p. 74680B.

[12] A. J. Devaney, *Mathematical foundations of imaging, tomography and wavefield inversion*. Cambridge University Press, 2012.

[13] X. Liu, J. U. Kang, X. Liu, J. Kang *et al.*, "Compressive sd-oct: the application of compressed sensing in spectral domain optical coherence tomography," *Optics express*, vol. 18, no. 21, p. 22010, 2010.

[14] J. Hunt, T. Driscoll, A. Mrozack, G. Lipworth, M. Reynolds, D. Brady, and D. R. Smith, "Metamaterial apertures for computational imaging," *Science*, vol. 339, no. 6117, pp. 310–313, 2013.

[15] D. J. Brady, K. Choi, D. L. Marks, R. Horisaki, and S. Lim, "Compressive holography," *Optics Express*, vol. 17, no. 15, pp. 13 040–13 049, 2009.

[16] C. F. Cull, D. A. Wikner, J. N. Mait, M. Mattheiss, and D. J. Brady, "Millimeter-wave compressive holography," *Applied optics*, vol. 49, no. 19, pp. E67–E82, 2010.

[17] Y. Rivenson, A. Stern, and B. Javidi, "Compressive fresnel holography," *Display Technology, Journal of*, vol. 6, no. 10, pp. 506–509, 2010.
[18] K. Choi, R. Horisaki, J. Hahn, S. Lim, D. L. Marks, T. J. Schulz, and D. J. Brady, "Compressive holography of diffuse objects," *Applied Optics*, vol. 49, no. 34, pp. H1–H10, 2010.
[19] A. A. Wagadarikar, D. L. Marks, K. Choi, R. Horisaki, and D. J. Brady, "Imaging through turbulence using compressive coherence sensing," *Optics letters*, vol. 35, no. 6, pp. 838–840, 2010.

Challenges in Optical Compressive Imaging and Some Solutions

Adrian Stern, [1]Yair Rivenson and Yitzhak August

Department of Electro-Optical Engineering
[1]Department of Electrical and Computer Engineering
Ben-Gurion University of the Negev, Israel
{stern, rivenson, augusty} @bgu.ac.il

***Abstract*—The theory of compressive sensing (CS) has opened up new opportunities in the field of optical imaging. However, its implementation in this field is often not straight-forward. We list the implementation challenges that might arise in compressive imaging and present some solutions to overcome them.**

I. Introduction

Compressive sensing (CS) theory introduced a new paradigm for sampling and, subsequently, stimulated interest in its application in various fields. Imaging is a natural field for the implementation of CS theory because typical images involve a large amount of data, which facilitates efficient compression. Compressive imaging (CI) techniques were developed for various purposes, such as reducing hardware [1, 2], shortening image scanning time [1, 3], increasing image resolution [4-6] [7] and improving other imaging performance parameters [8]. CI techniques have been developed for motion tracking [9], spectral imaging [10] and holography. A review of CI techniques may be found in [11].

Principles of CI system design differ drastically from the principles used for conventional imaging. Conventional imaging seeks to perform isomorphic mapping; that is, to create images that are exact replica of the object. Ideally, each object point is mapped to a single pixel sensor so that, besides simple geometrical transformation (e.g., inversion), the captured image is a sharp copy of the object. In contrast, CS acquisition guidelines prescribe some way of mixing the information so that multiple image points are projected onto a single pixel sensor. The preferred projection is a random one so that all object points are randomly spread on the image sensors.

When coming to apply the CS framework for optical imaging and sensing one needs to consider the special characteristics of the optical data collection systems. In Sec. II we discuss the special issues and implementation limitations arising in the application of CS for optical imaging and sensing. The implementation limitations can be significantly reduced by intelligently compromising the guidelines for optimal universal CS. For instance, instead of using random projections one may use some kind of structured pseudo random projection scheme. Random convolution [12] is such an example. In subsections III A, B we present another two examples. The CI implementation challenges may also be bypassed if a specific-task system is to be designed. For example, if the task is to track motion in the scene, a technique as described in Sec. IIIC can be efficiently applied. Fortunately, there are also cases in which the optical sensing mechanism fits the CS guidelines well. Such a case is demonstrated in Sec. IV.

II. Special aspects of application of CS for imaging

Let us consider a conventional CS measurement scheme:

$$\mathbf{g} = \mathbf{\Phi f} \tag{1}$$

where the signal $\mathbf{f}$ is assumed to be k-sparse (or at least compressible) in a domain defined by the sparsifying operation $\mathbf{a} = \mathbf{\Psi f}$. For universal imaging tasks, $\mathbf{\Psi}$ should perform some random projections. In incoherent imaging $\mathbf{f} \in \mathbb{R}^N, \mathbf{g} \in \mathbb{R}^M$ and $\mathbf{\Phi} \in \mathbb{R}^{M \times N}$while in coherent imaging $\mathbf{f} \in \mathbb{C}^N, \mathbf{g} \in \mathbb{C}^M$ and $\mathbf{\Phi} \in \mathbb{C}^{M \times N}$. In the following, we shall consider the particular features of the components of (1) in the context of optical imaging and sensing.

A. The input signal

In optical sensing, the input signal $\mathbf{f}$ represents the features of the "object", such as the spatial, spatio-temporal, spectral or polarimetic distributions of the electromagnetic field or of the radiant power. We shall list the special features of $\mathbf{f}$ and their consequences.

1) Sparsity: In most imaging scenarios, the object is indeed highly compressiable, as required for CS. For instance, 2D images in the visible may be compressible by a factor of $10-50$. 3D images and hyprespectral images may be even more compressible.

2) Physical representation dimensions: The object is typically represented as a 2D or 3D distribution. Therefore, in order to adjust to the matrix-vector formalism of (1) the signal is converted into the form of a vector by lexicographic ordering. By this, analytic and computational tools developed for (1) can be directly applied; however, part of the structural information is lost. For efficient implementation of CS one should attempt to employ the structural information intelligently in the sparsifying operator $\mathbf{\Psi}$ and by introducing appropriate priors in the reconstruction process.

3) Size: The signal **f** and measurements **g** are typically large. For example, in incoherent imaging in the visible, *N* can be easily of order of 10^7 and in multidimensional imaging (such as in 3D images and hyperspectral images) it can be much larger. Obviously, this leads to computational implications in terms of reconstruction speed.

4) Non-negativity: In incoherent imaging, the signal **f** is non-negative. For efficient CI, this fact should be considered in the reconstruction process by introducing appropriate constrains in the reconstruction problem or by working with centralized signals (with the average subtracted).

B. The System Matrix

1) Size of the matrix: The size of the system matrix is $M \times N$, where *N* and *M* may be of order of $10^5 - 10^7$. Therefore the size of the system matrix is huge, leading to the following significant challenges:
Computational - $\mathbf{\Phi}$ may require hundreds of Gigabytes of storage and the application of reconstruction algorithms with such large matrices is very difficult and time-consuming.
Optical realization - Realization of random $\mathbf{\Phi}$ requires building an imaging system with a space bandwidth product (SBP) larger than $M \times N$. In other words, the imaging system needs to have at least $M \times N$ almost independent modes, or degrees of freedom. It is not trivial to design a system with such a large SBP. For example, spatial light modulators that are commonly used in CI, have an SBP of $\mathcal{O}(N)$. Therefore, in order to realize $\times M$ times larger SBP, multiple measurements are required.
Optical Calibration - Sensing systems with a large SBP also require exhaustive and time-consuming calibration processes. In order to calibrate $\mathbf{\Phi}$, one needs to measure *N* point spread functions, each having *M* samples.

2) Non-negativity: In incoherent imaging, it is impossible to realize a system matrix $\mathbf{\Phi}$ with negative entries. This means that $\mathbf{\Phi}$ spans only the positive orthant. As a result, the mutual coherence of $\mathbf{\Phi}$ is lower, indicating lower compressibility. This problem may be addressed by applying preconditioning in the reconstruction process [13] or by doubling the number of measurements to generate measurements equivalent to that of a bipolar system matrix.

C. Measured signal

1) Size: Although the dimension of the measured image **g** is smaller than that of the signal **f** $(M < N)$, in typical CI systems it is still large. Therefore, similar computation issues as with **f** (see subsection II.A) are relevant for **g** too.

2) Realness and non-negativity: Optical sensors measure irradiance, which is real and non-negative. Negative and complex values can be measured indirectly, typically by acquiring multiple measurements. For example, in compressive holography [10] complex field amplitude is measured with temporal or spatial multiplexing.

3) Dynamic range: The dynamic range of optical sensors is typically limited. For example, conventional, uncooled optoelectronic sensors in the visible have a dynamic range of 8-12 bits. At longer wavelengths, the dynamic range may be even smaller. This may set significant limitations, particularly in incoherent imaging, where $\mathbf{\Phi}$ is no-negative.

III. Feasible sampling operators for optical CS

A. Separable Sensing Matrix

One way to alleviate the complexity associated with implementing CI systems with random projections is by designing sensing operators $\mathbf{\Phi}$ that are separable in the physical dimension of the optical signal [14, 15]. For instance, for capturing a typical 2D image, one may use a sampling operator that is separable in the x-y directions. Mathematically, such a sensing operator can be expressed by means of the Kronecker product of the sensing operators in each direction, $\mathbf{\Phi} = \mathbf{\Phi}_x \otimes \mathbf{\Phi}_x$. The sensing operators in each direction,$\mathbf{\Phi}_x, \mathbf{\Phi}_y$, can be designed to perform random projections.
The SBP of an $x - y$ separable $\mathbf{\Phi}$, is $\mathcal{O}(\sqrt{N \cdot M})$; thus the matrix storage requirements and the optical sensing complexity is reduced from $\mathcal{O}(N \cdot M)$ to $\mathcal{O}(\sqrt{N \cdot M})$. Employing separable $\mathbf{\Phi}$ can be useful also in the reconstruction step as it permits using block-iterative algorithms.
The price to be paid by using a separable sensing technique is in reducing the compressibility performance. For instance, a theoretical analysis in [14] showed that for 2D images, approximately $\sqrt{N}$ times more samples are needed to achieve similar performance as with a non-separable random system matrix. An empirical study in [16] showed more relaxed requirements, indicating that the minimum number of samples required for perfect recovery is $M \approx 1.25K \log(N/k + 1)$. Analysis of compressibility of signals separable in more than two dimensions may be found in [17].
Compressive imaging with a separable sensing operator has been demonstrated for 2D images [14, 16]. Recently, an optical scheme implementing *hyperspectral* imaging with a separable sensing operator was presented in [18].

B. Optical Radon Projections for Imaging

In [3], a CI technique is proposed that uses a cylindrical lens to perform a Radon projection of the object plane on a line array of sensors. The system performs a rotational scan to capture multiple Radon projections at various angles during the scanning process. By applying reconstruction algorithms based on ℓ_1 minimization, the image can be reconstructed from many fewer projections than are needed conventionally, e.g. with filtered back-projection algorithms.
The CI approach in [3] exhibits a very good trade-off between acquisition time and system complexity. Compared to the two other main CI approaches, it allows a much faster scan than with the "single pixel camera" [1], while, on the other hand, its implementation complexity is much lower than that of the "single shot compressive imaging camera" [4]. The imaging approach presented in [3] was further improved in [19], where it is shown that angular sampling with *golden angle* steps

allows progressive compressive image acquisition. Gradual improvement of the reconstructed image is obtained by adding new projections to the existing ones without re-sampling and recalculation. Each new measurement increases the quality of the previous reconstruction, as demonstrated in Fig. 1.

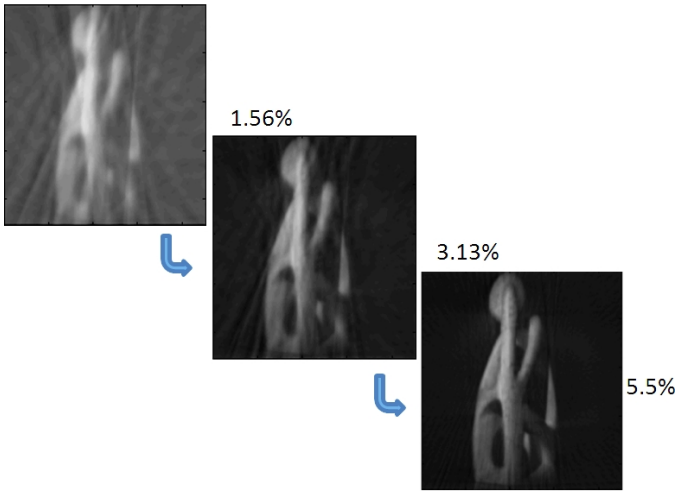


Fig. 1. Progressive compressive imaging with optical Radon projections, using obtained 1.56% (top), 3.13% (middle), 5.5% (bottom) of nominal samples (Nyquist). Image size 1280x1280 pixels.

The progressive compressive sensing approach is particularly useful when no prior knowledge about the required number of samples for good reconstruction is available. This means that the progressive Radon acquisition scheme is inherently adjustable to the type of the object imaged. The approach is also shown to be immune to sudden stopping of the scanning process, which otherwise would be intolerable with the uniform angular sampling scheme. An additional advantage of the approach is that it facilitates compressive imaging of large size images by employing ordered sets reconstruction algorithms on subsets of the data, thus remedying otherwise severe computation issues [19]. Note, for example, that the images in Fig. 1 are of megapixel size.

C. Optical Radon Projections for Motion Tracking

In the case that the task of the acquisition system is change detection or motion tracking, the signal is extremely sparse. Consider, for example, the task of tracking a point during 10 sec. with a temporal resolution of 20 milliseconds in a field of view of 1Megapixels. With conventional imagers, 500 Megapixels are acquired for this task, while here, the trajectory of the moving point can be described by only 500 pairs of Cartesian coordinates; thus $K/N = 0.5 \cdot 10^6$. Cartesian coordinates of moving objects can be obtained by measuring the temporal differences of two perpendicular Radon projections. As mentioned in Sec. IIIB, Radon projections can be obtained optically with anamorphic optical elements such as a cylindrical lens. Figure 2 depicts the concept behind change detection form two Radon projections. Consecutive temporal projections are subtracted from one another, indicating the projected location of the changes [Fig. 2 (c) and (d)]. Then the projections may be back projected to give the location of the changes on a Cartesian grid. Since the signal is extremely sparse, ℓ_1 minimization algorithms are particularly efficient.

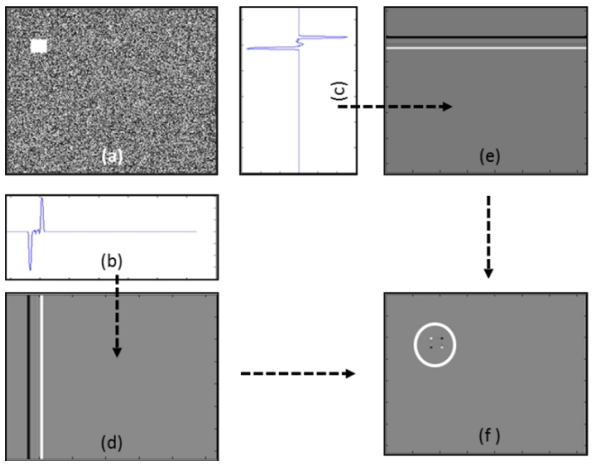


Fig. 2. Motion detection with two projections. (a) Original frame out of 2 consecutive frames; (b,c) difference between projection of two consecutive frames; (d,e) back projection of the frame difference; (f) intersection of the x,y back projections. The detected object is marked with white circle.

In practice, two projections are insufficient for detecting multiple moving objects in arbitrary directions. At least three projections are necessary to track objects moving in an arbitrary direction. In [9] we developed an optical system that essentialy perform, uses a superposition of four projections. Simulative experiments in [9] show that this system is able to track up to ten moving object points. Real experiments showed that objects can be tracked within a field of view of 500×500 pixels with approximately 250 times less samples than a conventional camera takes for the same task.

IV. Natural Optical Compressive Sensing Operators

There are cases in which the optical sensing operator fits the CS guidelines well. One such example is the free space propagation operator, described mathematically by the Fresnel transform. The Fresnel diffraction of the object field can be recorded by means of digital holography, which is found to be a physically realizable, quite simple and yet very efficient compressive sensing mechanism. Applying the CS paradigm for digital Fresnel holograms is attractive from the fact that the Fresnel and Fourier transforms are closely related. Therefore, Fourier subsampling schemes, studied extensively in CS literature, can be directly applied. In [20] it is shown that for a sufficiently large propagation distance the number of random samples in the hologram plane that is required for full reconstruction is $K \log N$, just like for the Fourier sensing case. Figure 3 shows an example of the dependence of the compressibility ratio M/N as a function of the imaging distance. From Fig. 3 it can be seen that the number of random Fresnel samples required to reconstruct the image exactly decreases with the imaging distance till it reaches an asymptote

in the region where the Fresnel propagator behaves as a Fourier transform.

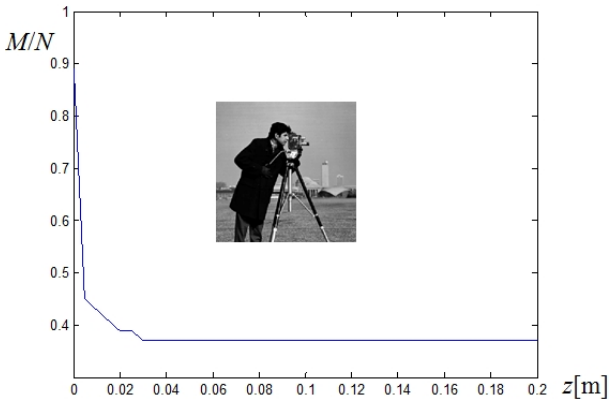


Fig. 3. Compressive sampling ratio required for full reconstruction of the Cameraman image (inset). M/N is the compressive sensing ratio and z is the recording distance.

For a recent review on compressive digital holography theory and applications the reader is referred to [10].

V. Conclusions

We have overviewed the characteristics of optical imaging that preclude straight-forward application of CS theory to imaging. In many cases, practical and physical limitations force the CI designer to deviate from basic CS guidelines. He has to compromise the randomness of the sensing operator required for universal CS by introducing some amount of structure. We presented two examples to demonstrate this.
The implementation limitations may be much less severe if a specific task is defined, as we have shown with our compressive motion detection and tracking system. In some particular cases, the particular optical sensing mechanism fits CS guidelines well. We have described compressive holography as an example of such a case.

Acknowledgment

Adrian Stern wishes to thank the Israel Science Foundation (grant No.1039/09) and Israeli Ministry of Science for supporting this research.

Reference

[1] D. Takhar, J. Laska, M. B. Wakin, M. F. Duarte, D. Baron, S. Sarvotham, K. Kelly and R. G. Baraniuk. A new compressive imaging camera architecture using optical-domain compression. Presented at Proc. IST/SPIE Symposium on Electronic Imaging. 2006,.

[2] C. F. Cull, D. A. Wikner, J. N. Mait, M. Mattheiss and D. J. Brady. Millimeter-wave compressive holography. Appl. Opt. 49(19), pp. 67-82. 2010.

[3] A. Stern. Compressed imaging system with linear sensors. *Opt. Lett. 32(21)*, pp. 3077-3079. 2007.

[4] A. Stern and B. Javidi. Random projections imaging with extended space-bandwidth product. *Journal of Display Technology 3(3)*, pp. 315-320. 2007.

[5] S. Gazit, A. Szameit, Y. C. Eldar and M. Segev. Super-resolution and reconstruction of sparse sub-wavelength images: Erratum. *Optics Express 17(25)*, pp. 23920-23946. 2009.

[6] Y. Shechtman, S. Gazit, A. Szameit, Y. C. Eldar and M. Segev. Super-resolution and reconstruction of sparse images carried by incoherent light. *Opt. Lett. 35(8)*, pp. 1148-1150. 2010.

[7] Y. Rivenson, A. Stern and B. Javidi. Single exposure super-resolution compressive imaging by double phase encoding. *Optics Express 18(14)*, pp. 15094-15103. 2010.

[8] R. Horisaki, K. Choi, J. Hahn, J. Tanida and D. J. Brady. Generalized sampling using a compound-eye imaging system for multi-dimensional object acquisition. *Opt.Express 18(18)*, pp. 19367-19378. 2010.

[9] Y. Kashter, O. Levi and A. Stern. Optical compressive change and motion detection. Appl. Opt. 51(13), pp. 2491-2496. 2012.

[10] Y. Rivenson, A. Stern and B. Javidi, "An overview of compressive sensing techniques applied in holography," *Appl. Opt., vol. 52*, pp. A423-A432, 2013.

[11] R. M. Willett, R. F. Marcia and J. M. Nichols. Compressed sensing for practical optical imaging systems: A tutorial. *Optical Engineering 50pp*. 072601. 2011.

[12] J. Romberg. Compressive sensing by random convolution. *SIAM Journal on Imaging Sciences 2(4)*, pp. 1098-1128. 2009.

[13] A. M. Bruckstein, M. Elad and M. Zibulevsky. On the uniqueness of nonnegative sparse solutions to underdetermined systems of equations. *Information Theory, IEEE Transactions on 54(11)*, pp. 4813-4820. 2008.

[14] Y. Rivenson and A. Stern. Compressed imaging with a separable sensing operator. Signal Processing Letters, *IEEE 16(6)*, pp. 449-452. 2009.

[15] M. F. Duarte and R. G. Baraniuk. Kronecker compressive sensing. *Image Processing, IEEE Transactions on 21(2)*, pp. 494-504. 2012.

[16] Y. Rivenson and A. Stern. Practical compressive sensing of large images. Presented at Digital Signal Processing, 2009 16th International Conference on. 2009,.

[17] Y. Rivenson and A. Stern. An efficient method for multi-dimensional compressive imaging. Presented at Computational Optical Sensing and Imaging. 2009,.

[18] Y. August, C. Vachman, Y. Rivenson and A. Stern, "Compressive hyperspectral imaging by random separable projections in both spatial and spectral domains," Submitted for Publication,.

[19] S. Evladov, O. Levi and A. Stern. Progressive compressive imaging from radon projections. *Optics Express 20(4)*, pp. 4260-4271. 2012.

[20] Y. Rivenson and A. Stern. Conditions for practicing compressive fresnel holography. *Opt. Lett. 36(17)*, pp. 3365-3367. 2011.

Finite-power spectral analytic framework for quantized sampled signals

Nguyen T. Thao
Dept. of Electrical Engineering
City College, CUNY, New York, NY 10031
Email: thao@ee.ccny.cuny.edu

***Abstract*—To be accurate, the theoretical spectral analysis of quantized sequences requires that the deterministic definition of power spectral density be used. We establish the functional space foundations for this analysis, which remarkably appear to be missing until now. With them, we then shed some new light on quantization error spectra in PCM and $\Sigma\Delta$ modulation.**

I. Introduction

The spectral analysis of quantized signals appears to miss clear functional foundations. In spite of their deterministic nature, quantized sequences are often theoretically described using a probabilistic definition of power spectral density. This however only leads to approximate statistical models that cannot predict quantization phenomena such as intermodulation, idle tones and limit cycles present in $\Sigma\Delta$ modulation for example. The first rigorous analysis of quantized signals in pulse code modulation (PCM) and $\Sigma\Delta$ modulation was performed by R. M. Gray [1], [2], [3] in the late 80's based on the deterministic time-averaged power function $\mathsf{M}(|x|^2)$ of a sequence $x[n]$, where

$$\mathsf{M}(x) := \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} x[n].$$

Similarly to the probabilistic case, power spectral density is obtained by taking the Fourier transform of the autocorrelation $r_x[m] := \mathsf{M}(x\,T^m x)$, where T is the shift sequence operator defined by $Tx[n] = x[n+1]$. From a functional space viewpoint, this appears at first sight as a mere extension of *energy spectral density* defined in the Hilbert space ℓ^2 of square-summable sequences. In this view, one would define the space of finite-power sequences as

$$\mathcal{P} := \big\{x \in \mathbb{C}^{\mathbb{N}} : \mathsf{M}(|x|^2) \text{ exists}\big\}$$

with the inner-product

$$\langle x, y\rangle_{\mathcal{P}} := \mathsf{M}(x^* y).$$

This is however doomed to fail as $\mathcal{P}$ is not even a linear space as shown in this paper.

The goal of this article is to rigorously establish Hilbert space foundations to the spectral analysis of finite-power sequences. Based on this, standard theorems can be applied such as the spectral properties of unitary operators. We thus provide a functional space background to the work of R. M. Gray, explaining for example why mixed quantization spectra are to be expected. We also indicate some possible generalization to overloaded $\Sigma\Delta$ modulators. The detailed proofs of the claimed results are included in [4].

II. Hilbert spaces of finite-power sequences

The first obstacle to a direct analogy between ℓ^2 and $\mathcal{P}$ is that the function $\langle\cdot,\cdot\rangle_{\mathcal{P}}$ is not defined everywhere in $\mathcal{P}\times\mathcal{P}$. Consider for example the sequence $s[n] := (-1)^{\lfloor \log_2(n)\rfloor}$ for $n \geq 1$, which is clearly an element of $\mathcal{P}$. One easily shows that $\mathsf{M}(s)$ does not converge, which is equivalent to saying that $\langle s, 1\rangle_{\mathcal{P}}$ does not exist, although the constant sequence 1 is also in $\mathcal{P}$. This simultaneously shows that $\mathsf{M}(|s+1|^2)$ does not exist, otherwise we would obtain $2\langle s, 1\rangle = \mathsf{M}(|s+1|^2) - \mathsf{M}(|s|^2) - \mathsf{M}(|1|^2)$. So $\mathcal{P}$ is not even a linear space.

To rigorously justify spectral analysis in the sense of finite power, one needs to explicitly build a Hilbert space within $\mathcal{P}$. One possible procedure is to come up with a known family of sequences $\{\varphi_k\}_{k\in K}$ of $\mathcal{P}$ such that $\langle\varphi_k, \varphi_{k'}\rangle_{\mathcal{P}}$ exists for all $k, k' \in K$ and is equal to $\delta_{k-k'}$ where δ is the Kronecker symbol. We say that $\langle\varphi_k, \varphi_{k'}\rangle_{\mathcal{P}}$ is orthonormal with respect to $\langle\cdot,\cdot\rangle_{\mathcal{P}}$. Up to some non-trivial theoretical considerations [4], it can be shown that the space of sequences of the form $x[n] = \sum_{k\in K} \alpha_k\, \varphi_k$ where $(\alpha_k)_{k\in K}$ is a family of complex coefficients whose nonzero values are in countable number and square summable, is a Hilbert space with respect to $\langle\cdot,\cdot\rangle_{\mathcal{P}}$ that is included in $\mathcal{P}$. We denote this space by

$$\mathcal{H} := \overline{\text{span}}\{\varphi_k\}_{k\in K}.$$

The most basic example of such a construction is obtained with the exponential sequences $e_\xi[n] := e^{\mathrm{i}2\pi\xi n}$. Note that $\langle e_\xi, e_{\xi'}\rangle_{\mathcal{P}} = \mathsf{M}(e_{\xi'-\xi}) = \delta_{\xi'-\xi}$ when $\xi, \xi' \in [0, 1)$, where δ is the Kronecker symbol. Thus, $\mathcal{B} := \overline{\text{span}}\{e_\xi\}_{\xi\in[0,1)}$ is a Hilbert space included in $\mathcal{P}$. This is called the space of *almost periodic sequences in the sense of Besicovitch* (Besicovitch-AP sequences) [5]. Its elements can be presented in the form

$$x[n] = \textstyle\sum_{k\in\mathbb{Z}} \alpha_k\, e_{\xi_k}[n] \qquad (1)$$

where $(\alpha_k)_{k\in\mathbb{Z}}$ is square summable and $(\xi_k)_{k\in\mathbb{Z}}$ are distinct values in $[0, 1)$.

III. Finite power by Weyl's criterion

Weyl's equidistribution criterion [6] states that a real sequence $s[n]$ is uniformly distributed modulo 1 (i.e., its fractional part is uniformly distributed in $[0, 1)$) if and only if

$\mathsf{M}(e^{\mathrm{i}2\pi ks}) = 0$ for all nonzero integers k, where $e^{\mathrm{i}2\pi ks}$ designates the sequence $(e^{\mathrm{i}2\pi ks[n]})_{n\geq 1}$. Noticing the relation $\langle e^{\mathrm{i}2\pi k\cdot s}, e^{\mathrm{i}2\pi k'\cdot s}\rangle_{\mathcal{P}} = \mathsf{M}(e^{\mathrm{i}2\pi l\cdot s})$ where $l := k-k'$, the uniform distribution of $s[n]$ is then equivalent to the orthonormality of the family of sequences $\{e^{\mathrm{i}2\pi ks}\}_{k\in\mathbb{Z}}$. We state below the multi-dimensional version [6] of this result.

Proposition 3.1: Let $\boldsymbol{s}[n]$ be a sequence of vectors in $\mathbb{R}^d$. Then $\boldsymbol{s}[n]$ is uniformly distributed modulo 1 (u.d. mod 1) *if and only if* $\left\{e^{\mathrm{i}2\pi \boldsymbol{k}\cdot\boldsymbol{s}}\right\}_{\boldsymbol{k}\in\mathbb{Z}^d}$ is an orthonormal family with respect to $\langle\cdot,\cdot\rangle_{\mathcal{P}}$ (where $\boldsymbol{k}\cdot\boldsymbol{s}$ is the dot product of $\boldsymbol{k}$ and $\boldsymbol{s}$ in $\mathbb{R}^d$).

As soon as a u.d. mod 1 sequence $\boldsymbol{s}[n]$ is found, one therefore generates a (separable) Hilbert space in $\mathcal{P}$ by forming the space

$$\mathcal{H}_{\boldsymbol{s}} := \overline{\mathrm{span}}\{e^{\mathrm{i}2\pi \boldsymbol{k}\cdot\boldsymbol{s}}\}_{\boldsymbol{k}\in\mathbb{Z}^d}.$$

The next proposition characterizes a useful subspace of $\mathcal{H}_{\boldsymbol{s}}$.

Proposition 3.2: Let $\boldsymbol{s}[n]$ be a sequence of vectors in $\mathbb{R}^d$ that is u.d. mod 1. Then, for any d-variable 1-periodic *Riemann integrable* function $h(\boldsymbol{u})$, the sequence $h(\boldsymbol{s}[n])$ belongs to $\mathcal{H}_{\boldsymbol{s}} \subset \mathcal{P}$ and yields the orthogonal expansion

$$h(\boldsymbol{s}[n]) = \sum_{\boldsymbol{k}\in\mathbb{Z}^d} \hat{h}_{\boldsymbol{k}}\, e^{\mathrm{i}2\pi \boldsymbol{k}\cdot\boldsymbol{s}[n]} \tag{2}$$

where $(\hat{h}_{\boldsymbol{k}})_{\boldsymbol{k}\in\mathbb{Z}^d}$ are the Fourier coefficients of $h(\boldsymbol{u})$.

The 1-periodicity of h implies that $h(\boldsymbol{u}+\boldsymbol{k}) = h(\boldsymbol{u})$ for all $\boldsymbol{u}\in\mathbb{R}^d$ and $\boldsymbol{k}\in\mathbb{Z}^d$. The proof of this proposition uses the more general equivalent criterion for equidistribution implying that $\mathsf{M}\big(f(\boldsymbol{s}[n])\big) = \int_{[0,1)^d} f(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u}$ for any d-variable 1-periodic Riemann integrable function $f(\boldsymbol{u})$ [6]. The conceptual difficulty of (2) is that the summation does not necessarily converge pointwise, although it converges in the sense of the norm $\|\cdot\|_{\mathcal{P}}$.

A simple case of interest is when $\boldsymbol{s}[n] = n\boldsymbol{\zeta}$ where $\boldsymbol{\zeta} = (\zeta_1,\cdots,\zeta_d)\in\mathbb{R}^d$. It is known that $\boldsymbol{s}[n] = n\boldsymbol{\zeta}$ is uniformly distributed modulo 1 if and only if $\zeta_1,\cdots,\zeta_d,1$ are rationally independent (i.e., no rational combination of $\zeta_1,\cdots,\zeta_d,1$ other than the zero combination is equal to zero) [6]. Assuming that this condition is realized and that h has the required property, Proposition 3.2 implies that $h(n\boldsymbol{\zeta})$ is a finite-power sequence that belongs to the closed space $\mathcal{H}_{\boldsymbol{s}}$ spanned by the orthonormal family $\{e_{\xi_{\boldsymbol{k}}}\}_{\boldsymbol{k}\in\mathbb{Z}^d}$ where

$$\xi_{\boldsymbol{k}} := \boldsymbol{k}\cdot\boldsymbol{\zeta} = k_1\,\zeta_1 + \cdots + k_d\,\zeta_d, \tag{3}$$

since $e^{\mathrm{i}2\pi \boldsymbol{k}\cdot\boldsymbol{s}[n]} = e^{\mathrm{i}2\pi \boldsymbol{k}\cdot(n\boldsymbol{\zeta})} = e^{\mathrm{i}2\pi\xi_{\boldsymbol{k}}n}$. In this case, $\mathcal{H}_{\boldsymbol{s}}\subset\mathcal{B}$. Hence, $h(n\boldsymbol{\zeta})$ is a Besicovitch-AP sequence. Its orthogonal expansion in $\mathcal{B}$ is explicitly

$$h(n\boldsymbol{\zeta}) = \textstyle\sum_{\boldsymbol{k}\in\mathbb{Z}^d} \hat{h}_{\boldsymbol{k}}\, e_{\xi_{\boldsymbol{k}}}[n]. \tag{4}$$

IV. Basics of spectral analysis

A. Power spectral measure

Similarly to the probabilistic approach, the power spectral density of a sequence $x[n]$ would be the Fourier transform of the autocorrelation

$$r_x[m] := \langle x, T^m x\rangle_{\mathcal{P}} \tag{5}$$

and T is the shift sequence operator defined in the introduction. The sequence $r_x[m]$ is guaranteed to exist if $x[n]$ is shown to be in some Hilbert space $\mathcal{H}\subset\mathcal{P}$ that is invariant by T. Assume that this is the case. Then, T is a *unitary* operator of $\mathcal{H}$ as it preserves the inner-product and is invertible. Now rigorously speaking, $r_x[m]$ may not have a Fourier transform. Given the positive definite property of $r_x[m]$, it is however shown to be at least the Fourier coefficients of a positive measure μ_x [7], which we call the *power spectral measure* of $x[n]$. By Lebesgue's decomposition theorem, μ_x has in general *three* components, a pure-point part (purely discrete measure composed of Dirac masses), an absolutely-continuous part (the actual and only part that yields a power spectral density by Radon-Nikodym derivative) and a singular-continuous part.

B. Spectral decomposition

A key to analyzing the measure structure of μ_x is to decompose $\mathcal{H}$ as an orthogonal sum

$$\mathcal{H} = \textstyle\bigoplus_{k\in K} \mathcal{I}_k \tag{6}$$

of Hilbert subspaces $\mathcal{I}_k$ that are invariant by T. Writing $x = \sum_{k\in K} x_k$ where x_k is the orthogonal projection of x onto $\mathcal{I}_k$, it is clear that $\langle x_k, T^m x_{k'}\rangle = 0$ when $k\neq k'$. This leads to the decompositions

$$r_x[m] = \textstyle\sum_{k\in K} r_{x_k}[m] \qquad \text{and} \qquad \mu_x = \sum_{k\in K}\mu_{x_k}.$$

In this reduction process, we will encounter in this paper two types of measure μ_{x_k}.

1) Simple discrete spectrum: This is the case where $\mathcal{I}_k$ is spanned by a single vector φ which we can assume of norm 1. By T-invariance of $\mathcal{I}_k$, φ must be an eigenfunction of T. Since T is unitary, it is known that the eigenvalue of φ must be of the form $e^{\mathrm{i}2\pi\xi}$ where $\xi\in[0,1)$. Since $T^m\varphi = e^{\mathrm{i}2\pi\xi m}\varphi$, the sequence x_k which is of the form $a\,\varphi$ yields the autocorrelation $r_{x_k}[m] = \langle a\varphi, aT^m\varphi\rangle_{\mathcal{P}} = |a|^2 e^{\mathrm{i}2\pi\xi m}$. The corresponding spectral measure μ_{x_k} is then discrete and equal to $|a|^2\delta_\xi$ where δ_ξ denotes the Dirac mass at frequency ξ.

2) Simple absolutely-continuous spectrum: This is the case where $\mathcal{I}_k$ is the closed span of an orthonormal family of the form $\{T^n\varphi\}_{n\in\mathbb{Z}}$. The sequence x_k is then of the form $\sum_{n\in\mathbb{Z}} a_n T^n\varphi$, where a_n is a square-summable sequence. By orthonormality of $\{T^n\varphi\}_{n\in\mathbb{Z}}$, one finds that

$$r_{x_k}[m] = \langle x_k, T^m x_k\rangle_{\mathcal{P}} = \textstyle\sum_{n\in\mathbb{Z}} a^*_{n+m}\, a_n.$$

This is precisely the *finite-energy* autocorrelation of a_{-n}. In this case, $r_{x_k}[m]$ yields a Fourier transform $R_{x_k}(\xi) = |A(-\xi)|^2$ where $A(\xi)$ is the Fourier transform of a_n in $L^2([0,1))$, making $R_{x_k}(\xi)$ a function in $L^1([0,1))$. This makes the measure μ_{x_k} absolutely-continuous.

C. Besicovitch almost-periodic sequences

The most straightforward example of spectral decomposition is achieved in the space $\mathcal{B}$. Every function of its orthonormal basis $\{e_\xi\}_{\xi\in[0,1)}$ turns out to be an eigenfunction of T since

$$Te_\xi = e^{\mathrm{i}2\pi\xi}\, e_\xi.$$

Once a Besicovitch-AP sequence $x[n]$ is written in the form (1), it can be presented as element of a space sum (6) with $K := \mathbb{Z}$ and $\mathcal{I}_k := \text{span}\{e_{\xi_k}\}$. This falls in the case of Section IV-B1. One then obtains the discrete power spectral measure $\mu_x = \sum_{k\in\mathbb{Z}} |\alpha_k|^2\, \delta_{\xi_k}$.

V. PCM with trigonometric polynomial input

We show that the quantizer error sequence $\epsilon[n]$ from the pulse code modulation (PCM) of a finite sum of sinusoids is Besicovitch almost-periodic, thus yielding a purely discrete power spectral measure. PCM consists in transforming every sample of a sequence $x[n]$ individually by a nonlinear memoryless scalar function $Q(\cdot)$ that is basically piecewise constant. This results in an error sequence $\epsilon[n] := x[n]-Q(x[n])$[1]. An input sequence $x[n]$ that is a finite sum of sinusoids can always be expanded as a trigonometric polynomial[2]. $x[n] = \sum_{k=1}^{p} \alpha_k\, e^{\mathrm{i}2\pi\zeta_k n}$ where ζ_k are distinct frequency values in $[0,1)$. By defining the p-variable 1-periodic function $\mathrm{x}(u_1,\cdots,u_p) = \sum_{k=1}^{p} \alpha_k\, e^{\mathrm{i}2\pi u_k}$, one can write $x[n] = \mathrm{x}(n\boldsymbol{\zeta})$ where $\boldsymbol{\zeta} := (\zeta_1,\cdots,\zeta_p)$. Thus, the error sequence is of the form

$$\epsilon[n] = h(n\boldsymbol{\zeta}) \qquad \text{where} \qquad h(\boldsymbol{u}) := Q(\mathrm{x}(\boldsymbol{u}))-\mathrm{x}(\boldsymbol{u}). \quad (7)$$

The function $h(\boldsymbol{u})$ is p-variable, 1-periodic and Riemann integrable since Q is piecewise constant and $\mathrm{x}(\boldsymbol{u})$ is continuous. Under the condition that $\zeta_1,\cdots,\zeta_p,1$ are rationally independent, we know from Section III that $\epsilon[n]$ is a Besicovitch-AP sequence of orthogonal expansion (4). From Section IV-C, we conclude that the power spectral measure of $\epsilon[n]$ is purely discrete with the autocorrelation expansion

$$r_\epsilon[m] = \sum_{\boldsymbol{k}\in\mathbb{Z}^r} |\hat{h}_{\boldsymbol{k}}|^2\, e^{\mathrm{i}2\pi\xi_{\boldsymbol{k}} m}.$$

The discrete frequencies of the spectrum of $\epsilon[n]$ are the values $\xi_{\boldsymbol{k}}$ given by (3) and are nothing but the *intermodulation products* of the fundamental input frequencies $\zeta_1,\cdots,\zeta_p$, as seen in (3).

VI. Quantization error in ideal ΣΔ modulation

In an ideal ΣΔ modulator with a polynomial trigonometric input, we show that the errors due to quantization can be presented as output of a system of the form

$$\begin{cases} \boldsymbol{s}[n] = \boldsymbol{M}\boldsymbol{s}[n-1] + \boldsymbol{\tau} \\ \epsilon[n] = h(\boldsymbol{s}[n]) \end{cases} \quad (8)$$

where h is a d-variable 1-periodic Riemann-integrable function and $\boldsymbol{M}$ is a square matrix that is *unimodular* (i.e., invertible with integer entries) and *unipotent* (i.e., with all eigenvalues equal to 1).

A. General equations

The general diagram of a ΣΔ modulator is shown in Figure 1 and defines the signal notation we will use. In this section, the modulator is assumed to be ideal, i.e., $h[n]$ is a purely differentiating sequence of z-transform $H(z) = (1-z^{-1})^r$ and the quantizer is not overloaded. Like in PCM, we define the quantizer error to be $\epsilon[n] = y[n]-q[n] = -e[n]$. Using the vector sequence $\boldsymbol{v}[n] = (v_1[n],\cdots,v_r[n])$ such that $v_i[n]$ is the $(r-i)$th order differentiation of $\epsilon[n]$, one can show the following system of equation

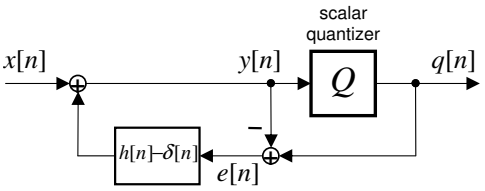


Fig. 1. General diagram of a ΣΔ modulator

$$\begin{cases} \boldsymbol{v}[n] = \boldsymbol{L}\,\boldsymbol{v}[n-1] + \boldsymbol{1}(x[n]-q[n]) \\ \epsilon[n] = \boldsymbol{j}\cdot\boldsymbol{v}[n] \end{cases} \quad (9)$$

where $\boldsymbol{L}$ is the lower-triangular matrix of 1's and size r, $\boldsymbol{1} := (1,\cdots,1) \in \mathbb{R}^r$ and $\boldsymbol{j} := (0,\cdots,0,1) \in \mathbb{R}^r$. The outstanding property of this state model is that $\boldsymbol{L}$, $\boldsymbol{1}$, $\boldsymbol{j}$ and $q[n]+\frac{1}{2}$ all have entries or coefficients that are integers. So, if we recursively construct the sequence

$$\boldsymbol{u}[n] = \boldsymbol{L}\,\boldsymbol{u}[n-1] + \boldsymbol{1}\,x'[n] \quad (10)$$

where $x'[n] := x[n]+\frac{1}{2}$ and $\boldsymbol{u}[0] = \boldsymbol{v}[0]$, then the relation $\boldsymbol{v}[n]-\boldsymbol{u}[n] \in \mathbb{Z}^r$ is maintained for all $n \geq 0$. Then,

$$\epsilon[n] \equiv \boldsymbol{j}\cdot\boldsymbol{u}[n] \mod 1. \quad (11)$$

B. Trigonometric polynomial inputs

Assuming that $x[n]$ is a trigonometric polynomial, we show that $\epsilon[n]$ can be at least determined modulo 1 via an *autonomous* system of the type

$$\begin{cases} \boldsymbol{s}[n] = \boldsymbol{M}\boldsymbol{s}[n-1] + \boldsymbol{\tau} \\ \epsilon[n] \equiv g(\boldsymbol{s}[n]) \mod 1 \end{cases} \quad (12)$$

where $\boldsymbol{M}$ is unimodular and unipotent, and g is a continuous function of $\mathbb{R}^d$ such that $g(\boldsymbol{s})-g(\boldsymbol{s}') \in \mathbb{Z}^d$ when $\boldsymbol{s}-\boldsymbol{s}' \in \mathbb{Z}^d$. When $x'[n]$ is equal to a constant $\bar{x}$, this is easily achieved from (10) and (11) by taking $\boldsymbol{s}[n] = \boldsymbol{u}[n]$, $\boldsymbol{M} = \boldsymbol{L}$, $\boldsymbol{\tau} = \boldsymbol{1}\bar{x}$ and $h(\boldsymbol{u}) = \langle \boldsymbol{j}\cdot\boldsymbol{u}\rangle_I$. When $x'[n]$ is not constant, one goes from (10-11) to (12) by the technique of *skew-product* first used in ΣΔ modulation in [8]. This requires the following preliminary result.

Proposition 6.1: Let $\mathrm{x}(\boldsymbol{u}) = \sum_{k=1}^{p} \alpha_k\, e^{\mathrm{i}2\pi u_k}$ and $\boldsymbol{\zeta} \in (0,1)^p$. An explicit solution to the equation

$$\boldsymbol{w}[n] = \boldsymbol{L}\,\boldsymbol{w}[n-1] + \boldsymbol{1}\,\mathrm{x}(n\boldsymbol{\zeta}) \quad (13)$$

is $\boldsymbol{w}[n] = \mathbf{x}(n\boldsymbol{\zeta})$, where $\mathbf{x}(\boldsymbol{u}) := \sum_{k=1}^{p} \alpha_k\, e^{\mathrm{i}2\pi u_k}\, \boldsymbol{x}_{\zeta_k}$ and $\boldsymbol{x}_\zeta := (x_\zeta, x_\zeta^2,\cdots,x_\zeta^r)$ with $x_\zeta := (1-e^{\mathrm{i}2\pi\zeta})^{-1}$.

Calling $\bar{x}$ the constant component of $x'[n]$, we express the "AC-component" $x'[n]-\bar{x}$ in the form $\mathrm{x}(n\boldsymbol{\zeta})$ as was done in Section V with the difference that $\boldsymbol{\zeta} \in (0,1)^p$. With the resulting function $\mathbf{x}(\boldsymbol{u})$ as obtained in the above proposition, we obtain the following result.

[1]The usual convention for a system error is $e[n] = Q(x[n])-x[n]$. Working with the sequence $\epsilon[n] := -e[n]$ will prove more convenient from a dynamical system viewpoint.

[2]This is in the largest sense of sequences of the form $x[n] := \sum_{k=1}^{N} \alpha\, e^{\mathrm{i}2\pi\xi_k n}$, where the ξ_k's are not necessarily harmonics of a single frequency.

Proposition 6.2: The sequence $\boldsymbol{s}[n] := \big(n\boldsymbol{\zeta}, \boldsymbol{u}[n]-\mathbf{x}(n\boldsymbol{\zeta})\big)$ achieves the system (12), with

$$\boldsymbol{M} := \begin{bmatrix} \boldsymbol{I} & 0 \\ 0 & \boldsymbol{L} \end{bmatrix}, \quad \boldsymbol{\tau} = (\boldsymbol{\zeta}, \mathbf{1}\bar{x}) \quad \text{and} \quad g(\boldsymbol{\theta}, \bar{\boldsymbol{u}}) := \boldsymbol{j} \cdot (\bar{\boldsymbol{u}}+\mathbf{x}(\boldsymbol{\theta}))$$

where $\boldsymbol{I}$ is the identity matrix of size p.

The central argument of the proof is that the second component $\bar{\boldsymbol{u}}[n] := \boldsymbol{u}[n]-\mathbf{x}(n\boldsymbol{\zeta})$ of $\boldsymbol{s}[n]$ satisfies the recursion $\bar{\boldsymbol{u}}[n] = \boldsymbol{L}\,\bar{\boldsymbol{u}}[n-1] + \mathbf{1}\,\bar{x}$.

C. Non-overloaded quantizer

The fact that the quantizer is not overloaded implies that $\epsilon[n]$ remains in the interval $I := [-\frac{1}{2}, \frac{1}{2})$. Let $\langle\cdot\rangle_I$ be the unique 1-periodic function that is identity in I (explicitly equal to $\langle\cdot+\frac{1}{2}\rangle - \frac{1}{2}$ where $\langle\cdot\rangle$ is the fractional part function). Since $\epsilon[n] = \langle\epsilon[n]\rangle_I$ and $\epsilon[n] \equiv g(\boldsymbol{s}[n]) \bmod 1$, then $\epsilon[n] = h(\boldsymbol{s}[n])$ where $h(\boldsymbol{s}) := \langle g(\boldsymbol{s})\rangle_I$. One easily verifies that h is 1-periodic and Riemann integrable. The system (8) is thus fully achieved.

VII. Spectral analysis in "unipotent" dynamical system

In this section, we perform the spectral analysis of sequences $\epsilon[n]$ output by systems of the type (8) with a unimodular and unipotent matrix $\boldsymbol{M}$.

A. State equidistribution

Proposition 7.1: Let $\boldsymbol{s}[n]$ be a sequence satisfying (8) where $\boldsymbol{M}$ is unimodular and unipotent. Then, $\boldsymbol{s}[n]$ is u.d. mod 1 *if and only if* $\boldsymbol{k} \cdot \boldsymbol{\tau} \notin \mathbb{Z}$ for all $\boldsymbol{k} \in \mathbb{Z}^d\backslash\{\mathbf{0}\}$ such that $\boldsymbol{M}^\top\boldsymbol{k} = \boldsymbol{k}$.

The basic ingredients of the proof are as follows. Due to the unipotent property of $\boldsymbol{M}$, the components of $\boldsymbol{s}[n]$ are shown to be polynomial sequences. Next, one uses the known fact that a polynomial sequence is u.d. mod 1 if and only if at least one of the coefficients of its non-constant terms is irrational [6]. In the setting of ideal $\Sigma\Delta$ modulation of the previous section, this proposition implies that $\bar{x}, \zeta_1, \cdots, \zeta_r, 1$ must be rationally independent for $\boldsymbol{s}[n]$ to be u.d. mod 1. In the constant input case, this reduces to the condition that $\bar{x}$ be irrational.

B. Spectral analysis

Assume that the condition of Proposition 7.1 is realized. Proposition 3.2 then implies that $\epsilon[n]$ is a finite-power sequence with the orthogonal expansion

$$\epsilon = \textstyle\sum_{\boldsymbol{k}\in\mathbb{Z}^d} \hat{h}_{\boldsymbol{k}}\, \varphi_{\boldsymbol{k}} \qquad \text{where} \qquad \varphi_{\boldsymbol{k}}[n] := e^{\mathrm{i}2\pi\boldsymbol{k}\cdot\boldsymbol{s}[n]}.$$

To derive the autocorrelation $r_\epsilon[m] := \langle\epsilon, T^m\epsilon\rangle_{\mathcal{P}}$, one needs to apply T^m on the basis vectors $\varphi_{\boldsymbol{k}}$. From the mere relation $\boldsymbol{s}[n+1] = \boldsymbol{M}\boldsymbol{s}[n]+\boldsymbol{\tau}$, one finds that

$$T\varphi_{\boldsymbol{k}} = e^{\mathrm{i}2\pi\xi_{\boldsymbol{k}}}\, \varphi_{\boldsymbol{k}'} \qquad \text{where} \qquad \xi_{\boldsymbol{k}} := \boldsymbol{k}\cdot\boldsymbol{\tau} \qquad (14)$$

and $\boldsymbol{k}' := \boldsymbol{M}^\top\boldsymbol{k} \in \mathbb{Z}^d$. This first implies that the space $\mathcal{H} := \overline{\text{span}}\{\varphi_{\boldsymbol{k}}\}_{\boldsymbol{k}\in\mathbb{Z}^d}$ is T-invariant. But as $\boldsymbol{M}$ is unimodular, it defines a permutation of $\mathbb{Z}^d$. Hence, the action of T on $\{\varphi_{\boldsymbol{k}}\}_{\boldsymbol{k}\in\mathbb{Z}^d}$ amounts to a permutation of the basis vectors plus some phase shift. One then obtains an orthogonal decomposition of $\mathcal{H}$ of the type (6) with the following definitions: $\mathcal{I}_{\boldsymbol{k}} := \overline{\text{span}}\{\varphi_{\boldsymbol{l}}\}_{\boldsymbol{l}\in\mathcal{O}(\boldsymbol{k})}$, $\mathcal{O}(\boldsymbol{k}) := \big\{\boldsymbol{k}_n : n \in \mathbb{Z}\big\}$, $\boldsymbol{k}_n := {\boldsymbol{M}^\top}^n\boldsymbol{k}$ and K equal to a subset of $\mathbb{Z}^r$ such that $\{\mathcal{O}(\boldsymbol{k})\}_{\boldsymbol{k}\in K}$ is a partition of $\mathbb{Z}^r$. We know from Section IV-B that $\mu_\epsilon = \sum_{\boldsymbol{k}\in K}\mu_{\epsilon_{\boldsymbol{k}}}$ where $\epsilon_{\boldsymbol{k}}$ is the orthogonal projection of ϵ onto $\mathcal{I}_{\boldsymbol{k}}$.

Proposition 7.2: For any $\boldsymbol{k} \in \mathbb{Z}^d$, the following statements are equivalent: (i) $\mathcal{O}(\boldsymbol{k})$ is a finite orbit, (ii) $\mathcal{O}(\boldsymbol{k}) = \{\boldsymbol{k}\}$, (iii) $\boldsymbol{k}$ belongs to the set $J_{\boldsymbol{M}} := \big\{\boldsymbol{l} \in \mathbb{Z}^d : \boldsymbol{M}^\top\boldsymbol{l} = \boldsymbol{l}\big\}$.

While $(ii) \Leftrightarrow (iii)$ is trivial, $(i) \Leftrightarrow (iii)$ uses the unipotent property of $\boldsymbol{M}$. When $\boldsymbol{k} \in J_{\boldsymbol{M}}$, $\mathcal{I}_{\boldsymbol{k}} = \text{span}\{\varphi_{\boldsymbol{k}}\}$, so $\mu_{\epsilon_{\boldsymbol{k}}}$ is a single Dirac mass according to Section IV-B1. When $\boldsymbol{k} \in K\backslash J_{\boldsymbol{M}}$, one easily sees from (14) that $\{T^n\varphi_{\boldsymbol{k}}\}_{n\in\mathbb{Z}}$ is equal to $\{\varphi_{\boldsymbol{k}_n}\}_{n\in\mathbb{Z}}$ up to some phase shifts, and is therefore an orthonormal basis of $\mathcal{I}_{\boldsymbol{k}}$. So $\mu_{\epsilon_{\boldsymbol{k}}}$ is absolutely continuous according to Section IV-B2. We conclude that the power spectral measure of $\epsilon[n]$ is *a priori* mixed, with a pure-point part and an absolutely continuous-part equal to $\mu_{\mathring{\epsilon}}$ and $\mu_{\bar{\epsilon}}$, respectively, where $\mathring{\epsilon} := \sum_{\boldsymbol{k}\in J_{\boldsymbol{M}}}\epsilon_{\boldsymbol{k}}$ and $\bar{\epsilon} := \sum_{\boldsymbol{k}\in K\backslash J_{\boldsymbol{M}}}\epsilon_{\boldsymbol{k}}$, but *no* singular-continuous part.

VIII. Discussion and extensions

In his work [1], [2], [3], R. M. Gray found with ideal $\Sigma\Delta$ modulators the more precise result that μ_ϵ is either purely discrete or uniform (white noise), which is a particular case of absolutely-continuous spectral measure. The reason for this special result is particular to the specific function $h(\boldsymbol{s}) = \langle g(\boldsymbol{s})\rangle_I$ found in Section VI-C. It is however shown in [4] that systems of the type (8) are achieved with a class of $\Sigma\Delta$ modulators that are more representative of practical configurations [9] including quantizer overloading, and yield truly mixed spectra. Finally, although this paper constantly assumed the uniform distribution of $\boldsymbol{s}[n]$ modulo 1, similar results can be obtained with absolutely no condition on $\boldsymbol{s}[n]$. This uses the result that the closure of the set of points $\boldsymbol{s}[n]$ modulo 1 is a compact group and the generalized notion of uniform distribution in a compact group [6].

Acknowledgment

A large number of the mathematical results have been proved in collaboration with Sinan Güntürk [4].

References

[1] R. M. Gray, "Spectral analysis of quantization noise in a single loop sigma-delta modulator with dc input," *IEEE Trans. Commun.*, vol. 37, pp. 588-599, June 1989.

[2] R. M. Gray, "Quantization noise spectra," *IEEE Trans. Inform. Th.*, vol. 36, pp. 1220-1244, Nov. 1990.

[3] P.-W. Wong and R. M. Gray, "Two-stage sigma-delta modulation," *IEEE Trans. Acoust., Speech, Signal Proc.*, vol. 38, pp. 1937-1952, Nov. 1990.

[4] C. S. Güntürk and N. T. Thao, "Signal-analytic theory of $\Sigma\Delta$ quantization," *in preparation*.

[5] C. Corduneanu, *Almost Periodic Oscillations and Waves*, Springer, 2009.

[6] L. Kuipers and H. Niederreiter, *Uniform Distribution of Sequences*, Wiley, 1974.

[7] M. Queffélec, *Substitution Dynamical Systems - Spectral Analysis*, 2nd ed., Springer, 2010.

[8] A. Teplinsky, E. Condon and O. Feely, "Driven interval shift dynamics in sigma-delta modulators and phase-locked loops," IEEE Trans. Circuits and Systems I, vol. 52, pp. 1224-1235, June 2005.

[9] R. Schreier, "An empirical study of high-order single-bit delta-sigma modulators," IEEE Trans. Circuits and Systems II, vol. 40, pp. 461-466, Aug. 1993.

Non-Convex Decoding for $\Sigma\Delta$-Quantized Compressed Sensing

Evan Chou
Courant Institute of Mathematical Sciences, NYU
Email: chou@cims.nyu.edu

***Abstract*—Recently Güntürk et al. showed that $\Sigma\Delta$ quantization is more effective than memoryless scalar quantization (MSQ) when applied to compressed sensing measurements of sparse signals. MSQ with the l^1 decoder recovers an approximation to the original sparse signal with an error proportional to the quantization step size δ_Q. For an r-th order $\Sigma\Delta$ scheme the reconstruction accuracy can be improved by a factor of $(m/k)^{\alpha(r-1/2)}$ for any $0 < \alpha < 1$ if $m \gtrsim k(\log N)^{1/(1-\alpha)}$, with high probability on the measurement matrix. The method requires a preliminary support recovery stage for which r cannot be too large and δ_Q must be sufficiently small. In this paper, we remove this requirement, showing that the constrained l^0 and l^τ (for sufficiently small τ) minimization problems subject to a $\Sigma\Delta$-type quantization constraint would approximate the original signal from the $\Sigma\Delta$ quantized measurements with a comparable reconstruction accuracy. We note that these results allow us to achieve root-exponential reconstruction accuracy while using a fixed quantization alphabet.**

I. Introduction

The robust recovery results in compressed sensing, e.g. [3], [6], [11] showed that sparse vectors could be recovered from compressed sensing measurements even when the measurements are perturbed. Quantization of these measurements introduces such a perturbation from which the robust recovery result allows us to recover.

To fix notation, let N be the ambient dimension of the sparse signal x that we wish to recover. Define the sparsity measure $\|x\|_0 := |\{i : x(i) \neq 0\}|$ and let Σ_k^N be the set of all k-sparse vectors in N dimensions $\Sigma_k^N := \{x \in \mathbb{R}^N : \|x\|_0 \leq k\}$. We will use Φ to denote the $m \times N$ measurement matrix, where we wish to recover x from the quantization of the measurements $y = \Phi x$.

Mathematically, a quantizer maps the measurement space $\mathbb{R}^m$ to a finite set, which we will assume to be of the form $\mathcal{A}^m$, where the quantization alphabet $\mathcal{A}$ is a finite arithmetic progression of step size δ_Q. For memoryless scalar quantization (MSQ), each measurement is simply rounded to the nearest element of $\mathcal{A}$. For an r-th order $\Sigma\Delta$ scheme, the quantization is found by solving a difference equation

$$y - q = D^r u \tag{1}$$

for $q \in \mathcal{A}^m$ and $u \in \mathbb{R}^m$, where $\|u\|_\infty$ should be bounded independently of m. D^r is the r-th difference operator: in matrix form, D is 1 on the diagonal and -1 on the subdiagonal, with zeros elsewhere. Note that MSQ corresponds to the case $r = 0$.

The authors in [8] investigated the use of $\Sigma\Delta$ quantization for a specific class of compressed sensing matrices: the random $m \times N$ matrices with each entry drawn independently from the standard Gaussian distribution, $\mathcal{N}(0, 1)$. For MSQ, the quantization introduces an error of at most $\delta_Q/2$ per measurement, and the corresponding recovery error is a constant multiple of δ_Q. For r-th order $\Sigma\Delta$, the quantization introduces an error of at most $2^{r-1}\delta_Q$ per measurement, but the error vector is highly structured. Once the support is recovered, for instance via l^1 minimization, the Sobolev-dual approximation (Equation 6) yields an error of at most $\delta_Q(m/k)^{-\alpha(r-1/2)}$ for some $0 < \alpha < 1$, when $m \gtrsim k(\log N)^{1/(1-\alpha)}$. However, the method to recover the support requires that $5\sqrt{2} \cdot 2^r \delta_Q < \min_{i:x_i \neq 0} |x_i|$ [8]. Thus δ_Q needs to be small, and r cannot be too large.

Suppose we use r-th order $\Sigma\Delta$ with step size δ_Q to produce $q \in \mathcal{A}^m$ with $\|u\|_\infty \leq \mu$. Rearranging Equation 1 shows that $\|D^{-r}(y-q)\|_2 \leq \sqrt{m}\mu$. We will show in Proposition II.2 that the sparsest solution satisfying this quantization constraint

$$x^{0,\mu} := \underset{\|D^{-r}(\Phi z - q)\|_2 \leq \sqrt{m}\mu}{\operatorname{Argmin}} \|z\|_0 \ , \tag{2}$$

approximates the original sparse vector with the same accuracy up to a constant as the Sobolev-dual approximation. Then we will show in Theorem IV.3 that if we solve the non-convex minimization

$$x^{\tau,\mu} := \underset{\|D^{-r}(\Phi z - q)\|_2 \leq \sqrt{m}\mu}{\operatorname{Argmin}} \|z\|_\tau \ , \tag{3}$$

where $\|z\|_\tau = \left(\sum_{i=1}^N |z(i)|^\tau\right)^{1/\tau}$, then there is a value of $\tau > 0$ sufficiently small so that the minimizer approximates the original sparse vector with the same accuracy up to a constant as the Sobolev-dual approximation. In Section V we note that given a bit budget R for quantizing the measurements, we can now achieve a reconstruction accuracy of the form $\exp(-c(R/k)^\alpha)$ where c is an absolute constant. Previously, Krahmer et al. showed a similar result for $\Sigma\Delta$ quantization of frame coefficients for specially designed frames [9]. Finally in section VI we briefly discuss approaches for tackling the minimization problems.

II. $\Sigma\Delta$-Quantization and Sobolev Dual Recovery

Suppose we quantize the measurements $y = \Phi x$ with r-th order $\Sigma\Delta$, i.e. we solve Equation (1) for $q \in \mathcal{A}^m$ and $u \in \mathbb{R}^m$. We highlight two approaches for accomplishing this, where details can be found in [5], [7]:

A. The simplest greedy method chooses q_{i+1} which would minimize the corresponding value of $|u_{i+1}|$ in the equation. The result is a solution that requires an alphabet of size $2^r + 2\|y\|_\infty/\delta_Q$ and has the bound $\|u\|_\infty \le \delta_Q/2$.

B. An alternative method which can be viewed as a greedy method on a different but related difference equation decreases the required alphabet size to

$$C_1 + 2\|y\|_\infty/\delta_Q \tag{4}$$

for some absolute constant C_1 but increases the bound to

$$\|u\|_\infty \lesssim (C_2 r)^r \delta_Q \tag{5}$$

This method allows us to use a fixed quantization alphabet for all r.

Consider any solution with $\|u\|_\infty \le \mu$. The difference equation can be rewritten as

$$D^{-r}\Phi x - D^{-r} q = u \ .$$

We review the results from [2] concerning the Sobolev dual. Suppose that an oracle tell us the support T of x. We can then focus on just Φ_T, the $m \times k$ submatrix with columns corresponding to the index set T. Taking the pseudoinverse,

$$x - (D^{-r}\Phi_T)^\dagger D^{-r} q = (D^{-r}\Phi_T)^\dagger u \ .$$

Note that if $r = 0$, the quantizer is MSQ, and u is the quantization error vector with norm $\sqrt{m}\delta_Q/2$. Taking the pseudoinverse of Φ_T recovers an approximation $\hat{x}^{(0)} = \Phi_T^\dagger q$ with error

$$\|x - \hat{x}^{(0)}\|_2 \le \|\Phi_T^\dagger\|_2 \sqrt{m}\delta_Q/2 \ .$$

From the restricted isometry property, the singular values of every submatrix of Φ with $|T| = k$ columns is concentrated around $\sqrt{m}$ with high probability if the entries of Φ are drawn independently from $\mathcal{N}(0,1)$; so $\|\Phi_T^\dagger\|_2 \sim 1/\sqrt{m}$ and the error bound is proportional to δ_Q and does not decrease with m, as stressed in [8].

For $r > 0$, we see that $\hat{x}^{(r)} = (D^{-r}\Phi_T)^\dagger D^{-r} q$ recovers an approximation with error

$$\|x - \hat{x}^{(r)}\|_2 \le \frac{\sqrt{m}\mu}{\sigma_{\min}(D^{-r}\Phi_T)} \ . \tag{6}$$

Note $(D^{-r}\Phi_T)^\dagger D^{-r}$ is precisely the r-th order Sobolev dual of Φ_T. Here we will restate the relevant result from [8, Theorem 3.8] about the smallest singular value:

Theorem II.1. *Let Φ be an $m \times N$ random matrix whose entries are i.i.d. $\mathcal{N}(0,1)$. Let $0 < \alpha < 1$ and suppose for some $C_3 = C_3(r)$*

$$\frac{m}{s} \ge C_3 (\log N)^{1/(1-\alpha)} \ .$$

Then there exist constants C_4, C_5 depending only on r such that with probability at least $1 - \exp(-C_4 m^{1-\alpha} s^\alpha)$ on the draw of Φ, every $m \times s$ submatrix E of Φ satisfies

$$\sigma_{\min}(D^{-r}E) \ge C_5 \sqrt{m} (m/s)^{\alpha(r-1/2)}$$

This theorem implies that given the support, the error for the Sobolev dual recovery (6) becomes $C_5 (m/k)^{-\alpha(r-1/2)}\mu$.

We now use Theorem II.1 to show that solving (2) will recover the support and have an accuracy matching that of the Sobolev dual.

Proposition II.2. *Let Φ be an $m \times N$ random matrix whose entries are i.i.d $\mathcal{N}(0,1)$. Let α, m, r, and $s = 2k$ satisfy the conditions of Theorem II.1. Suppose $x \in \Sigma_k^N$ and let q be the quantization of Φx using r-th order $\Sigma\Delta$. Let $\|u\|_\infty \le \mu$. Then the minimizer $x^{0,\mu}$ of (2) recovers an approximation of x with error*

$$\|x^{0,\mu} - x\|_2 \le \frac{2}{C_5} \left(\frac{m}{2k}\right)^{-\alpha(r-1/2)} \mu \ .$$

Proof: Suppose T is the support of x, and let $x' = x^{0,\mu}$ with support T'. Since x, x' are both feasible and x' is the sparsest feasible point, $|T'| = |x'|_0 \le |x|_0 \le k$. Then by Theorem II.1,

$$\|x - x'\|_2 \le \frac{\|D^{-r}\Phi_{T\cup T'}(x - x')\|_2}{C_5 \sqrt{m} \left(\frac{m}{2k}\right)^{\alpha(r-1/2)}} \ .$$

Note that $D^{-r}\Phi_{T\cup T'}(x - x') = D^{-r}\Phi(x - x')$. Using the triangle inequality and feasibility conditions,

$$\begin{aligned}\|D^{-r}\Phi(x - x')\|_2 &\le \|D^{-r}(\Phi x' - q)\|_2 + \|D^{-r}(\Phi x - q)\|_2 \\ &\le 2\sqrt{m}\mu \ .\end{aligned}$$

The result follows from substitution. ∎

III. Robustness of l^τ minimization

We will follow the approaches of [6], [12] to study $\|x\|_\tau$ minimization as stated in (3). As in [6], we will state our results in terms of the condition numbers of submatrices of the measurement matrix:

Definition III.1. Define $a_s(A)$ to be the largest a and $b_s(B)$ to be smallest b such that the following holds:

$$a\|z\|_2 \le \|Az\|_2 \le b\|z\|_2 \text{ for all } z \in \Sigma_s^N \ .$$

The next result combines ideas from the analysis in [6], [12] which will show that constrained l^τ minimization recovers an approximation with error proportional to $1/a$:

Theorem III.2. *Let A be an $m \times N$ matrix, $0 < \tau \le 1$, and let $x \in \Sigma_k^N, w \in \mathbb{R}^m$ satisfy $\|Ax - w\|_2 \le \epsilon$. Define $\rho := k/J$ and $\gamma := \frac{b_J(A)}{a_{k+J}(A)}$. If $\gamma\rho^{1/\tau - 1/2} < 1$ holds, then the minimizer*

$$x^\sharp := \operatorname*{Argmin}_{\|Az - w\|_2 \le \epsilon} \|z\|_\tau$$

satisfies the bound

$$\|x^\sharp - x\|_2 \le \frac{\sqrt{1 + \frac{1}{2/\tau - 1}\left(\frac{k}{k+J}\right)^{2/\tau - 1}}}{1 - \gamma\rho^{1/\tau - 1/2}} \cdot \frac{2\epsilon}{a_{k+J}(A)} \ .$$

Proof: Define $\eta := x^\sharp - x$, and T to be the support of x. Using Hölder's inequality and the fact that $x^{\tau,\mu}$ is the l^τ minimizer (see (25) of [12]),

$$\|\eta_{T^c}\|_\tau \le \|\eta_T\|_\tau \le k^{1/\tau - 1/2}\|\eta_T\|_2 \ . \tag{7}$$

Block η_{T^c} into disjoint blocks of size J of decreasing magnitudes, i.e. $\eta_{T^c} = \sum_{i=1}^{L} \eta_{T_i}$ with $|T_i| = J$ and $|\eta_{T_i}(j)| \leq |\eta_{T_{i-1}}(j')|$ for $j \in T_i$, $j' \in T_{i-1}$ and $i > 1$. Using the constraint and singular value conditions,

$$\begin{aligned}\|\eta_{T\cup T_1}\|_2 &\leq \frac{1}{a_{k+J}(A)}\|A\eta_{T\cup T_1}\|_2 \\ &\leq \frac{1}{a_{k+J}(A)}\left(\|A\eta\|_2 + \sum_{i=2}^{L}\|A\eta_{T_i}\|_2\right) \\ &\leq \frac{2\epsilon}{a_{k+J}(A)} + \gamma\sum_{i=2}^{L}\|\eta_{T_i}\|_2 \ . \end{aligned} \tag{8}$$

Using 4.2.II of [12], bound $\|\eta_{T_i}\|_2 \leq J^{\frac{1}{2}-\frac{1}{\tau}}\|\eta_{T_{i-1}}\|_\tau$. Combined with the reversed triangle inequality (for $\tau < 1$ and non-negative vectors), we have $\sum_{i=2}^{L}\|\eta_{T_i}\|_2 \leq J^{\frac{1}{2}-\frac{1}{\tau}}\|\eta_{T^c}\|_\tau$. Finally using (7),

$$\sum_{i=2}^{L}\|\eta_{T_i}\|_2 \leq \rho^{1/\tau-1/2}\|\eta_{T\cup T_1}\|_2 \ . \tag{9}$$

Combining equation (30) of [12] with (7) gives

$$\|\eta\|_2 \leq \sqrt{1 + \frac{1}{2/\tau - 1}\rho^{2/\tau-1}}\|\eta_{T\cup T_1}\|_2 \ .$$

The result follows from substituting (9) into (8), solving for $\|\eta_{T\cup T_1}\|_2$ and substituting into the last equation. ■

IV. l^τ MINIMIZATION WITH $\Sigma\Delta$ AND COMPRESSED SENSING

Finally we put together our two main observations and state the known bounds and conditions for recovery. We state precisely the results concerning the singular value of submatrices of $D^{-r}\Phi$ to use with Theorem III.2. We already know that Theorem II.1 covers the smallest singular values of the submatrices. For the largest singular values, we can first use Gershgorin's circle theorem for eigenvalues on $(D^{-1})^T D^{-1}$, to show that $\sigma_{\max}(D^{-1}) \leq \sqrt{m + \frac{(m-1)m}{2}} \leq m$. Then using the bound $\sigma_{\max}(AB) \leq \sigma_{\max}(A)\sigma_{\max}(B)$,

$$\sigma_{\max}(D^{-r}E) \leq m^r\sigma_{\max}(E) \ . \tag{10}$$

The standard restricted isometry property allows us to bound the largest singular value of submatrices of Φ, which has a simple proof in [1]:

Theorem IV.1 (e.g. Theorem 5.2 of [1])**.** *Let Φ be an $m \times N$ matrix whose entries are i.i.d. $\mathcal{N}(0,1)$, and suppose*

$$\frac{m}{s} \geq C_6 \log(N/s)$$

for some absolute constant C_6. Then there exists an absolute constant C_7 such that $b_s(\Phi) < 2\sqrt{m}$ with probability $\geq 1 - 2e^{-C_7 m}$.

We can now combine these two results to obtain upper and lower singular value bounds:

Corollary IV.2. *Let Φ be an $m \times N$ random matrix whose entries are i.i.d $\mathcal{N}(0,1)$. Let $0 < \alpha < 1$ and suppose that*

$$\frac{m}{s} \geq C_8(\log N)^{\frac{1}{1-\alpha}} \tag{11}$$

for $C_8 = \max(C_3, C_6)$ from Theorems II.1 and IV.1. Then for some constant C_9 depending on r, with probability $\geq 1 - 3\exp(-C_9 m^{1-\alpha}s^\alpha)$ the following holds: For all $z \in \Sigma_s^N$,

$$C_5\sqrt{m}(m/s)^{\alpha(r-1/2)}\|z\|_2 \leq \|D^{-r}\Phi_T z\|_2 \leq 2m^{r+1/2}\|z\|_2$$

where C_5 is from Theorem II.1. In other words,

$$\begin{aligned} a_s(D^{-r}\Phi) &\geq C_5\sqrt{m}(m/s)^{\alpha(r-1/2)} \\ b_s(D^{-r}\Phi) &\leq 2m^{r+1/2} \ . \end{aligned}$$

Proof: Note that the condition (11) implies that the conditions for both Theorems $II.1$ and $IV.1$ are satisfied. Then with the union bound, both conclusions hold with probability $\geq 1 - \exp(-C_4 m^{1-\alpha}s^\alpha) - 2\exp(-C_7 m)$. Since $m \geq m^{1-\alpha}s^\alpha$, we can bound this by $\geq 1 - 3\exp(-C_9 m^{1-\alpha}s^\alpha)$ with $C_9 = \min(C_4, C_7)$. The conclusion of Theorem II.1 gives the lower inequality (a_s), and the conclusion of Theorem IV.1 along with observation (10) gives the upper inequality (b_s). ■

The following result is then immediate from Theorem III.2 and Corollary IV.2 using $A = D^{-r}\Phi$, $w = D^{-r}q$, $J = 2k$ and $\epsilon = \sqrt{m}\mu$:

Theorem IV.3. *Let Φ be an $m \times N$ matrix whose entries are i.i.d $\mathcal{N}(0,1)$, and let $0 < \alpha < 1$. Suppose for k and $0 < \tau \leq 1$, the following conditions are satisfied:*

i. $$\frac{m}{k} \geq 3C_8(\log N)^{\frac{1}{1-\alpha}}$$

ii. $$\frac{1}{\tau} > \frac{1}{2} + \log_2(2/C_5) + r\log_2 m$$

Then with probability $\geq 1 - \exp(-3^\alpha C_9 m^{1-\alpha}k^\alpha)$, the following holds:
For every $x \in \Sigma_k^N$, if r-th order $\Sigma\Delta$ is used to quantize Φx, with q being the quantization and $\|u\|_\infty < \mu$ in the corresponding difference equation, then the minimizer $x^{\tau,\mu}$ of (3) satisfies the bound

$$\|x^{\tau,\mu} - x\|_2 \leq C_{10}\mu\left(\frac{m}{k}\right)^{-\alpha(r-1/2)}$$

for some r-dependent constant C_{10}.

Remark IV.4. Note that for any fixed δ_Q and order r, there is a τ sufficiently small for which the conditions for recovery hold, and in the recovery error μ will typically have a linear dependence on δ_Q.

V. ROOT-EXPONENTIAL ACCURACY

Suppose we impose a bit budget of R bits for quantizing measurements from unit-norm vectors in Σ_k^N. Define $R_{\text{eff}} := R/k$, the effective bit-rate per sparse dimension. We will also work with a *fixed* quantization alphabet $\mathcal{A}$ of spacing δ_Q. This requires that we use the quantization method (II.B) and that the measurements be bounded independently from the number

of measurements. Unfortunately, Gaussian measurements do not satisfy this criteria [8], but there is also ongoing work that would allow us to use alternative matrix ensembles which are bounded, such as the Bernoulli-$\{\pm 1\}$ matrices [10]. For what follows we will assume the bound $\|y\|_\infty \leq M$ for some absolute constant M.

Also, by inspecting the proofs in [8] we can expand the r-dependent constants in the paper so that in Theorem IV.3, C_8 does not actually depend on r and $C_{10} \leq (C_{11}r)^r$ where C_{11} is now an absolute constant. Substituting (5) for μ in the conclusion of Theorem IV.3 gives a reconstruction accuracy of

$$\delta_Q (C_{12} r^2)^r (3k/m)^{\alpha(r-1/2)}$$

with $C_{12} = C_2 C_{11}$, and the number of bits needed for quantization is $R_{\text{eff}} = C_{13}\frac{m}{k}$ with $C_{13} = \log_2(C_1 + 2M/\delta_Q)$ from (4). Solving for m/k in the rate and substituting, the accuracy becomes

$$\delta_Q (R_{\text{eff}}/C_{13})^{\alpha/2} (C_{14} r^2 / R_{\text{eff}}^\alpha)^r$$

with $C_{14} = C_{12}(3C_{13})^\alpha$. Then optimizing over r, or choosing $r = \sqrt{R_{\text{eff}}^\alpha/(eC_{14})}$ gives

$$\delta_Q (R_{\text{eff}}/C_{13})^{\alpha/2} \exp(-C_{15} R_{\text{eff}}^{\alpha/2})$$

with $C_{15} = 1/\sqrt{eC_{14}}$.

VI. Algorithms

Solving the constrained l^τ minimization problem (3) is tricky given the non-convexity of the $\|\cdot\|_\tau$, but there are several approaches. In [11], Saab et al. use a modification to iterative reweighted least squares with encouraging numerical results. If we want a weight that encourages minimization of the sparsity measure $\|x\|_0$ instead, [13] mentions a weighting scheme that is non-separable which could potentially be used in this situation. Other approaches involve projected gradient, and different regularizations of the l^τ norm [4].

We conclude with a sample plot from the approach of [11], which uses the iteration

$$w_i^{(n)} = (|\hat{x}_i^{(n)}|^2 + \epsilon_w)^{\tau/2-1}$$
$$\hat{x}^{(n+1)} = W^{-1}A'(AW^{-1}A' + \lambda I)^{-1} D^{-\tau} q$$

where $W = W^{(n)}$ is diagonal with entries $w_i^{(n)}$, and $w^{(0)} \equiv 1$. Fixing $\epsilon_w = 10^{-10}$ and $\lambda = 1$, we start with $\tau = 1$ and decrease τ to 0.1. With $N = 200$ and $k = 3$, we generate a k-sparse signal and a 180×200 Bernoulli random matrix. For a range of m, we take the first m measurements, quantize and recover, recording the resulting error. In figure 1 we plot the result, comparing the iterative method with l^1 minimization and with the Sobolev dual (assuming a support oracle). What we observe in many cases is that after a certain number of measurements, the error starts tracking that of the Sobolev dual. In fact, if $w_i^{(n)} \to \infty$ for $i \notin \text{supp}(x)$, $\hat{x}^{(n+1)}$ converges to a small perturbation of the Sobolev dual reconstruction. Thus the success of the method hinges on a reweighting scheme that can detect the support of the source signal. We emphasize that with such a coarse quantization step size, l^1 minimization generally will fail to detect the support, a crucial requirement for the results in [8].

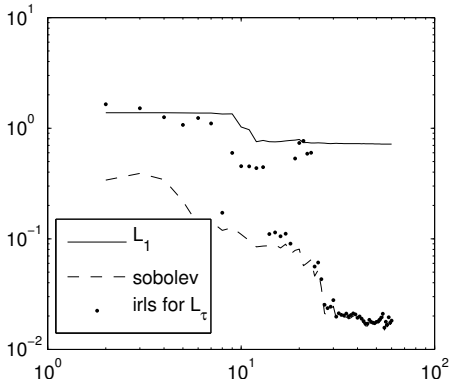


Fig. 1. Log-log plot comparing accuracy vs oversampling ratio m/k for a fixed k-sparse signal and Bernoulli measurements for $r = 1$ and $\delta_Q = 2$. In this example, $N = 200$, and $k = 3$.

Acknowledgement

The author would like to thank Sinan Güntürk and the reviewers for the valuable discussion and input for this paper.

References

[1] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 28(3):253–263, 2008.

[2] J. Blum, M. Lammers, A.M. Powell, and Ö. Yılmaz. Sobolev duals in frame theory and sigma-delta quantization. *Journal of Fourier Analysis and Applications*, 16(3):365–381, 2010.

[3] E.J. Candes, J.K. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on pure and applied mathematics*, 59(8):1207–1223, 2006.

[4] R. Chartrand. Fast algorithms for nonconvex compressive sensing: Mri reconstruction from very few data. In *Biomedical Imaging: From Nano to Macro, 2009. ISBI '09. IEEE International Symposium on*, pages 262 –265, 28 2009-july 1 2009.

[5] P. Deift, F. Krahmer, and C.S. Güntürk. An optimal family of exponentially accurate one-bit sigma-delta quantization schemes. *Communications on Pure and Applied Mathematics*, 64(7):883–919, 2011.

[6] S. Foucart and M.J. Lai. Sparsest solutions of underdetermined linear systems via lq-minimization for $0 < q \leq 1$. *Applied and Computational Harmonic Analysis*, 26(3):395–407, 2009.

[7] C.S. Güntürk. One-bit sigma-delta quantization with exponential accuracy. *Communications on Pure and Applied Mathematics*, 56(11):1608–1630, 2003.

[8] S. Güntürk, A. Powell, R. Saab, and Ö. Yılmaz. Sobolev duals for random frames and sigma-delta quantization of compressed sensing measurements. *FoCM*, 13(1):1–36, 2013.

[9] F. Krahmer, R. Saab, and R. Ward. Root-exponential accuracy for coarse quantization of finite frame expansions. *Information Theory, IEEE Transactions on*, 58(2):1069–1079, 2012.

[10] F. Krahmer, R. Saab, and O. Yilmaz. Personal communication, 2013.

[11] R. Saab, R. Chartrand, and O. Yilmaz. Stable sparse approximations via nonconvex optimization. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 3885–3888. IEEE, 2008.

[12] R. Saab and Ö. Yılmaz. Sparse recovery by non-convex optimization–instance optimality. *Applied and Computational Harmonic Analysis*, 29(1):30–48, 2010.

[13] D. Wipf, S. Nagarajan, et al. Solving sparse linear inverse problems: Analysis of reweighted l1 and l2 methods. In *SPARS-Signal Processing with Adaptive Sparse Structured Representations*, 2009.

Quantized Iterative Hard Thresholding:
Bridging 1-bit and High-Resolution Quantized Compressed Sensing

Laurent Jacques*, Kévin Degraux and Christophe De Vleeschouwer*
ICTEAM Institute, ELEN Department, Université catholique de Louvain (UCL)

***Abstract*—In this work, we show that reconstructing a sparse signal from quantized compressive measurement can be achieved in an unified formalism whatever the (scalar) quantization resolution, *i.e.*, from 1-bit to high resolution assumption. This is achieved by generalizing the *iterative hard thresholding* (IHT) algorithm and its binary variant (BIHT) introduced in previous works to enforce the consistency of the reconstructed signal with respect to the quantization model. The performance of this algorithm, simply called *quantized* IHT (QIHT), is evaluated in comparison with other approaches (*e.g.*, IHT, *basis pursuit denoise*) for several quantization scenarios.**

I. Introduction

Since the advent of Compressed Sensing (CS) almost 10 years ago [1, 2], many works have treated the problem of inserting this theory into an appropriate quantization scheme. This step is indeed mandatory for transmitting, storing and even processing any compressively acquired information, and more generally for sustaining the embedding of the CS principle in sensor design.

In its most popular version, CS provides uniform theoretical guarantees for stably recovering any sparse (or compressible) signal at a sensing rate proportional to the signal intrinsic dimension (*i.e.*, its *sparsity* level) [1, 2]. In this context, scalar quantization of compressive measurements has been considered along two main directions.

First, under a high-resolution quantization assumption, *i.e.*, when the number of bits allocated to encode each measurement is high, the quantization impact is often modeled as a mere additive Gaussian noise whose variance is adjusted to the quantization ℓ_2-distortion [3]. In short, under this high-rate model, the CS stability guarantees under additive Gaussian noise, *i.e.*, as derived from the $\ell_2 - \ell_1$ instance optimality [2], are used to bound the reconstruction error obtained from quantized observations. Variants of these works handle quantization saturation [4], prequantization noise [5], ℓ_p-distortion models ($p \geq 2$) for improved reconstruction in oversampled regimes [6, 7], optimize the high-resolution quantization procedure [8] or integrate more evolved $\Sigma\Delta$-quantization models departing from scalar PCM quantization [9].

Second, and more recently, extreme 1-bit quantization recording only the sign of the compressive measurement, *i.e.*, an information encoded in a single bit, has been considered [10–13]. New guarantees have been developed to tackle the non-linear nature of the sign operation thanks to the replacement of the *restricted isometric property* (RIP) by the quasi-isometric *binary ϵ-stable embedding* (BϵSE) [11], or to more general characterization of the binary embedding of sets based on their Gaussian Mean Width [12, 13]. In this context, iterative methods such as the *binary iterative hard thresholding* [11] or linear programming optimization [12] have been introduced for estimating the 1-bit sensed signal.

This work proposes a general procedure for handling the reconstruction of sparse signals observed according to a standard non-uniform scalar quantization of the compressive measurements. The novelty of this scheme is its ability to handle any resolution level, from 1-bit to high-resolution, in a progressive fashion. Conversely to the Bayesian approach of [16], our method relies on a generalization of the *iterative hard thresholding* (IHT) [17] that we simply called *quantized iterative hard thresholding*. Actually, QIHT reduces to BIHT for 1-bit sensing and it converges to IHT at high resolution.

Conventions: Most of domain dimensions (*e.g.*, M, N) are denoted by capital roman letters. Vectors and matrices are associated to bold symbols while lowercase light letters are associated to scalar values. The i^{th} component of a vector $\boldsymbol{u}$ is u_i or $(\boldsymbol{u})_i$. The identity matrix is $\mathbf{Id}$. The set of indices in $\mathbb{R}^D$ is $[D] = \{1, \cdots, D\}$. Scalar product between two vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^D$ reads $\boldsymbol{u}^*\boldsymbol{v} = \langle \boldsymbol{u}, \boldsymbol{v} \rangle$ (using the transposition $(\cdot)^*$), while the Hadamard product $\boldsymbol{u} \odot \boldsymbol{v}$ is such that $(\boldsymbol{u} \odot \boldsymbol{v})_i = u_i v_i$. For any $p \geq 1$, $\|\cdot\|_p$ represents the ℓ_p-norm such that $\|\boldsymbol{u}\|_p^p = \sum_i |u_i|^p$ with $\|\boldsymbol{u}\| = \|\boldsymbol{u}\|_2$ and $\|\boldsymbol{u}\|_\infty = \max_i |u_i|$. The ℓ_0 "norm" is $\|\boldsymbol{u}\|_0 = \#\operatorname{supp} \boldsymbol{u}$, where $\#$ is the cardinality operator and $\operatorname{supp} \boldsymbol{u} = \{i : u_i \neq 0\} \subseteq [D]$. For $\mathcal{S} \subseteq [D]$, $\boldsymbol{u}_\mathcal{S} \in \mathbb{R}^{\#\mathcal{S}}$ (or $\boldsymbol{\Phi}_\mathcal{S}$) denotes the vector (resp. the matrix) obtained by retaining the components (resp. columns) of $\boldsymbol{u} \in \mathbb{R}^D$ (resp. $\boldsymbol{\Phi} \in \mathbb{R}^{D' \times D}$) belonging to $\mathcal{S} \subseteq [D]$. The operator $\mathcal{H}_K$ is the hard thresholding operator setting all the coefficients of a vector to 0 but those having the K strongest amplitudes. The set of canonical K-sparse vectors in $\mathbb{R}^N$ is $\Sigma_K = \{\boldsymbol{v} \in \mathbb{R}^N : \|\boldsymbol{v}\|_0 \leq K\}$ while $\Sigma_\mathcal{T}$ denotes the set of vectors whose support is $\mathcal{T} \subseteq [N]$. Moreover, $\Sigma_K^* = \Sigma_K \cap S^{N-1}$ and $\Sigma_\mathcal{T}^* = \Sigma_\mathcal{T}^* \cap S^{N-1}$ with S^{N-1} the $(N-1)$-sphere in $\mathbb{R}^N$. Finally, $\chi_\mathcal{I}$ is the characteristic function on $\mathcal{I} \subset \mathbb{R}$, $\operatorname{sign} \lambda$ equals 1 if λ is positive and -1 otherwise, $(\lambda)_+ = (\lambda + |\lambda|)/2$ and $(\lambda)_- = -(-\lambda)_+$ project λ on $\mathbb{R}_+$ and $\mathbb{R}_-$, respectively, with all these operators being applied component wise onto vectors.

II. Noisy Compressed Sensing Framework

The *iterative hard thresholding* (IHT) algorithm has been introduced for iteratively reconstructing a sparse or compressible signal $\boldsymbol{x} \in \mathbb{R}^N$ from compressible observations $\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{x} + \boldsymbol{n}$, where $\boldsymbol{\Phi} \in \mathbb{R}^{M \times N}$ is the sensing matrix and $\boldsymbol{n} \in \mathbb{R}^M$ stands for a possible observational noise with bounded energy $\|\boldsymbol{n}\| \leq \varepsilon$. IHT is an alternative to the *basis pursuit denoise* (BPDN) method [18] which aims at solving a global convex

*LJ and CDV are funded by the Belgian F.R.S-FNRS. Part of this research is supported by the DETROIT project (WIST3), Walloon Region, Belgium. *Acknowledgements:* We thank Prasad Sudhakar (UCL/ICTEAM) and the anonymous reviewers of SAMPTA 2013 for their useful comments.

minimization promoting a ℓ_1-sparse data prior model under the constraint of reproducing the compressive observation.

Assuming that $\boldsymbol{x}$ is K-sparse in the canonical basis $\boldsymbol{\Psi} = \mathbf{Id}$, *i.e.*, $\boldsymbol{x} \in \Sigma_K$, the IHT algorithm is designed to approximately solve the (LASSO-type) problem

$$\min_{\boldsymbol{u}\in\mathbb{R}^N} \tfrac{1}{2}\|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{u}\|^2 \text{ s.t. } \|\boldsymbol{u}\|_0 \leq K. \tag{1}$$

It proceeds by computing the following recursion

$$\boldsymbol{x}^{(n+1)} = \mathcal{H}_K\big[\boldsymbol{x}^{(n)} + \mu\boldsymbol{\Phi}^*(\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x}^{(n)})\big], \tag{IHT}$$

where $\boldsymbol{x}^{(0)} = \mathbf{0}$, and $\mu > 0$ must satisfy $\mu^{-2} > \|\boldsymbol{\Phi}\| := \sup_{\boldsymbol{u}:\|\boldsymbol{u}\|=1} \|\boldsymbol{\Phi}\boldsymbol{u}\|$ for guaranteeing convergence [19].

In other words, at each iteration, starting from the previous estimation $\boldsymbol{x}^{(n)}$, the fidelity function $\mathcal{E}(\boldsymbol{u}) := \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{u}\|^2$ is decreased by a gradient descent step with gradient $\boldsymbol{\nabla}\mathcal{E}(\boldsymbol{x}^{(n)}) = \boldsymbol{\Phi}^*(\boldsymbol{\Phi}\boldsymbol{x}^{(n)} - \boldsymbol{y})$, followed by a "projection" on Σ_K accomplished by the hard thresholding $\mathcal{H}_K$.

In [17], it is shown that if $\boldsymbol{\Phi}$ respects the *restricted isometry property* (RIP) of order $3K$ with radius $\delta_{3K} < 1/15$, which means that for all $\boldsymbol{u} \in \Sigma_{3K}$, $(1-\delta_{3K})\|\boldsymbol{u}\|^2 \leq \|\boldsymbol{\Phi}\boldsymbol{u}\|^2 \leq (1+\delta_{3K})\|\boldsymbol{u}\|^2$, then, at iteration $n^* = \lceil \log_2 \|\boldsymbol{x}\|/\varepsilon \rceil$, the reconstruction error satisfies $\|\boldsymbol{x} - \boldsymbol{x}^{(n^*)}\| \leq 5\varepsilon$.

III. Quantized Sensing Model

For the sake of simplicity, let us consider a unit K-sparse signal $\boldsymbol{x}_0 \in \Sigma_K^*$ observed through the following Quantized Compressed Sensing (QCS) model

$$\boldsymbol{y} = \mathcal{Q}_b[\boldsymbol{\Phi}\boldsymbol{x}_0], \tag{2}$$

where $\boldsymbol{\Phi} \in \mathbb{R}^{M\times N}$ is the sensing matrix and $\mathcal{Q}_b$ the quantization operator defined at a *resolution* of b-bits per measurement, *i.e.*, with no further encoding treatment, $\boldsymbol{y}$ requires a total of $\mathfrak{B} = bM$ bits. In this work, we will not consider any prequantization noise in (2).

The quantization $\mathcal{Q}_b$ is assumed optimal with respect to the distribution of each component of $\boldsymbol{z} = \boldsymbol{\Phi}\boldsymbol{x}_0 \in \mathbb{R}^M$. In particular, by considering only random Gaussian matrices $\boldsymbol{\Phi} \sim \mathcal{N}^{M\times N}(0,1)$, *i.e.*, where each matrix entry follows $\Phi_{ij} \sim_{\text{iid}} \mathcal{N}(0,1)$, we have $z_i \sim \mathcal{N}(0, \|\boldsymbol{x}_0\|^2 = 1)$ and we adjust $\mathcal{Q}_b$ to an optimal b-bits Gaussian Quantizer minimizing the quantization distortion, *e.g.*, using a Lloyd-Max optimization [20]. This provides a set of thresholds $\{\tau_i \in \bar{\mathbb{R}} : 1 \leq i \leq 2^b + 1\}$ (with $-\tau_1 = \tau_{2^b+1} = +\infty$) defining 2^b quantization bins $\mathcal{R}_i = [\tau_i, \tau_{i+1})$, and a set of levels $\{q_i \in \mathcal{R}_i : 1 \leq i \leq 2^b\}$ such that

$$\mathcal{Q}_b[\lambda] = q_k \;\Leftrightarrow\; \lambda \in \mathcal{R}_k,$$

with $2\tau_i = q_{i-1} + q_i$ and $q_i = \mathbb{E}[g_x | g_x \in \mathcal{R}_i]$ with $g_x \sim \mathcal{N}(0,1)$. Notice that this QCS model includes 1-bit CS scheme since $\mathcal{Q}_1[\lambda] = q_0 \operatorname{sign}(\lambda)$ with $q_0 := q_2 = -q_1 = \sqrt{2/\pi}$.

IV. Quantized Iterative Hard Thresholding

In this section, we propose a generalization of the IHT algorithm taking into account the particular nature of the scalar quantization model introduced in Sec. III. The idea is to enforce the consistency of the iterates with the quantized observations. This is first achieved by defining an appropriate cost measuring deviation from quantization consistency.

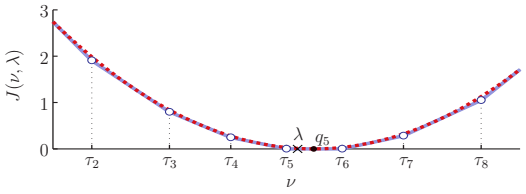


Fig. 1: (plain curve) Plot of J as a function of $\nu \in \mathbb{R}$ for $b = 3$ ($\tau_5 = 0$) and $\lambda \in \mathcal{R}_5$. (dashed curve) Plot of $\frac{1}{2}(\nu - q_5)^2$.

Given $\nu, \lambda \in \mathbb{R}$ and using the levels and thresholds associated to $\mathcal{Q}_b$, we first define

$$J(\nu,\lambda) = \sum_{j=2}^{2^b} w_j \left|\big(\operatorname{sign}(\lambda - \tau_j)(\nu - \tau_j)\big)_-\right|, \tag{3}$$

with $w_j = q_j - q_{j-1}$. Equivalently, given $\mathcal{I}(\nu,\lambda) := [\min(\nu,\lambda), \max(\nu,\lambda)]$, $J(\nu,\lambda) = \sum_{j=2}^{2^b} w_j \chi_{\mathcal{I}}(\tau_j)\,|\nu - \tau_j|$. The non-zero terms are therefore determined by the thresholds lying between λ and ν, *i.e.*, for which $\operatorname{sign}(\lambda - \tau_j) \neq \operatorname{sign}(\nu - \tau_j)$. Interestingly, $J(\nu;\lambda) = J(\nu;\mathcal{Q}_b(\lambda))$ since $\operatorname{sign}(\lambda - \tau_j) = \operatorname{sign}(\mathcal{Q}_b(\lambda) - \tau_j)$ for all $j \in [2^b + 1]$.

Then, our quantization consistency function between two vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^M$ reads

$$\mathcal{J}(\boldsymbol{u},\boldsymbol{v}) := \sum_{k=1}^{M} J(u_k, v_k) = \mathcal{J}(\boldsymbol{u}, \mathcal{Q}_b(\boldsymbol{v})). \tag{4}$$

This cost, which is convex with respect to $\boldsymbol{u}$, has two interesting limit cases. First, for $b = 1$, it reduces to the cost on which relies the *binary iterative hard thresholding* algorithm (BIHT) adapted to 1-bit CS [11]. In this context, the sum in (3) has only one term (for $j = 2$) and $\mathcal{J}(\boldsymbol{u},\boldsymbol{v}) = 2q_0\,\|(\operatorname{sign}(\boldsymbol{v}) \odot \boldsymbol{u})_-\|_1$. Up to a normalization by $2q_0$, this is the ℓ_1-sided norm minimized by BIHT which vanishes when $q_0 \operatorname{sign}(\boldsymbol{u}) = \mathcal{Q}_1(\boldsymbol{u}) = \mathcal{Q}_1(\boldsymbol{v}) = q_0 \operatorname{sign}(\boldsymbol{v})$, with q_0 defined in Sec. III.

Second, in the high resolution limit when $b \gg 1$, $\mathcal{J}(\boldsymbol{u},\boldsymbol{v})$ tends to $\frac{1}{2}\|\boldsymbol{u} - \boldsymbol{v}\|^2$. Indeed, in this case $w_j \ll 1$ and, the sum in (3) tends to

$$J(\nu,\lambda) \simeq \left|\textstyle\int_\nu^\lambda (\nu - t)\,\mathrm{d}t\right| = \tfrac{1}{2}(\nu - \lambda)^2.$$

This asymptotic quadratic behavior of J is illustrated in Fig. 1.

Given the quantization consistency cost $\mathcal{J}$, we can now formulate a generalization of (1) for estimating a K-sparse signal $\boldsymbol{x}_0$ observed by the model (2):

$$\min_{\boldsymbol{u}\in\mathbb{R}^N} \mathcal{E}_b(\boldsymbol{u}) \text{ s.t. } \|\boldsymbol{u}\|_0 \leq K, \tag{5}$$

with $\mathcal{E}_b(\boldsymbol{u}) := \mathcal{J}(\boldsymbol{\Phi}\boldsymbol{u}, \boldsymbol{y}) = \mathcal{J}(\boldsymbol{\Phi}\boldsymbol{u}, \mathcal{Q}_b[\boldsymbol{\Phi}\boldsymbol{x}_0])$.

Following the procedure determining the IHT algorithm from (1) (Sec. II), our aim is to find an IHT variant which minimizes the quantization inconsistency, as measured by $\mathcal{E}_b$, instead of the quadratic cost $\mathcal{E}$. This is done by first determining a *subgradient* of the convex but non-smooth function $\mathcal{E}_b$ [21].

A quick calculation shows that a subdifferential of $J(\nu,\lambda)$ with respect to ν reads

$$\sum_{j=k_-+1}^{k_+} \tfrac{w_j}{2}\big(\operatorname{sign}(\nu - \tau_j) - \operatorname{sign}(\lambda - \tau_j)\big), \tag{6}$$

where $k_- = \min(k_\nu, k_\lambda)$, $k_+ = \max(k_\nu, k_\lambda)$, and k_ν and k_λ are the bin indices of $\mathcal{Q}_b(\nu)$ and $\mathcal{Q}_b(\lambda)$ respectively. From the definition of the w_j, the sum simplifies to $q_{k_\nu} - q_{k_\lambda}$. Therefore, a subgradient of $\mathcal{J}(\boldsymbol{u}, \boldsymbol{v})$ with respect to $\boldsymbol{u}$ reads simply $\mathcal{Q}_b(\boldsymbol{u}) - \mathcal{Q}_b(\boldsymbol{v})$, so that a subgradient of $\mathcal{J}(\boldsymbol{\Phi u}, \boldsymbol{y})$ with respect to $\boldsymbol{u}$ corresponds to $\boldsymbol{\Phi}^*(\mathcal{Q}_b(\boldsymbol{\Phi u}) - \boldsymbol{y})$.

Therefore, from this last ingredient, we define the *quantized iterative hard thresholding algorithm* (QIHT) by the recursion

$$\boldsymbol{x}^{(n+1)} = \mathcal{H}_K\big[\boldsymbol{x}^{(n)} + \mu \boldsymbol{\Phi}^*\big(\boldsymbol{y} - \mathcal{Q}_b(\boldsymbol{\Phi x}^{(n)})\big)\big], \qquad \textbf{(QIHT)}$$

where $\boldsymbol{x}^{(0)} = 0$ and μ is set hereafter.

V. QIHT ANALYSIS

Despite successful simulations of sparse signal recovery from quantized measurements (see Sec. VI), we were not able to prove the stability and the convergence of the QIHT algorithm yet. However, there exist a certain number of promising properties suggesting the existence of such a result. The first one comes from a limit case analysis. Except for the normalizing factor μ, QIHT at 1-bit ($b = 1$) reduces to BIHT [11]. Moreover, when $b \gg 1$, $\mathcal{Q}_b[\boldsymbol{z}] \simeq \boldsymbol{z}$ for $\boldsymbol{z} \in \mathbb{R}^M$ and we recover the IHT algorithm. These limit cases are consistent with the previous observations made above on the asymptotic behaviors of $\mathcal{J}$ in these two cases.

Second, as for the modified Subspace Pursuit algorithm [3], QIHT is designed for improving the quantization consistency of the current iterate with the quantized observations. For the moment, the importance of this improvement can only be understood in 1-bit. Given $\delta > 0$, when $M = O(\delta^{-1} K \log N)$ and with high probability on the drawing of a random Gaussian matrix $\boldsymbol{\Phi} \sim \mathcal{N}^{M \times N}(0, 1)$, $\| \frac{\boldsymbol{a}}{\|\boldsymbol{a}\|} - \frac{\boldsymbol{b}}{\|\boldsymbol{b}\|} \| \leq \delta$ if $\mathcal{Q}_1(\boldsymbol{\Phi a}) = \mathcal{Q}_1(\boldsymbol{\Phi b})$ for all $\boldsymbol{a}, \boldsymbol{b} \in \Sigma_K$ [11]. Actually, it can be shown[1] that if no more than r components differ between $\mathcal{Q}_1(\boldsymbol{\Phi a})$ and $\mathcal{Q}_1(\boldsymbol{\Phi b})$, then, with high probability on $\boldsymbol{\Phi}$,

$$\| \tfrac{\boldsymbol{a}}{\|\boldsymbol{a}\|} - \tfrac{\boldsymbol{b}}{\|\boldsymbol{b}\|} \| \leq (\tfrac{K+r}{K})\, \delta,$$

for $M = O(\delta^{-1} K \log MN)$. We understand then the beneficial impact of any increase of consistency between $\mathcal{Q}_1(\boldsymbol{\Phi x}^{(n)})$ and $\boldsymbol{y}$ at each QIHT iteration.

Third, the adjustment of μ, which is decisive for QIHT efficiency, leads also to some interesting observations. Extensive simulations not presented here pointed us that, for $\boldsymbol{\Phi} \sim \mathcal{N}^{M \times M}(0, 1)$, $\mu \propto 1/M$ seems to be a universal rule of efficiency at any bit rate. Interestingly, this setting was already characterized for IHT where $\mu \simeq 1/(1 + \delta_{2K})$ if the sensing matrix respects the RIP property with radius δ_{2K} [19]. Since $\boldsymbol{\Phi}/\sqrt{M}$ is RIP for $\boldsymbol{\Phi} \sim \mathcal{N}^{M \times N}(0, 1)$ as soon as $M = O(K \log N/K)$ this is equivalent to impose $\mu \simeq 1/M$.

At the other extreme, the rule $\mu \propto 1/M$ is also consistent with the following 1-bit analysis. In [13], it is shown that the mapping $\boldsymbol{u} \to \operatorname{sign}(\boldsymbol{\Phi u})$ respects an interesting property that we arbitrary call *sign product embedding*[2] (SPE):

Proposition 1. *Given $0 < \delta < 1$, there exist two constants $c, C > 0$ such that, if $M \geq C\delta^{-6} K \log N/K$, then, with a probability higher than $1 - 8\exp(-c\delta^2 M)$, $\boldsymbol{\Phi} \sim \mathcal{N}^{M \times N}(0, 1)$ satisfies*

$$\big|\mu^* \langle \operatorname{sign}(\boldsymbol{\Phi u}), \boldsymbol{\Phi v} \rangle - \langle \boldsymbol{u}, \boldsymbol{v} \rangle\big| \leq \delta, \quad \forall \boldsymbol{u}, \boldsymbol{v} \in \Sigma_K^*, \qquad (7)$$

with $\mu^* = 1/(q_0\, M)$. When $\boldsymbol{u}$ is fixed, the condition on M is relaxed to $M \geq C\delta^{-2} K \log N/K$.

When $\boldsymbol{\Phi}$ respects (7), we simply write that $\boldsymbol{\Phi}$ is SPE(Σ_K^*, δ). When $\boldsymbol{u}$ is fixed, we say that $\boldsymbol{\Phi}$ is locally SPE(Σ_K^*, δ) on $\boldsymbol{u}$. This SPE property leads to an interesting phenomenon.

Proposition 2. *Given $\boldsymbol{x} \in \Sigma_K^*$ and let $\boldsymbol{\Phi} \in \mathbb{R}^{M \times N}$ be a matrix respecting the local SPE(Σ_{2K}^*, δ) on $\boldsymbol{x}$ for some $0 < \delta < 1$. Then, given $\boldsymbol{y} = \mathcal{Q}_1[\boldsymbol{\Phi x}] = q_0 \operatorname{sign}(\boldsymbol{\Phi x})$, the vector*

$$\hat{\boldsymbol{x}} := \tfrac{1}{q_0^2 M} \mathcal{H}_K(\boldsymbol{\Phi}^* \boldsymbol{y}),$$

satisfies $\|\boldsymbol{x} - \hat{\boldsymbol{x}}\| \leq 2\delta$.

Proof: Let us define $\mathcal{T}_0 = \operatorname{supp} \boldsymbol{x}$, $\mathcal{T} = \mathcal{T}_0 \cup \operatorname{supp} \hat{\boldsymbol{x}}$, and $\boldsymbol{a} = \frac{1}{q_0^2 M} \boldsymbol{\Phi}^* \boldsymbol{y} = \mu^* \boldsymbol{\Phi}^* \operatorname{sign}(\boldsymbol{\Phi x})$ with $\hat{\boldsymbol{x}} = \mathcal{H}_K(\boldsymbol{a})$. Then $\hat{\boldsymbol{x}}$ is also the best K-term approximation $\boldsymbol{a}_\mathcal{T} = \boldsymbol{\Phi}_\mathcal{T}^* \boldsymbol{y}$, so that $\|\boldsymbol{x} - \hat{\boldsymbol{x}}\| \leq \|\boldsymbol{x} - \boldsymbol{a}_\mathcal{T}\| + \|\hat{\boldsymbol{x}} - \boldsymbol{a}_\mathcal{T}\| \leq 2\|\boldsymbol{x} - \boldsymbol{a}_\mathcal{T}\|$. Therefore, since $\|\boldsymbol{x} - \boldsymbol{a}_\mathcal{T}\| = \sup_{\boldsymbol{w} \in \Sigma_\mathcal{T}^*} \langle \boldsymbol{w}, \boldsymbol{x} - \boldsymbol{a}_\mathcal{T} \rangle$ and $\boldsymbol{\Phi}$ is SPE(Σ_{2K}^*, δ), $\|\boldsymbol{x} - \hat{\boldsymbol{x}}\| \leq 2 \sup_{\boldsymbol{w} \in \Sigma_\mathcal{T}^*} \big(\langle \boldsymbol{w}, \boldsymbol{x} \rangle - \mu^* \langle \boldsymbol{\Phi w}, \operatorname{sign}(\boldsymbol{\Phi x}) \rangle\big) \leq 2 \sup_{\boldsymbol{w} \in \Sigma_\mathcal{T}^*} \big(\langle \boldsymbol{w}, \boldsymbol{x} \rangle - \langle \boldsymbol{w}, \boldsymbol{x} \rangle + \delta\big) = 2\delta$, using $\operatorname{supp}(\boldsymbol{x} - \boldsymbol{a}_\mathcal{T}) \subseteq \mathcal{T}$ with $\#\mathcal{T} \leq 2K$. ∎

This proposition shows that a single hard thresholding of $\frac{1}{q_0^2 M} \boldsymbol{\Phi}^* \boldsymbol{y}$ already provides a good estimation of $\boldsymbol{x}$. Actually, from the condition on M for reaching the local SPE, we deduce that $\|\boldsymbol{x} - \hat{\boldsymbol{x}}\| = O(\sqrt{K/M})$. This is quite satisfactory for such a simple $\boldsymbol{x}$ estimation and it suggests setting $\mu \propto 1/M$ in QIHT for $b = 1$ where $\hat{\boldsymbol{x}}$ is related to $\boldsymbol{x}^{(1)}$.

Noticeably, it has been recently observed in [14] that $\hat{\boldsymbol{x}}' := \hat{\boldsymbol{x}}/\|\hat{\boldsymbol{x}}\|$ is actually solution of $\operatorname{argmax}_{\boldsymbol{u}} \langle \boldsymbol{y}, \boldsymbol{\Phi u} \rangle$ s.t. $\|\boldsymbol{u}\|_0 \leq K$, for which there exists the weaker error bound $\|\boldsymbol{x} - \hat{\boldsymbol{x}}'\|^2 = O(\sqrt{K/M})$ when $\boldsymbol{x}$ is fixed [13].

VI. EXPERIMENTS

An extensive set of simulations has been designed for evaluating the efficiency of QIHT in comparison with two other methods more suited to high-resolution quantization, namely, IHT and BPDN. Our objective is to show that QIHT provides better quality results at least at small quantization levels. For all experiments, we set $N = 1024$, $K = 16$ and the K-sparse signals were generated by choosing their supports uniformly at random amongst the $\binom{N}{K}$ available ones, while their non-zero coefficients were drawn uniformly at random on the sphere $S^{K-1} \subseteq \mathbb{R}^K$. For each algorithm, 100 initial such sparse vectors were generated and the reconstruction method was tested for $1 \leq b \leq 5$ and for $\mathfrak{B} = bM \in \{64, 128, \cdots, 1280\}$, *i.e.*, approximately fixing $M = \lfloor \mathfrak{B}/b \rfloor$. For each experimental condition, the quantized M-dimensional measurement vectors $\boldsymbol{y}_b$ was generated as in (2) with a random sensing matrix $\boldsymbol{\Phi} \sim \mathcal{N}^{M \times N}(0, 1)$ and according to an optimal Lloyd-Max b-bits Quantizer $\mathcal{Q}_b$ (Sec. III). IHT and QIHT iterations were both stopped at step n as soon as $\|\boldsymbol{x}^{(n+1)} - \boldsymbol{x}^{(n)}\| \|\boldsymbol{x}^{(n+1)}\|^{-1} < 10^{-4}$ or if $n = 1000$. The BPDN algorithm was solved with the SPGL1

[1] The interested reader can find the proof in a related technical report [15].
[2] In [13], more general embeddings than this of Σ_K are studied.

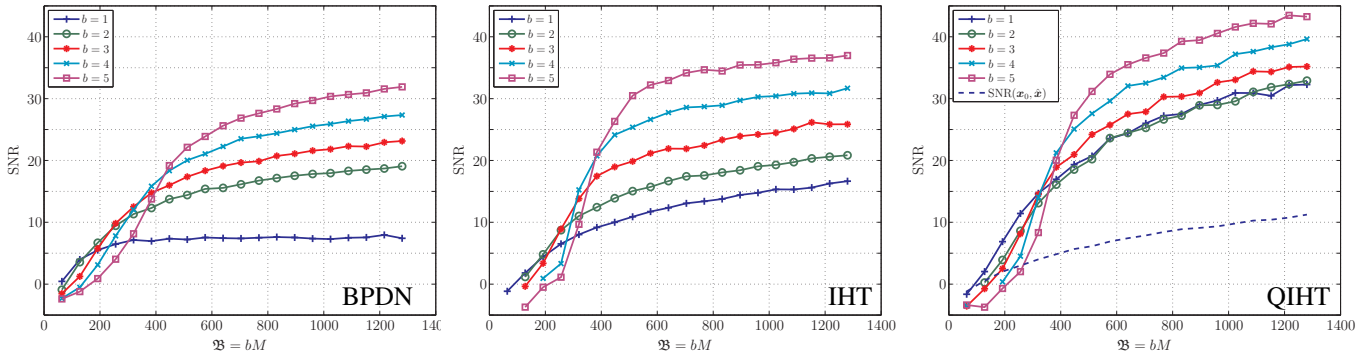


Fig. 2: Comparison between (from left to right) BPDN, IHT and QIHT for several quantization scenarios. The SNR is expressed in dB as a function of the bit budget $\mathfrak{B}$ and the number of bits b used to quantize each measurement.

MATLAB toolbox [22]. In IHT and QIHT, signal sparsity K was assumed known and both were set with $\mu = \frac{1}{M}\big(1 - \sqrt{\frac{2K}{M}}\big)$. This fits the IHT condition $\mu < 1/(1+\delta_{2K})$ mentioned in Sec. V by assuming that the RIP radius δ_{2K} behaves like $\sqrt{2K/M}$, which is a common assumption in CS. For BPDN, the noise energy was given by an oracle installing BPDN in the best reconstruction scenario, *i.e.*, $\epsilon = \|\boldsymbol{\Phi}\boldsymbol{x}_0 - \boldsymbol{y}\|_2$. Whatever the reconstruction method, given an initial signal $\boldsymbol{x}_0 \in \Sigma_K^*$ and its reconstruction $\boldsymbol{x}^*$, the reconstruction quality was measured by $\mathrm{SNR}(\boldsymbol{x}_0, \boldsymbol{x}^*) = -20\log_{10}\big\|\boldsymbol{x}_0 - \|\boldsymbol{x}^*\|^{-1}\boldsymbol{x}^*\big\|$. In other words, we focus here on a good "angular" estimation of the signals, adopting therefore a common metric for $b > 1$ and for $b = 1$, where amplitude information is lost. Finally, for each method and each couple of (M, b), the SNR was averaged over the 100 test signals and expressed in dB.

Fig. 2 gathers the SNR performances of the 3 methods as a function of $\mathfrak{B}$. QIHT outperforms both BPDN and IHT for the selected scenarios, especially for low bit quantizers. At high resolution, the gain between QIHT and IHT decreases as expected from the limit case analysis of QIHT. We can also notice that, first, there is almost no quality difference between QIHT at $b = 1$ and $b = 2$. This could be due to a non-optimality of the Lloyd-Max quantizer with respect to QIHT reconstruction error minimization. Second, BPDN and IHT asymptotically present the "6dB per bit" gain, while QIHT hardly exhibits such behavior only when $b = 4 \rightarrow 5$.

Finally, in order to test Prop. 2, the SNR reached by the single thresholding solution $\hat{\boldsymbol{x}}$ is plotted in dashed in Fig 2-right. Despite its poor behavior compared to QIHT at $b = 1$, it outperforms BPDN at high $\mathfrak{B} = M$ with a $\mathrm{SNR} \geq 10$dB at $M = N = 1024$. A curve fitting (no shown here) shows that this SNR increases a bit faster than $20\log_{10}\sqrt{K/M} + O(1)$.

VII. Conclusion

We have introduced the QIHT algorithm as a generalization of the BIHT and IHT algorithms aiming at enforcing consistency with quantized observations at any bit resolution. In particular, we showed that the almost obvious inclusion of the quantization operator in the IHT recursion is actually related to the implicit minimization of a particular inconsistency cost $\mathcal{E}_b$. This function generalizes the one-sided ℓ_1 cost of BIHT and asymptotically converges to the quadratic fidelity minimized by IHT. There is still a hard work to be performed in order to prove QIHT convergence and stability. However, the different ingredients defining it, as $\mathcal{E}_b$, deserve independent analysis extending previous 1-bit embeddings developed in [11–13].

References

[1] D. L. Donoho, "Compressed Sensing," *IEEE Trans. Inf. Th.*, **52**(4):1289–1306, 2006.

[2] E. J. Candès, J. Romberg and T. Tao. "Stable signal recovery from incomplete and inaccurate measurements," *Comm. Pure Appl. Math.*, **59**(8):1207–1223, 2006

[3] W. Dai, H. V. Pham, and O. Milenkovic, "Information theoretical and algorithmic approaches to quantized compressive sensing," *IEEE Trans. Comm.*, **59**(7):1857–1866, 2011.

[4] J. Laska, P. Boufounos, M. Davenport, and R. Baraniuk, "Democracy in action: Quantization, saturation, and compressive sensing," *App. Comp. Harm. Anal.*, **31**(3): 429–443, 2011.

[5] A. Zymnis, S. Boyd, and E. Candès, "Compressed sensing with quantized measurements," *IEEE Sig. Proc. Let.*, **17**(2): 149–152, 2010.

[6] L. Jacques, D. K. Hammond, and M. J. Fadili, "Dequantizing Compressed Sensing: When Oversampling and Non-Gaussian Constraints Combine.," *IEEE Trans. Inf. Th.*, **57**(1): 559–571, 2011.

[7] L. Jacques, D. K. Hammond, and M. J. Fadili, "Stabilizing nonuniformly quantized compressed sensing with scalar companders," *arXiv:1206.6003*, 2012.

[8] J. Z. Sun and V. K. Goyal, "Optimal quantization of random measurements in compressed sensing," in *Int. Symp. Inf. Th. (ISIT)*, June 2009.

[9] C. Sinan Güntürk, M Lammers, A. M. Powell, R. Saab, and Ö. Yılmaz, "Sobolev duals for random frames and $\Sigma\Delta$ quantization of compressed sensing measurements," *Found. Comp. Math.*, **13**(1):1–36, 2013.

[10] P. Boufounos and R. Baraniuk, "1-bit compressive sensing," in *Proc. Conf. Inform. Sc. Sys. (CISS)*, Princeton, NJ, Mar. 2008.

[11] L. Jacques, J.N. Laska, P.T. Boufounos, and R.G. Baraniuk, "Robust 1-bit compressive sensing via binary stable embeddings of sparse vectors," *IEEE Trans. Inf. Th.*, **59**(4):2082-2102, 2013.

[12] Y. Plan and R. Vershynin, "One-bit compressed sensing by linear programming," *Comm. Pure Appl. Math.*, 2013.

[13] Y. Plan and R. Vershynin, "Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach," *IEEE Trans. Inf. Th.*, **59**(1):482–494, 2013.

[14] S. Bahmani, P. T. Boufounos and B. Raj, "Robust 1-bit Compressive Sensing via Gradient Support Pursuit," *arXiv:1304.6627*, 2013.

[15] L. Jacques, K. Degraux, C. De Vleeschouwer, "Quantized Iterative Hard Thresholding: Bridging 1-bit and High-Resolution Quantized Compressed Sensing" TR-LJ-2013-02, *arXiv:1305.1786*, 2013

[16] Z. Yang, L. Xie, and C. Zhang, "Unified framework and algorithm for quantized compressed sensing," *arXiv:1203.4870*, 2012.

[17] T. Blumensath and M.E. Davies, "Iterative hard thresholding for compressed sensing," *App. Comp. Harm. Anal.*, **27**(3):265–274, 2009.

[18] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic Decomposition by Basis Pursuit," *SIAM J. Sc. Comp.*, **20**(1):33–61, 1998.

[19] T. Blumensath, "Accelerated iterative hard thresholding," *Sig. Proc.*, **92**(3):752–756, 2012.

[20] R. M. Gray and D. L. Neuhoff, "Quantization," *IEEE Trans. Inf. Th.*, **44**(6):2325–2383, 1998.

[21] R.T. Rockafellar, *Convex analysis*, vol. 28, Princeton Univ. Press, 1970.

[22] E. van den Berg and M. P. Friedlander, "SPGL1: A solver for large-scale sparse reconstruction," June 2007, http://www.cs.ubc.ca/labs/scl/spgl1.

Sigma-Delta quantization of sub-Gaussian compressed sensing measurements

Felix Krahmer
Institute for Numerical and Applied Mathematics
University of Göttingen
Lotzestraße 16-18
37085 Göttingen, Germany
Email: f.krahmer@math.uni-goettingen.de

Rayan Saab
Department of Mathematics
Duke University
Durham, NC 27708, USA
Email: rayans@math.duke.edu

Özgür Yılmaz
Department of Mathematics
University of British Columbia
1984 Mathematics Road
Vancouver, BC V6T 1Z2, Canada
Email: oyilmaz@math.ubc.ca

***Abstract*—Recently, it has been shown that for the setup of compressed sensing with Gaussian measurements that $\Sigma\Delta$ quantization can be effectively incorporated into the sensing mechanism [1]. In contrast to independently quantized measurements, the resulting schemes yield better reconstruction accuracy with a higher number of measurements even at a constant number of bits per signal. The original analysis of this method, however, crucially depends on the rotation invariance of the Gaussian measurements and hence does not directly generalize to other classes of measurements. In this note, we present a refined analysis that allows for a generalization to arbitrary sub-Gaussian measurements.**

I. Introduction

Compressed Sensing [2], [3] is a recent paradigm in signal processing based on the observation that many natural signals are approximately sparse in suitable representation systems, that is, they have only few significant coefficients. The underlying idea is that such signals are intrinsically low-dimensional, so the number of linear measurements necessary to allow for recovery of the signal should be considerably smaller than the signal dimension. Here taking m linear measurements of a signal $x \in \mathbb{R}^N$ is to be understood as considering the measurement vector $y = Ax$, where $A \in \mathbb{R}^{m\times N}$ is a fixed measurement matrix. As it turns out, a number of measurements proportional to $s\log(N/s)$ can allow for stable and robust recovery of signals with s non-vanishing entries in dimension N, provided the measurement matrix is suitably chosen. As no deterministic constructions for such matrices are known, this choice typically involves a random matrix construction.

Note that the resulting linear system to be solved to recover the signal is underdetermined, so the regularizing assumption of sparsity is crucial. However, once it has been determined which s coefficients are significant, the system becomes redundant by at least a logarithmic factor.

In order for the signal to be processed digitally, the measurements must, in a second step, be *quantized*. That is, the measurement vector, whose entries can a priori take arbitrary real values, must be represented by a sequence of values from a given finite alphabet. At this stage, the redundancy mentioned above can be exploited by applying a Sigma-Delta quantization scheme. Such coarse quantization schemes, originally designed for quantizing oversampled bandlimited signals [4], translate redundancy in a signal representation to more accurate quantized representations even though the alphabet size representing each sample is fixed. The idea is that the quantized representations are chosen dynamically using a feedback loop such that the quantization error made in a given sample partly compensates for the error made in previous samples.

This idea has been transferred to the setup of quantizing frame representations in $\mathbb{R}^N$ in [5]. As it turned out in subsequent works, higher accuracy can be achieved if instead of the Moore-Penrose pseudoinverse of the frame matrix, the so called canonical dual frame, an alternative dual frame, the so-called *Sobolev dual* is used for reconstruction [6]. In compressed sensing, once the support is identified, the measurement vector is nothing but a frame representation of the signal, so similar ideas apply.

For measurement matrices with independent standard normal entries, this scenario has been analyzed in [1]. For recovery, the authors proceed via a two-stage approach. In a first stage, standard compressed sensing techniques are used to estimate the support of the signal from the quantized measurements. In a second step, once the support has been identified, a Sobolev dual is used to obtain a more precise estimate of the signal coefficients.

The analysis in [1] is specific to Gaussian measurements. In this note, we generalize their results to arbitrary sub-Gaussian measurements. This more general setup includes important examples like Bernoulli matrices.

II. Sigma-Delta quantization

A. Greedy quantization schemes

Denote by $\mathcal{A}$ the $2L$ level *mid-rise* alphabet

$$\mathcal{A} = \Big\{ \pm (2j+1)\delta/2, \quad j \in \{0, ..., L-1\} \Big\}$$

and let $Q : \mathbb{R} \mapsto \mathcal{A}$ denote the scalar quantizer, which is defined via its action

$$Q(x) = \arg\min_{q\in\mathcal{A}} |x - q|.$$

The rth order *greedy* $\Sigma\Delta$ quantization scheme, defined via

$$q_i = Q\Big(\sum_{j=1}^{r}(-1)^{j-1}\binom{r}{j}u_{i-j} + y_i\Big)$$
$$u_i = \sum_{j=1}^{r}(-1)^{j-1}\binom{r}{j}u_{i-j} + y_i - q_i, \qquad (1)$$

maps a sequence of inputs $(y_i)_{i=1}^m$ to a sequence $(q_i)_{i=1}^m$ whose elements take on values from $\mathcal{A}$. Note that condition (1) can be rewritten as

$$(\Delta^r u)_i = y_i - q_i,$$

where Δ is the finite difference operator.

It is easily seen by induction that for bounded input sequences $\|y\|_\infty < (L - 2^{r-1} - 3/2)$, such schemes satisfy

$$\|u\|_\infty \leq \delta/2.$$

In other words, the scheme is *stable*, that is, its state sequence remains bounded. Note that to satisfy this stability condition, the number of levels L must increase with r.

B. Sigma-Delta error analysis

If $y = Ex \in \mathbb{R}^m$ is a vector of frame coefficients that is $\Sigma\Delta$ quantized to yield the vector $q \in \mathcal{A}^m$, then linear reconstruction of x from q using some dual frame F of E (i.e., $FE = I$) produces the estimate $\hat{x} := Fq$. We would like to control the reconstruction error $\eta := x - \hat{x}$. Writing the state variable equations (II-A) in vector form, we have

$$D^r u = y - q,$$

where D is the $m \times m$ difference matrix with entries given by

$$D_{ij} = \begin{cases} 1 & i = j \\ -1 & i = j+1 \\ 0 & \text{otherwise} \end{cases}.$$

Thus,

$$\eta = x - Fq = F(y - q) = FD^r u.$$

Working with with stable $\Sigma\Delta$ schemes, one can control $\|u\|_2$ via $\|u\|_\infty$. Thus, it remains to bound the operator norm $\|FD^r\| := \|FD^r\|_{\ell_2^m \mapsto \ell_2^k}$ and a natural choice for F is

$$F := \arg\min_{G:GE=I} \|GD^r\| = (D^{-r}E)^\dagger D^{-r}. \qquad (2)$$

This so-called Sobolev dual frame was first proposed in [6]. Here $A^\dagger := (A^*A)^{-1}A^*$ is the $k \times m$ Moore-Penrose (left) inverse of the $m \times k$ matrix A. Since (2) implies that $FD^r = (D^{-r}E)^\dagger$, the singular values of $D^{-r}E$ will play a key role in this paper.

An important property of the matrix D is given in the following proposition .

Proposition 1 ([1], Proposition 3.1): There are constants c_1, c_2 depending only on r such that the singular values of the matrix D^{-r} satisfy

$$c_1(r)\left(\frac{m}{j}\right)^r \leq \sigma_j(D^{-r}) \leq c_2(r)\left(\frac{m}{j}\right)^r.$$

III. Preliminaries

Here and throughout, $x \sim \mathcal{D}$ denotes that the random variable x is drawn according to a distribution $\mathcal{D}$. Furthermore, $\mathcal{N}(0, \sigma^2)$ denotes the zero-mean Gaussian distribution with variance σ^2. The following definition provides a means to compare the tail decay of two distributions.

Definition 2: If two random variables $\eta \sim \mathcal{D}_1$ and $\xi \sim \mathcal{D}_2$ satisfy $P(|\eta| > t) \leq KP(|\xi| > t)$ for some constant K and all $t \geq 0$, then we say that η is K-dominated by ξ (or, alternatively, by $\mathcal{D}_2$).

Definition 3: A random variable is sub-Gaussian with parameter $c > 0$ if it is e-dominated by $\mathcal{N}(0, c^2)$.

Remark 4: One can also define sub-Gaussian random variables via their moments or, in case of zero mean, their moment generating functions. See [7] for a proof that all these definitions are equivalent.

Remark 5: Exmaples of sub-Gaussian random variables include Gaussian random variables, all bounded random variables (such as Bernoulli), and their linear combinations.

Definition 6: We say that a matrix E is sub-Gaussian with parameter c if its entries are independent sub-Gaussian random variables with mean zero, variance one, and parameter c.

IV. Main results

In this section, we present our main results, generalizing the theorems of [1] on the singular values of $D^{-r}E$ to sub-Gaussian matrix entries and leveraging these results to establish recovery guarantees from $\Sigma\Delta$ quantized compressed sensing measurements.

Proposition 7: Let E be an $m \times k$ sub-Gaussian matrix with parameter c, let $S = \mathrm{diag}(s)$ be a diagonal matrix, and let V be an orthonormal matrix, both of size $m \times m$. Further, let $r \in \mathbb{Z}^+$ and suppose that $s_j \geq C_1^r\left(\frac{m}{j}\right)^r$, where C_1 is a positive constant that may depend on r. Then there exist constants $C_2, C_3 > 0$ (depending on c and C_1) such that for $0 < \alpha < 1$ and $\lambda := \frac{m}{k} \geq C_2^{\frac{1}{1-\alpha}}$

$$\mathbb{P}\left(\sigma_{\min}(\tfrac{1}{\sqrt{m}}SV^*E) \leq \lambda^{\alpha(r-1/2)}\right) \leq 2\exp(-C_3 m^{1-\alpha}k^\alpha).$$

In particular, C_3 depends only on c, while C_2 can be expressed as $f(c)C_1^{-\frac{2r}{2r-1}}$ provided $C_1 \leq 1/2$.

Proof: The matrix SV^*E has dimensions m and k, so by the Courant min-max principle applied to the transpose one has

$$\sigma_{\min}(SV^*E) = \min_{\substack{W\subset\mathbb{R}^m \\ \dim W = m-k+1}} \sup_{z\in W:\|z\|_2=1} \|E^*VSz\|_2$$

Noting that, for $m \geq \widetilde{k} := C_4 m^{1-\alpha}k^\alpha > k$, where the constant C_4 will be determined later, each $m - k + 1$-dimensional subspace intersects the span $V_{\widetilde{k}}$ of the first $\widetilde{k}$ standard basis vectors in at least a $\widetilde{k} - k + 1$-dimensional

space, this expression is bounded from below by

$$\begin{aligned}\min_{\substack{W\subset V_{\widetilde{k}}\\ \dim W=\widetilde{k}-k+1}} \sup_{z\in W:\|z\|_2=1} \|E^*VSz\|_2 &\geq \min_{\substack{W\subset V_{\widetilde{k}}\\ \dim W=\widetilde{k}-k+1}} \sup_{z\in W:\|z\|_2=s_{\widetilde{k}}} \|E^*Vz\|_2\\ &= \min_{\substack{W\subset \mathbb{R}^{\widetilde{k}}\\ \dim W=\widetilde{k}-k+1}} \sup_{z\in W:\|z\|_2=1} s_{\widetilde{k}}\|E^*VP_{\widetilde{k}}^*z\|_2.\end{aligned} \tag{3}$$

The inequality follows from the observation that $V_{\widetilde{k}}$ is invariant under S and the smallest singular value of $S|_{V_{\widetilde{k}}}$ is $s_{\widetilde{k}}$. In the last step, $P_{\widetilde{k}}$ denotes the projection of an m-dimensional vector onto its first $\widetilde{k}$ components. We note that (3), again by the Courant min-max principle, is equal to

$$s_{\widetilde{k}}\sigma_k(E^*VP_{\widetilde{k}}^*) = s_{\widetilde{k}}\sigma_{min}(P_{\widetilde{k}}V^*E) = s_{\widetilde{k}}\inf_{y\in S^{k-1}}\|P_{\widetilde{k}}V^*Ey\|_2$$

Now, as $\mathbb{E}\|P_{\widetilde{k}}V^*Ey\|_2^2 = \widetilde{k}$,

$$\begin{aligned}&\inf_{y\in S^{k-1}}\|P_{\widetilde{k}}V^*Ey\|_2^2\\ &\qquad\geq \Big(\widetilde{k} - \sup_{y\in S^{k-1}}\big|\|P_{\widetilde{k}}V^*Ey\|_2^2 - \mathbb{E}\|P_{\widetilde{k}}V^*Ey\|_2^2\big|\Big).\end{aligned}$$

Thus, noting that

$$\begin{aligned}\frac{\lambda^{\alpha(r-1/2)}}{s_{\widetilde{k}}} &< m^{\alpha(r-1/2)}k^{-\alpha(r-1/2)}C_1^{-r}m^{-r}\widetilde{k}^r\\ &= C_1^{-r}C_4^{r-\frac{1}{2}}\frac{\sqrt{\widetilde{k}}}{\sqrt{m}}\end{aligned}$$

and that by choosing $C_4 = \min(\frac{1}{2}C_1^{\frac{2r}{2r-1}}, \frac{1}{2})$ we ensure that $1 - C_1^{-2r}C_4^{2r-1} \geq \frac{1}{2}$,

$$\begin{aligned}&\mathbb{P}(\sigma_{\min}(\tfrac{1}{\sqrt{m}}SV^*E) \leq \lambda^{\alpha(r-1/2)})\\ &\leq \mathbb{P}(\sup_{y\in S^{k-1}}\big|\|\tfrac{1}{\sqrt{m}}P_{\widetilde{k}}V^*Ey\|_2^2 - \mathbb{E}\|\tfrac{1}{\sqrt{m}}P_{\widetilde{k}}V^*Ey\|_2^2\big| \geq \frac{\widetilde{k}}{2m}).\end{aligned} \tag{4}$$

Note that this choice of C_4 also ensures $\widetilde{k} \leq m$, which is required above. We will estimate (4) using the chaos bounds of [9], similarly to the proof of [9, Thm. A.1]. Indeed, we can write

$$\tfrac{1}{\sqrt{m}}P_{\widetilde{k}}V^*Ey = W_y\xi,$$

where ξ is a vector of length km with independent subgaussian entries of mean zero and variance 1, and

$$W_y = \frac{1}{\sqrt{m}}P_{\widetilde{k}}V^*\begin{pmatrix} y^T & 0 & \cdots & 0\\ 0 & y^T & \cdots & 0\\ \vdots & \vdots & \vdots & \vdots\\ 0 & \cdots & 0 & y^T\end{pmatrix}.$$

In order to apply [9, Thm. 3.1], we need to estimate, for $\mathcal{A} = \{W_y : y \in S^{k-1}\}$, $d_F(\mathcal{A}) := \sup_{A\in\mathcal{A}}\|A\|_F$, $d_{2\to2}(\mathcal{A}) := \sup_{A\in\mathcal{A}}\|A\|_{2\to2}$, and the Talagrand functional $\gamma_2(\mathcal{A}, \|\cdot\|_{2\to2})$ (see [8] for its definition). We obtain for $A = W_y \in \mathcal{A}$:

$$\|A\|_F^2 = \frac{1}{m}\sum_{j=1}^{k}\sum_{\ell_1,\ell_2=1}^{\widetilde{k},m} y_j^2V_{\ell_1,\ell_2}^2 = \frac{\widetilde{k}}{m}, \quad \text{so} \quad d_F(\mathcal{A}) = \sqrt{\frac{\widetilde{k}}{m}}.$$

Furthermore, we have, for $z \in \mathbb{R}^k$,

$$\begin{aligned}\|W_z\|_{2\to2} &= \left\|\frac{1}{\sqrt{m}}P_{\widetilde{k}}V^*\begin{pmatrix} z^T & 0 & \cdots & 0\\ 0 & z^T & \cdots & 0\\ \vdots & \vdots & \vdots & \vdots\\ 0 & \cdots & 0 & z^T\end{pmatrix}\right\|_{2\to2}\\ &\leq \left\|\frac{1}{\sqrt{m}}\begin{pmatrix} z^T & 0 & \cdots & 0\\ 0 & z^T & \cdots & 0\\ \vdots & \vdots & \vdots & \vdots\\ 0 & \cdots & 0 & z^T\end{pmatrix}\right\|_{2\to2},\end{aligned}$$

so the quantities $d_{2\to2}(\mathcal{A})$ and $\gamma_2(\mathcal{A}, \|\cdot\|_{2\to2})$ can be estimated in exact analogy to [9, Thm. A.1]. This yields $d_{2\to2}(\mathcal{A}) = \frac{1}{\sqrt{m}}$ and $\gamma_2(\mathcal{A}, \|\cdot\|_{2\to2}) \leq C_5\sqrt{\frac{k}{m}}$ for some constant C_5 depending only on c. With these estimates, we obtain for the quantities E, U, V in [9, Thm. 3.1]

$$\begin{aligned}E &\leq (2C_5+2)\frac{\sqrt{k\widetilde{k}}}{m}\\ U &\leq \frac{1}{m}\\ V &\leq (C_5+1)\frac{\sqrt{\widetilde{k}}}{m},\end{aligned}$$

so the resulting tail bound reads

$$\begin{aligned}&\mathbb{P}\Big(\sup_{y\in S^{k-1}}\big|\|\tfrac{1}{\sqrt{m}}P_{\widetilde{k}}V^*Ey\|_2^2 - \mathbb{E}\|\tfrac{1}{\sqrt{m}}P_{\widetilde{k}}V^*Ey\|_2^2\big| \geq\\ &\qquad\qquad c_1(2C_5+2)\frac{\sqrt{k\widetilde{k}}}{m} + t\Big)\\ &\leq e^{-c_2\min\left(\frac{t^2m^2}{(C_5+1)\widetilde{k}}, mt\right)}.\end{aligned}$$

where c_1 and c_2 are the constants depending only on c as they appear in [9, Thm. 3.1]. Note that $k = \widetilde{k}\frac{\lambda^{-(1-\alpha)}}{C_4}$, so for oversampling rates $\lambda > \big((4c_1(2C_5+2))^2/C_4\big)^{\frac{1}{1-\alpha}} =: C_2^{\frac{1}{1-\alpha}}$, we obtain $c_1E \leq \frac{\widetilde{k}}{4m}$ and hence, choosing $t = \frac{\widetilde{k}}{4m}$, we obtain the result

$$\begin{aligned}&\mathbb{P}(\sup_{y\in S^{k-1}}\big|\|\tfrac{1}{\sqrt{m}}P_{\widetilde{k}}V^*Ey\|_2^2 - \mathbb{E}\|\tfrac{1}{\sqrt{m}}P_{\widetilde{k}}V^*Ey\|_2^2\big| \geq \frac{\widetilde{k}}{2m})\\ &\leq e^{-C_3\widetilde{k}}\end{aligned}$$

where, as desired, the constant $C_3 := \frac{c_2}{16(C_5+1)}$ depends only on the subgaussian parameter c. ∎

Analogously to [1], we can use the above bounds to establish guarantees for recovery from $\Sigma\Delta$ quantized compressed sensing measurements.

Theorem 8: Let $\widetilde{\Phi}$ be an $m \times N$ sub-Gaussian matrix with parameter c and set $\Phi := \frac{1}{\sqrt{m}}\widetilde{\Phi}$, let $r \in \mathbb{Z}^+$, and let $0 < \alpha <$

1. Then exist constants C_6, C_7, C_8, C_9 depending only on r and c such that the following holds. Suppose that

$$\lambda := \frac{m}{k} \geq \left(C_6 \log(eN/k)\right)^{\frac{1}{1-\alpha}}.$$

Consider the $2L$-level rth order greedy $\Sigma\Delta$ schemes with step-size δ, denote by q the quantization output resulting from Φz where $z \in \mathbb{R}^N$, and denote by Δ a standard compressed sensing decoder. Then with probability exceeding $1 - 2e^{-C_7 m^{1-\alpha} k^{\alpha}}$ for all $z \in \Sigma_k^N$ having $\min_{j \in \mathrm{supp(z)}} |z_j| > C_8 \delta$:

1) the support of z, T, coincides with the support of the best k-term approximation of $\Delta(q)$.
2) denoting by E and F the sub-matrix of Φ corresponding to the support of z and its rth order Sobolev dual respectively, and by $x \in \mathbb{R}^k$ the restriction of z to its support, we have

$$\|x - Fq\|_2 \leq C_9 \lambda^{-\alpha(r-1/2)} \delta.$$

The proof traces the same steps as in [1]. Namely, 1) is a direct consequence of standard RIP-based recovery guarantees and 2) follows from a union bound over all submatrices consisting of k columns of Φ. This union bound determines the condition on λ and the probability. As the all the proof ingredients established above are identical to the corresponding results in [1], we omit the details.

V. Conclusion

Theorem 8 is a complete generalization of the main result of [1] to the scenario of sub-Gaussian matrices. Up to constants, the resulting embedding dimensions are the same as in the Gaussian case.

References

[1] C. Güntürk, M. Lammers, A. Powell, R. Saab, and Ö. Yılmaz, Sobolev duals for random frames and quantization of compressed sensing measurements," *Foundations of Computational Mathematics*, vol. 13, pp. 1–36, 2013.

[2] E. J. Candès, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Comm. Pure Appl. Math.*, vol. 59, pp. 1207–1223, 2006.

[3] D. Donoho, "Compressed sensing." *IEEE Trans. Inform. Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[4] H. Inose and Y. Yasuda, "A unity bit coding method by negative feedback," *Proceedings of the IEEE*, vol. 51, no. 11, pp. 1524–1535, 1963.

[5] J. Benedetto, A. Powell, and O. Yılmaz, "Sigma-delta ($\Sigma\Delta$) quantization and finite frames," *IEEE Trans. Inform. Theory*, vol. 52, no. 5, pp. 1990–2005, 2006.

[6] J. Blum, M. Lammers, A. Powell, and O. Yılmaz, "Sobolev duals in frame theory and sigma-delta quantization," *J. Fourier Anal. and Appl.*, vol. 16, no. 3, pp. 365–381, 2010.

[7] R. Vershynin, "Introduction to the non-asymptotic analysis of random matrices," in *Compressed Sensing: Theory and Applications*, Y. Eldar and G. Kutyniok, Eds. Cambridge: Cambridge Univ Press, 2012, pp. xii+544.

[8] M. Ledoux and M. Talagrand, *Probability in Banach Spaces: Isoperimetry and Processes*. New York: Springer, 1991.

[9] F. Krahmer, S. Mendelson, and H. Rauhut, Suprema of chaos processes and the restricted isometry property, *Communications on Pure and Applied Mathematics*, to appear.

Stable Recovery with Analysis Decomposable Priors

Jalal M. Fadili
GREYC
CNRS-ENSICAEN-Univ. Caen
Caen, France

Gabriel Peyré and Samuel Vaiter
CEREMADE
CNRS-Univ. Paris-Dauphine
Paris, France

Charles-Alban Deledalle
IMB
CNRS-Univ. Bordeaux 1
Bordeaux, France

Joseph Salmon
LTCI
CNRS-Télécom ParisTech
Paris, France

***Abstract*—In this paper, we investigate in a unified way the structural properties of solutions to inverse problems. These solutions are regularized by the generic class of semi-norms defined as a decomposable norm composed with a linear operator, the so-called analysis type decomposable prior. This encompasses several well-known analysis-type regularizations such as the discrete total variation (in any dimension), analysis group-Lasso or the nuclear norm. Our main results establish sufficient conditions under which uniqueness and stability to a bounded noise of the regularized solution are guaranteed. Along the way, we also provide a strong sufficient uniqueness result that is of independent interest and goes beyond the case of decomposable norms.**

I. Introduction

A. Problem statement

Suppose we observe

$$y = \Phi x_0 + w, \quad \text{where} \quad \|w\|_2 \leqslant \varepsilon \ ,$$

where Φ is a linear operator from $\mathbb{R}^N$ to $\mathbb{R}^M$ that may have a non-trivial kernel. We want to robustly recover an approximation of x_0 by solving the optimization problem

$$x^\star \in \operatorname*{Argmin}_{x \in \mathbb{R}^N} \tfrac{1}{2}\|y - \Phi x\|_2^2 + \lambda R(x) \ , \tag{1}$$

where

$$R(x) := \|L^* x\|_{\mathcal{A}} \ ,$$

with $L : \mathbb{R}^P \to \mathbb{R}^N$ a linear operator, and $\|\cdot\|_{\mathcal{A}} : \mathbb{R}^P \to \mathbb{R}^+$ is a decomposable norm in the sense of [1]. Decomposable regularizers are intended to promote solutions conforming to some notion of simplicity/low complexity that complies with that of $u_0 = L^* x_0$. This motivates the following definition of these norms. Throughout the paper, given a subspace $V \subset \mathbb{R}^P$, we will use the shorthand notation $L_V = L\mathcal{P}_V$, $L_V^* = \mathcal{P}_V L^*$, and $\alpha_V = \mathcal{P}_V \alpha$ for any vector $\alpha \in \mathbb{R}^P$, where $\mathcal{P}_V$ (resp. $\mathcal{P}_{V^\perp}$) is the orthogonal projector on V (resp. on its orthogonal complement $V^\perp$).

Definition 1. *A norm $\|\cdot\|_{\mathcal{A}}$ is* decomposable *at $u \in \mathbb{R}^P$ if:*

(i) there is a subspace $T \subset \mathbb{R}^P$ and a vector $e \in T$ such that

$$\partial\|\cdot\|_{\mathcal{A}}(u) = \left\{\alpha \in \mathbb{R}^P \mid \alpha_T = e \quad \text{and} \quad \|\alpha_{T^\perp}\|_{\mathcal{A}}^* \leqslant 1\right\}$$

(ii) and for any $z \in T^\perp$, $\|z\|_{\mathcal{A}} = \sup_{v \in T^\perp, \|v\|_{\mathcal{A}}^* \leqslant 1} \langle v, z\rangle$, where $\|\cdot\|_{\mathcal{A}}^*$ is the dual norm of $\|\cdot\|_{\mathcal{A}}$.

From this definition, it can be easily proved, using Fenchel identity, that $u \in T$ whenever $\|\cdot\|_{\mathcal{A}}$ is decomposable at u. Popular examples covered by decomposable regularizers are the ℓ_1-norm, the ℓ_1-ℓ_2 group sparsity norm, and the nuclear norm [1].

B. Contributions and relation to prior work

In this paper, we give a strong sufficient condition under which (1) admits a unique minimizer. From this, sufficient uniqueness conditions are derived. Then we develop results guaranteeing a stable approximation of x_0 from the noisy measurements y by solving (1), with an ℓ_2-error that comes within a factor of the noise level ε. This goes beyond [1] who considered identifiability under a generalized irrepresentable condition in the noiseless case with $L = \mathrm{Id}$. ℓ_2-stability for a class of decomposable priors closely related to Definition 1, is also studied in [8] for $L = \mathrm{Id}$ and general sufficiently smooth data fidelity. Their stability results require however stronger assumptions than ours (typically a restricted strong convexity which becomes a type of restricted eigenvalue property for linear regression with quadratic data fidelity). The authors in [3] provide sharp estimates of the number of generic measurements required for exact and ℓ_2-stable recovery of models from random partial information by solving a constrained form of (1) regularized by atomic norms. This is however restricted to the compressed sensing scenario. Our results generalize the stability guarantee of [7] established when the decomposable norm is ℓ_1 and L^* is the analysis operator of a frame. A stability result for general sublinear functions R is given in [6]. The stability is however measured in terms of R, and ℓ_2-stability can only be obtained if R is coercive, *i.e.,* L^* is injective.

At this stage, we would like to point out that although we carry out our analysis on the penalized form (1), our results remain valid for the data fidelity constrained version but obviously with different constants in the bounds. We omit these results for obvious space limitations.

II. Uniqueness

A. Main assumptions

We first note that traditional coercivity and convexity arguments allow to show that the set of (global) minimizers of (1) is a non-empty compact set if, and only if, $\ker(\Phi) \cap \ker(L^*) = \{0\}$.

The following assumptions will play a pivotal role in our analysis.

Assumption (SC_x) There exist $\eta \in \mathbb{R}^M$ and $\alpha \in \partial\|\cdot\|_{\mathcal{A}}(L^*x)$ such that the following so-called source (or range) condition is verified:

$$\Phi^*\eta = L\alpha \in \partial R(x) \ .$$

Assumption (INJ_T) For a subspace $T \subset \mathbb{R}^P$, Φ is injective on $\ker(L^*_{T^\perp})$.

It is immediate to see that since $\ker(L^*) \subseteq \ker(L^*_{T^\perp})$, ($\mathrm{INJ}_T$) implies that the set of minimizers is indeed non-empty and compact.

B. Strong Null Space Property

We shall now give a novel strong sufficient uniqueness condition under which problem (1) admits exactly one minimizer.

Theorem 1. *For a minimizer $x^\star$ of* (1)*, let T and e be the subspace and vector in Definition 1 associated to $u^\star = L^*x^\star$, and denote $S = T^\perp$. $x^\star$ is the unique minimizer of* (1) *if*

$$\langle L_T^* h,\, e\rangle < \|L_S^* h\|_{\mathcal{A}}^*, \qquad \forall h \in \ker(\Phi) \setminus \{0\} \ .$$

The above condition is a strong generalization of the Null Space Property well known in ℓ_1 regularization [4].

C. Sufficient uniqueness conditions

1) General case: A direct consequence of the above theorem is the following corollary.

Corollary 1. *For a minimizer $x^\star$ of* (1)*, let T and e be the subspace and vector in Definition 1 associated to $u^\star = L^*x^\star$, and denote $S = T^\perp$. Assume that* ($\mathrm{SC}_{x^\star}$) *is verified with $\|\alpha_S\|_{\mathcal{A}}^* < 1$, and that* ($\mathrm{INJ}_T$) *holds. Then, $x^\star$ is the unique minimizer of* (1)*.*

In fact, it turns out that the above two results are proved without requiring some restrictive implications of Definition 1(ii) of decomposable norms, and are therefore valid for a much larger class of regularizations. This can be clearly checked in the arguments used in the proofs.

2) Separable case:

Definition 2. *The decomposable norm $\|\cdot\|_{\mathcal{A}}$ is separable on the subspace $T^\perp = S = V \oplus W \subset \mathbb{R}^P$ if for any $u \in \mathbb{R}^P$, $\|u_{T^\perp}\|_{\mathcal{A}} = \|u_V\|_{\mathcal{A}} + \|u_W\|_{\mathcal{A}}$.*

Separability as just defined is fulfilled for several decomposable norms such as the ℓ_1 or $\ell_1 - \ell_p$ norms, $1 \leqslant p < +\infty$.

The non-saturation condition on the dual certificate required in Corollary 1 can be weakened to hold only on a subspace $V \subset S$ and the conclusions of the corollary remain valid, and assuming a stronger restricted injectivity assumption. We have the following corollary.

Corollary 2. *Assume that $\|\cdot\|_{\mathcal{A}}$ is also separable, with $S = V \oplus W$, such that* ($\mathrm{SC}_{x^\star}$) *is verified with $\|\alpha_V\|_{\mathcal{A}}^* < 1$, and* ($\mathrm{INJ}_V$) *holds. Then, $x^\star$ is the unique minimizer of* (1)*.*

III. Stability to noise

A. Main result

1) General case: We are now ready to state our main stability results.

Theorem 2. *Let T_0 and e_0 be the subspace and vector in Definition 1 associated to $u_0 = L^*x_0$, and denote $S_0 = T_0^\perp$. Assume that* (SC_{x_0}) *is verified with $\|\alpha_{S_0}\|_{\mathcal{A}}^* < 1$, and that* ($\mathrm{INJ}_{T_0}$) *holds. Then, choosing $\lambda = c\varepsilon$, $c > 0$, the following holds for any minimizer $x^\star$ of* (1)

$$\|x^\star - x_0\|_2 \leqslant C\varepsilon \ ,$$

where $C = C_1\,(2 + c\|\eta\|_2) + C_2\frac{(1+c\|\eta\|_2/2)^2}{c\left(1-\|\alpha_{S_0}\|_{\mathcal{A}}^*\right)}$, and $C_1 > 0$ and $C_2 > 0$ are constants independent of η and α.

Remark 1 (Separable case)**.** *When the decomposable norm is also separable (see Corollary 2), the stability result of Theorem 2 remains true assuming that $\|\alpha_V\|_{\mathcal{A}}^* < 1$ for $V \subset S_0$. This however comes at the price of the stronger restricted injectivity assumption* (INJ_V)*. To show this, the only thing to modify is the statement and the proof of Lemma 2 which can be done easily using similar arguments to those in the proof of Corollary 2.*

2) Case of frames: Suppose that L^* is the analysis operator of a frame ($\ker(L^*) = \{0\}$) with lower bound $a > 0$, let $\tilde{L}$ be a dual frame. The following stability bound can be obtained whose proof is omitted for space limitations.

Proposition 1. *Let T_0 and e_0 be the subspace and vector in Definition 1 associated to $u_0 = L^*x_0$, and denote $S_0 = T_0^\perp$. Assume that* (SC_{x_0}) *is verified with $\|\alpha_{S_0}\|_{\mathcal{A}}^* < 1$, and that Φ is injective on $\mathrm{Im}(\tilde{L}_{T_0})$. Then, choosing $\lambda = c\varepsilon$, $c > 0$, the following holds for any minimizer $x^\star$ of* (1)

$$\|x^\star - x_0\|_2 \leqslant C'\varepsilon \ ,$$

where $C' = C_1\,(2 + c\|\eta\|_2) + C_2'\frac{(1+c\|\eta\|_2/2)^2}{c\left(1-\|\alpha_{S_0}\|_{\mathcal{A}}^*\right)}$, and $C_1 > 0$ and $C_2' > 0$ are constants independent of η and α.

Since $\ker(L^*_{S_0}) \subseteq \mathrm{Im}(\tilde{L}_{T_0})$, the required restricted injectivity assumption is more stringent than (INJ_{T_0}). On the positive side, the constant C_2' is in general better than C_2. More precisely, the constant C_L, see the proof of Theorem 2, is replaced with $\sqrt{a}$. Note also that coercivity of R in this case allows to derive a bound similar to ours from the results in [6]. His restricted injectivity assumption is however different and our constants are sharper.

B. Generalized irrepresentable condition

In the following corollary, we provide a stronger sufficient stability condition that can be viewed as a generalization of the irrepresentable condition introduced in [5] when R is the ℓ_1 norm. It allows to construct dual vectors η and α which obey the source condition and are computable, which in turn yield explicit constants in the bound.

Definition 3. *Let $T \subset \mathbb{R}^P$ and $e \in \mathbb{R}^P$, and denote $S = T^\perp$. Suppose that* (INJ$_T$) *is verified. Define for any $u \in \ker(L_S)$ and $z \in \mathbb{R}^M$ such that $\Phi^* z \in \mathrm{Im}(L_S)$*

$$\mathrm{IC}_{u,z}(T,e) \;=\; \|\Gamma e + u_S + (L_S)^+ \Phi^* z\|_{\mathcal{A}}^*$$

where

$$\begin{aligned} \Gamma &= (L_S)^+(\Phi^*\Phi\Xi - \mathrm{Id})L_T \\ \Xi : h \mapsto \Xi h &= \operatorname*{argmin}_{x \in \ker(L_S^*)} \tfrac{1}{2}\|\Phi x\|_2^2 - \langle h,\, x\rangle \;, \end{aligned}$$

and M^+ is the Moore-Penrose pseudoinverse of M. Let $\bar{u}$, $\bar{z}$ and $\underline{u}$ defined as

$$\begin{aligned} (\bar{u},\bar{z}) &= \operatorname*{argmin}_{u \in \ker(L_S),\{z \mid \Phi^* z \in \mathrm{Im}(L_S)\}} \mathrm{IC}_{u,z}(T,e) \\ \textit{and} \quad \underline{u} &= \operatorname*{argmin}_{u \in \ker(L_S)} \mathrm{IC}_{u,0}(T,e) \;. \end{aligned}$$

Obviously, we have

$$\mathrm{IC}_{\bar{u},\bar{z}}(T,e) \leqslant \mathrm{IC}_{\underline{u},0}(T,e) \leqslant \mathrm{IC}_{0,0}(T,e) \;.$$

The convex programs defining $\mathrm{IC}_{\bar{u},\bar{z}}(T,e)$ and $\mathrm{IC}_{\underline{u},0}(T,e)$ can be solved using primal-dual proximal splitting algorithms whenever the proximity operator of $\|\cdot\|_{\mathcal{A}}$ can be easily computed [2]. The criterion $\mathrm{IC}_{\underline{u},0}(T,e)$ specializes to the one developed in [10] when $\|\cdot\|_{\mathcal{A}}$ is the ℓ_1 norm. $\mathrm{IC}_{0,0}(T,e)$ is a generalization of the coefficient involved in the irrepresentable condition introduced in [5] when R is the ℓ_1 norm, and to the one in [1] for decomposable priors with $L = \mathrm{Id}$.

Corollary 3. *Assume that* (INJ$_{T_0}$) *is verified and $\mathrm{IC}_{\bar{u},\bar{z}}(T_0,e_0) < 1$. Then, taking $\eta = \Phi\Xi L_{T_0} e_0 + \bar{z}$, one can construct α such that* (SC$_{x_0}$) *is satisfied and $\|\alpha_{S_0}\|_{\mathcal{A}}^* < 1$. Moreover, the conclusion of Theorem 2 remains true substituting $1 - \mathrm{IC}_{\bar{u},\bar{z}}(T_0,e_0)$ for $1 - \|\alpha_{S_0}\|_{\mathcal{A}}^*$.*

IV. Proofs

A. Proof of Theorem 1

A key observation is that by strong (hence strict) convexity of $\mu \mapsto \|y-\mu\|_2^2$, all minimizers of (1) share the same image under Φ. Therefore any minimizer of (1) takes the form $x^\star + h$ where $h \in \ker(\Phi)$. Furthermore, it can be shown by arguments from convex analysis that any proper convex function R has a unique minimizer $x^\star$ (if any) over a convex set C if its directional derivative satisfies

$$R'(x^\star; x - x^\star) > 0, \quad x \in C, x \neq x^\star \;.$$

Applying this to (1) with $C = x^\star + \ker(\Phi)$, and using the fact that the directional derivative is the support function of the subdifferential, we get that $x^\star$ is the unique minimizer of (1) if $\forall\, h \in \ker(\Phi) \setminus \{0\}$

$$\begin{aligned} 0 < R'(x^\star; h) &= \sup_{v \in \partial R(x^\star)} \langle v,\, h\rangle \\ &= \sup_{\alpha \in \partial\|\cdot\|_{\mathcal{A}}(L^* x^\star)} \langle \alpha,\, L^* h\rangle \\ &= \langle e,\, L_T^* h\rangle + \sup_{\|\alpha_S\|_{\mathcal{A}}^* \leqslant 1} \langle \alpha_S,\, L_S^* h\rangle \\ &= \langle e,\, L_T^* h\rangle + \|L_S^* h\|_{\mathcal{A}} \;. \end{aligned}$$

We conclude using symmetry of the norm and the fact that $\ker(\Phi)$ is a subspace. ∎

B. Proof of Corollary 1

The source condition (SC$_{x^\star}$) implies that $\forall\, h \in \ker(\Phi)\setminus\{0\}$

$$\langle h,\, L\alpha\rangle = \langle h,\, \Phi^*\eta\rangle = \langle \Phi h,\, \eta\rangle = 0 \;.$$

Moreover

$$\langle h,\, L\alpha\rangle = \langle L^* h,\, \alpha\rangle = \langle L_T^* h,\, e\rangle + \langle L_S^* h,\, \alpha_S\rangle \;.$$

Thus, applying the dual-norm inequality we get

$$\langle L_T^* h,\, e\rangle \leqslant \|L_S^* h\|_{\mathcal{A}} \|\alpha_S\|_{\mathcal{A}}^* < \|L_S^* h\|_{\mathcal{A}} \;,$$

where the last inequality is strict since $L_S^* h$ does not vanish owing to (INJ$_T$), and $\|\alpha_S\|_{\mathcal{A}}^* < 1$. ∎

C. Proof of Corollary 2

We follow the same lines as the proof of Corollary 1 and get

$$\langle L^* h,\, \alpha\rangle = \langle L_T^* h,\, e\rangle + \langle L_V^* h,\, \alpha_V\rangle + \langle L_W^* h,\, \alpha_W\rangle \;.$$

We therefore obtain

$$\begin{aligned} \langle L_T^* h,\, e\rangle &\leqslant \|L_V^* h\|_{\mathcal{A}} \|\alpha_V\|_{\mathcal{A}}^* + \|L_W^* h\|_{\mathcal{A}} \|\alpha_W\|_{\mathcal{A}}^* \\ &< \|L_V^* h\|_{\mathcal{A}} + \|L_W^* h\|_{\mathcal{A}} = \|L_S^* h\|_{\mathcal{A}}^* \;, \end{aligned}$$

where we used that $h \notin \ker(L_V^*)$, $\|\alpha_V\|_{\mathcal{A}}^* < 1$, separability and $\|\alpha_W\|_{\mathcal{A}}^* \leqslant \|\alpha_V\|_{\mathcal{A}}^* + \|\alpha_W\|_{\mathcal{A}}^* = \|\alpha_S\|_{\mathcal{A}}^* \leqslant 1$. ∎

D. Proof of Theorem 2

We first define the Bregman distance/divergence.

Definition 4. *Let $D_s^R(x,x_0)$ be the Bregman distance associated to R with respect to $s \in \partial R(x_0)$,*

$$D_s^R(x,x_0) = R(x) - R(x_0) - \langle s,\, x - x_0\rangle \;.$$

Define $D_\alpha^{\mathcal{A}}(u,u_0)$ as the Bregman distance associated to $\|\cdot\|_{\mathcal{A}}$ with respect to $\alpha \in \partial\|\cdot\|_{\mathcal{A}}(u_0)$.

Observe that by convexity, the Bregman distance is non-negative.

Preparatory lemmata We first need the following key lemmata.

Lemma 1 (Prediction error and Bregman distance convergence rates)**.** *Suppose that* (SC$_{x_0}$) *is satisfied. Then, for any minimizer $x^\star$ of* (1)*, and with $\lambda = c\varepsilon$ for $c > 0$, we have*

$$\begin{aligned} D_{\Phi^*\eta}^R(x^\star,x_0) = D_\alpha^{\mathcal{A}}(L^* x^\star, L^* x_0) &\leqslant \varepsilon \frac{(1 + c\|\eta\|_2/2)^2}{c} \;, \\ \|\Phi x^\star - \Phi x_0\|_2 &\leqslant \varepsilon(2 + c\|\eta\|_2) \;. \end{aligned}$$

The proof follows the same lines as that for any sublinear regularizer, see e.g. [9], where we additionally use the source condition (SC$_{x_0}$) and $D_{\Phi^*\eta}^R(x,x_0) = D_{L\alpha}^R(x,x_0) = D_\alpha^{\mathcal{A}}(L^* x, L^* x_0)$.

Now since $\|\cdot\|_{\mathcal{A}}$ is a norm, it is coercive, and thus

$$\exists\, C_{\mathcal{A}} > 0 \quad \text{s.t.} \quad \forall x \in \mathbb{R}^P,\; \|x\|_{\mathcal{A}} \geqslant C_{\mathcal{A}} \|x\|_2 .$$

We get the following inequality.

Lemma 2 (From Bregman to ℓ_2 bound)**.** *Suppose that* (SC_{x_0}) *holds with* $\|\alpha_{S_0}\|^*_{\mathcal{A}} < 1$. *Then,*

$$\|L^*_{S_0}(x^\star - x_0)\|_2 \leqslant \frac{D^{\mathcal{A}}_\alpha(L^*x^\star, L^*x_0)}{C_{\mathcal{A}}\,(1 - \|\alpha_{S_0}\|^*_{\mathcal{A}})} \ ,$$

Proof: Decomposability of $\|\cdot\|_{\mathcal{A}}$ implies that $\exists v \in S_0$ such that $\|v\|^*_{\mathcal{A}} \leqslant 1$ and $\|L^*_{S_0}(x^\star - x_0)\|_{\mathcal{A}} = \langle L^*_{S_0}(x^\star - x_0),\, v\rangle$. Moreover, $v + e_0 \in \partial\|\cdot\|_{\mathcal{A}}(L^*x_0)$. Thus

$$\begin{aligned}
D^{\mathcal{A}}_\alpha(L^*x^\star, L^*x_0) &\geqslant D^{\mathcal{A}}_\alpha(L^*x^\star, L^*x_0) \\
&\qquad - D^{\mathcal{A}}_{v+e_0}(L^*x^\star, L^*x_0) \\
&= \langle v + e_0 - \alpha,\, L^*(x^\star - x_0)\rangle \\
&= \langle v - \alpha_{S_0},\, L^*_{S_0}(x^\star - x_0)\rangle \\
&= \|L^*_{S_0}(x^\star - x_0)\|_{\mathcal{A}} \\
&\qquad - \langle \alpha_{S_0},\, L^*_{S_0}(x^\star - x_0)\rangle \\
&\geqslant \|L^*_{S_0}(x^\star - x_0)\|_{\mathcal{A}}(1 - \|\alpha_{S_0}\|^*_{\mathcal{A}}) \\
&\geqslant C_{\mathcal{A}}\|L^*_{S_0}(x^\star - x_0)\|_2(1 - \|\alpha_{S_0}\|^*_{\mathcal{A}}) \ .
\end{aligned}$$

■

Proof of the main result

$$\begin{aligned}
\|x^\star - x_0\|_2 &\leqslant \|\mathcal{P}_{\ker(L^*_{S_0})}(x^\star - x_0)\|_2 \\
&\qquad + \|\mathcal{P}_{\mathrm{Im}(L^*_{S_0})}(x^\star - x_0)\|_2 \\
&\leqslant {C_\Phi}^{-1}\|\Phi\mathcal{P}_{\ker(L^*_{S_0})}(x^\star - x_0)\|_2 \\
&\qquad + \|\mathcal{P}_{\mathrm{Im}(L^*_{S_0})}(x^\star - x_0)\|_2 \\
&\leqslant {C_\Phi}^{-1}\|\Phi(x^\star - x_0)\|_2 \\
&\qquad + (1 + {C_\Phi}^{-1}\|\Phi\|_{2,2})\|\mathcal{P}_{\mathrm{Im}(L^*_{S_0})}(x^\star - x_0)\|_2 \ ,
\end{aligned}$$

where we used assumption (INJ_{T_0}), *i.e.,*

$$\exists\ C_\Phi > 0 \quad \text{s.t.} \quad \|\Phi x\|_2 \geqslant C_\Phi\|x\|_2, \quad \forall x \in \ker(L^*_{S_0}) \ .$$

Since $L^*_{S_0}$ is injective on the orthogonal of its kernel, there exists $C_L > 0$ such that

$$\begin{aligned}
\|x^\star - x_0\|_2 &\leqslant {C_\Phi}^{-1}\|\Phi(x^\star - x_0)\|_2 \\
&\qquad + \tfrac{\|\Phi\|_{2,2} + C_\Phi}{C_L C_\Phi}\|L^*_{S_0}\mathcal{P}_{\mathrm{Im}(L^*_{S_0})}(x^\star - x_0)\|_2 \ .
\end{aligned}$$

Noticing that

$$\|L^*_{S_0}(x^\star - x_0)\|_2 = \|L^*_{S_0}\mathcal{P}_{\mathrm{Im}(L^*_{S_0})}(x^\star - x_0)\|_2,$$

we apply Lemma 2 to get

$$\begin{aligned}
\|x^\star - x_0\|_2 &\leqslant {C_\Phi}^{-1}\|\Phi(x^\star - x_0)\|_2 \\
&\qquad + \tfrac{\|\Phi\|_{2,2} + C_\Phi}{C_L C_\Phi\left(1 - \|\alpha_{S_0}\|^*_{\mathcal{A}}\right)} D^{\mathcal{A}}_\alpha(L^*x^\star, L^*x_0) \ .
\end{aligned}$$

Using Lemma 1 yields the desired result. ■

E. Proof of Corollary 3

Take $\alpha = e_0 + \Gamma e_0 + \bar{u}_{S_0} + (L_{S_0})^+\Phi^*\bar{z}$. First, $\alpha_{T_0} = e_0$ since $e_0 \in T_0$ and $\mathrm{Im}(\Gamma) \subseteq \mathrm{Im}((L_{S_0})^+) = \mathrm{Im}(L^*_{S_0})$. Then $\|\alpha_{S_0}\|^*_{\mathcal{A}} = \mathrm{IC}_{\bar{u},\bar{z}}(T_0, e_0) < 1$, whence we get that $\alpha \in \partial\|\cdot\|_{\mathcal{A}}(L^*x_0)$.

Now, we observe by definition of Ξ that $\mathcal{P}_{\ker(L^*_{S_0})}(\Phi^*\Phi\Xi - \mathrm{Id})L_{T_0} = 0$, which implies that $\mathrm{Im}((\Phi^*\Phi\Xi - \mathrm{Id})L_{T_0})) \subseteq \mathrm{Im}(L_{S_0})$. In turn, $L_{S_0}\Gamma = L_{S_0}(L_{S_0})^+\left((\Phi^*\Phi\Xi - \mathrm{Id})L_{T_0}\right) = \mathcal{P}_{\mathrm{Im}(L_{S_0})}\left((\Phi^*\Phi\Xi - \mathrm{Id})L_{T_0}\right) = (\Phi^*\Phi\Xi - \mathrm{Id})L_{T_0}$. This, together with the fact that $\bar{u} \in \ker(L_{S_0})$ and $\Phi^*\bar{z} \in \mathrm{Im}(L_{S_0})$ yields

$$\begin{aligned}
L_{S_0}\alpha &= (\Phi^*\Phi\Xi - \mathrm{Id})L_{T_0}e_0 + \Phi^*\bar{z} \\
&= \Phi^*\eta - L_{T_0}\alpha \iff \Phi^*\eta = L\alpha \ ,
\end{aligned}$$

which implies that $\Phi^*\eta = L\alpha \in \partial R(x_0)$. We have just shown that the vectors α and η as given above satisfy the source condition (SC_{x_0}) and the dual non-saturation condition. We conclude by applying Theorem 2 using (INJ_{T_0}). ■

V. Conclusion

We provided a unified analysis of the structural properties of regularized solutions to linear inverse problems through a class of semi-norms formed by composing decomposable norms with a linear operator. We provided conditions that guarantee uniqueness, and also those ensuring stability to bounded noise. The stability bound was achieved without requiring (even partial) recovery of T_0 and e_0. Recovery of T_0 and e_0 for analysis-type decomposable priors and beyond is currently under investigation. Another perspective concerns whether the ℓ_2 bound on $x^\star - x_0$ can be extended to cover more general low complexity-inducing regularizers beyond decomposable norms.

References

[1] E. J. Candès and B. Recht. Simple bounds for recovering low-complexity models. *Mathematical Programming*, pages 1–13, 2012.

[2] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011.

[3] V. Chandrasekaran, B. Recht, P. Parrilo, and A. Willsky. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12:805–849, 2012.

[4] D. L. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Transactions on Information Theory*, 47(7):2845–2862, 2001.

[5] J.-J. Fuchs. On sparse representations in arbitrary redundant bases. *IEEE Trans. Info. Theory*, 50(6):1341–1344, 2004.

[6] M. Grasmair. Linear convergence rates for Tikhonov regularization with positively homogeneous functionals. *Inverse Problems*, 27:075014, 2011.

[7] M. Haltmeier. Stable signal reconstruction via ℓ^1-minimization in redundant, non-tight frames. *IEEE Trans. on Sig. Proc.*, 2012. to appear.

[8] S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, December 2012.

[9] O. Scherzer, M. Grasmair, H. Grossauer, M. Haltmeier, and F. Lenzen. *Variational Methods in Imaging*. Applied Mathematical Sciences. Springer, 1st edition, 2009.

[10] S. Vaiter, G. Peyré, C. Dossal, and M.J. Fadili. Robust sparse analysis regularization. *to appear in IEEE Trans. Inf. Theo.*, 2012.

FRI-based Sub-Nyquist Sampling and Beamforming in Ultrasound and Radar

Tanya Chernyakova, Omer Bar-Ilan, Yonina C. Eldar
Department of Electrical Engineering
Technion, Israel Institute of Technology
Haifa 32000
Email: ctanya@tx.technion.ac.il, omerba@tx.technion.ac.il, yonina@ee.technion.ac.il

***Abstract*—Signals consisting of short pulses are present in many applications including ultrawideband communication, object detection and navigation (radar, sonar) and medical imaging. The structure of such signals, effectively captured within the finite rate of innovation (FRI) framework, allows for significant reduction in sampling rates, required for perfect reconstruction. In this work we consider two applications, ultrasound imaging and radar, where the FRI signal structure allows to reduce both sampling and processing rates. Furthermore, we show how the FRI framework inspires new processing techniques, such as beamforming in the frequency domain and Doppler focusing. In both applications a pulse of a known shape or a stream of such pulses is transmitted into the respective medium, and the received echoes are sampled and digitally processed in a way referred to as beamforming. Applied either spatially or temporally, beamforming allows to improve signal-to-noise ratio. In radar applications it also allows for target Doppler frequency estimation. Using FRI modeling both for detected and beamformed signals, we are able to reduce sampling rates and to perform digital beamforming directly on the low-rate samples.**

I. Introduction

When sampling an analog signal, we aim to represent it by discrete-time coefficients, while capturing its important features. According to the classic Shannon-Nyquist theorem the minimal sampling rate required for perfect reconstruction of bandlimited signals is twice the the maximal frequency. The required sampling rate can be significantly reduced when additional information about the signal structure is available. An interesting class of structured signals was suggested by Vetterli et al. [1], who considered signals with a finite number of degrees of freedom per unit time - signals with finite rate of innovation (FRI). One of the most studied cases of FRI signals is a stream of pulses, namely, a signal consisting of a stream of short pulses where the pulse shape is known. Such signals are presented in abundance in ultrawideband communication, object detection and navigation (radar, sonar) and medical imaging.

In this work we consider two applications where the FRI signal structure allows to reduce both sampling and processing rates and inspires new processing techniques. In particular, we show how different forms of beamforming, used to improve resolution and increase signal-to-noise-ratio (SNR), can be implemented directly on reduced rate samples. This is achieved by replacing the standard time-domain beamforming by a frequency domain approach and relying on previous FRI sampling techniques in frequency [2]–[4].

The first application is medical ultrasound, where the known pulse shape is transmitted into the tissue and the echoes reflected off scatterers form a stream of pulses signal detected by the elements of the transducer. Signals detected at each element are sampled and digitally processed by beamforming in time, exploiting the array geometry. Such a beamformed signal forms a line in the image. Treating both detected and beamformed signals in the FRI framework and performing beamforming in frequency allows to reduce the sampling rate far below standard rates that are required to improve the system's beamforming resolution.

The second application is radar. Similar to ultrasound, a stream of known pulses is transmitted into space and reflected off any targets. Whereas in ultrasound digital beamforming is performed spatially, i.e. combining a single pulse from different transducers, in the single transceiver radar model we consider beamforming is performed temporally between different pulses on the same transceiver. This beamforming process, besides improving SNR, allows for target Doppler frequency estimation as well. Here again we show how beamforming, and consequently, radar detection, can be performed efficiently at sub-Nyquist rates by using sub-Nyquist sampling methods in frequency [4], [5].

II. Ultrasound

Modern imaging systems use multiple transducer elements to transmit and receive acoustic pulses. The imaging process is described as follows: An energy pulse is transmitted along a narrow beam. During its propagation echoes are scattered by acoustic impedance perturbations in the tissue, and detected by the elements of the transducer. Collected data are sampled and digitally beamformed, resulting in an image line.

Rates up to 4 times the Nyquist rate, dictated by the bandwidth of the individual signal, are required in order to improve the system's beamforming resolution and to avoid artifacts caused by digital implementation. From now on we will denote this sampling rate as the beamforming rate f_s.

To get a sense of the sampling and processing rates involved in ultrasound imaging, we can evaluate the number of samples taken at each transducer element based on the imaging setup used to acquire in vivo cardiac data. The acquisition was performed with a GE breadboard ultrasonic scanner of 64

acquisition channels. The radiated depth $r = 16$ cm and the speed of the sound $c = 1540$ m/sec yield a signal of duration $T = 2r/c \simeq 210$ μsec. The acquired signal is characterized by a narrow bandpass bandwidth of 2 MHz, centered at the carrier frequency $f_0 \approx 3.1$ MHz, leading to a beamforming rate of $f_s \approx 16$ MHz and $Tf_s = 3360$ real-valued samples.

We now show that the number of samples can be reduced significantly since the oversampling dictated by the digital implementation of beamforming in time can be bypassed, when the beamformed signal is treated within the FRI framework and beamforming is performed in the frequency domain.

A. Signal Model

According to [2], [4], the beamformed signal in ultrasound imaging obeys an FRI model:

$$\Phi(t;\theta) \simeq \sum_{l=1}^{L} \tilde{b}_l h(t - t_l), \tag{1}$$

where $h(t)$ is the transmitted pulse-shape, L is the number of scattering elements in direction θ, $\{\tilde{b}_l\}_{l=1}^{L}$ are the unknown amplitudes of the reflections and $\{t_l\}_{l=1}^{L}$ denote the unknown delays. Sampling both sides of (1) at the rate f_s and quantizing the delays $\{t_l\}_{l=1}^{L}$ with quantization step $1/f_s$, such that $t_l = q_l/f_s, q_l \in \mathbb{Z}$, we can rewrite (1) as follows:

$$\Phi[n;\theta] \simeq \sum_{l=1}^{L} \tilde{b}_l h[n - q_l] = \sum_{l=0}^{N-1} b_l h[n - l], \tag{2}$$

where

$$b_l = \begin{cases} \tilde{b}_l & \text{if } l = q_l \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

Calculating the Discrete Fourier Transform (DFT) using (2):

$$c_k = \sum_{n=0}^{N-1} \Phi[n;\theta] e^{-i\frac{2\pi}{N}kn} = h_k \sum_{l=0}^{N-1} b_l e^{-i\frac{2\pi}{N}kl}, \tag{4}$$

where h_k is the DFT coefficient of $h[n]$. The transmitted pulse $h(t)$ is a narrowband baseband waveform, $g(t)$, modulated by a carrier at frequency f_0. When such a pulse is sampled at rate f_s, most of its DFT coefficients are zero. Obviously, (4) implies that the only non-zero DFT coefficients are in the bandwidth of the transmitted pulse. When a set κ of these non-zero DFT coefficients is known we can reconstruct the signal perfectly by zero-padding and then performing an inverse DFT (IDFT). In the cardiac imaging setup described above the bandwidth of $g(t)$ is equal to 2 MHz, the modulation frequency $f_0 = 3.1$ MHz, and the sampling rate $f_s = 16$ MHz, leading to $K = |\kappa| \approx 360$.

As we show further in Section II-C, sampling rates are proportional to the number of DFT coefficients of the beamformed signal that we want to calculate. Hence, to reduce the sampling rates, we aim to obtain only a subset $\mu \subset \kappa$, $|\mu| = M < K = |\kappa|$, of non-zero DFT coefficients of the beamformed signal and propose a method to reconstruct κ from its subset μ.

B. Beamformed Signal Reconstruction

Defining a K-length vector $\mathbf{c}$ with k-th entry $c_k/h_k, k \in \kappa$, we can rewrite (4) in matrix form:

$$\mathbf{c} = \mathbf{D}\mathbf{b}, \tag{5}$$

where $\mathbf{D}$ is a $K \times N$ matrix formed by taking the set κ of rows from an $N \times N$ DFT matrix, and vector $\mathbf{b}$ is of length N with l-th entry b_l. Since from now on only subset μ is given, define an M-length vector $\mathbf{c}_\mu$ with k-th entry $c_k/h_k, k \in \mu$ and rewrite (5) as follows:

$$\mathbf{c}_\mu = \mathbf{A}\mathbf{D}\mathbf{b}, \tag{6}$$

where $\mathbf{A}$ is an $M \times K$ measurement matrix which picks the subset μ of rows from $\mathbf{D}$.

We propose an analysis approach [6], namely, we aim to reconstruct the set κ from its subset μ, while assuming that the analyzed vector $\mathbf{D}^*\mathbf{c}$ is compressible. This assumption is justified as follows: A typical beamformed ultrasound signal is comprised of a relatively small number of strong reflections and a bunch of much weaker scattered echoes. It is, therefore, natural to assume that $\mathbf{b}$ from (5) is compressible, implying that $\mathbf{c}$ has a compressible expansion in $\mathbf{D}$. Since $\mathbf{D}$ is a partial DFT matrix, its Gram matrix is nearly diagonal, implying that $\mathbf{D}^*\mathbf{c}$ is also compressible [6]. The analysis approach can be translated into the l_1 optimization problem:

$$\min_{\mathbf{c}} \|\mathbf{D}^*\mathbf{c}\|_1 \quad \text{subject to} \quad \|\mathbf{A}\mathbf{c} - \mathbf{c}_\mu\|_2 \le \varepsilon. \tag{7}$$

Under certain conditions which are satisfied in our ultrasound setup [6], [7] the solution of (7) yields the set $\tilde{\kappa}$ of non-zero DFT coefficients of the beamformed signal which is sufficiently close to the true values of κ.

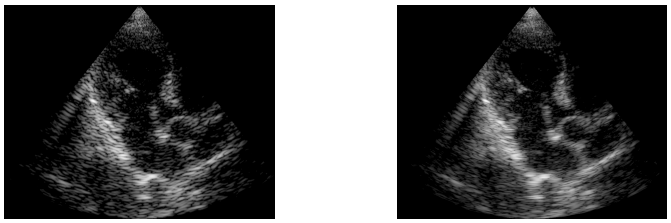
C. Sampling Scheme and Beamforming in Frequency

We now address the following question: how many samples of the individual signals should be taken in order to compute the subset μ of non-zero DFT coefficients of the beamformed signal?

To answer this question we introduce a recently developed technique, referred to as beamforming in frequency. This method was proposed in [4] and [7], where it was shown that beamforming can be performed directly in the frequency domain, namely, a set μ of the DFT coefficients of the beamformed signal can be calculated as a linear combination of a set ν of the DFT coefficients of each individual signal. Experimental results show that $|\nu| \approx |\mu|$, implying that we can calculate the desired set of beamformed DFT coefficients from a small number of DFT coefficients of each individual signal.

As it was shown in [2], [4], [7], a set ν of the DFT coefficients of each individual signal can be obtained by the sub-Nyquist Xampling ("compressed sampling") [8] method, an analog-to-digital conversion (ADC) which performs analog prefiltering of the signal before taking low-rate point-wise samples. The number of samples taken from the individual signal in this case is $|\nu| \approx |\mu|$.

To demonstrate the proposed method and evaluate the rate reduction, a subset μ of 100 DFT coefficients corresponding to the central frequency samples in the bandwidth of the transmitted pulse were chosen. To calculate μ we need approximately $|\mu| = 100$ samples per individual signal, implying 30 fold reduction in sampling rate. The result is shown in Fig. 1 (a). We compare it with an image created by a standard technique using 3360 samples per individual signal in Fig. 1 (b). As can be seen, we obtain sufficient image quality with more than 30 fold reduction in sampling rate.



(a) (b)

Fig. 1: Cardiac images. (a) Proposed method, 100 samples per image line. (b) Standard method, 3360 samples per image line.

III. RADAR

We next consider target detection and feature extraction in a single transceiver, monostatic, narrow-band pulse-train radar system. We show that both sampling and processing can be performed at sub-Nyquist rates, when an appropriate signal model is used. Targets are non-fluctuating point targets, sparsely populated in the radar's unambiguous time-frequency region: delays up to the Pulse Repetition Interval (PRI) and Doppler frequencies up to its reciprocal the Pulse Repetition Frequency (PRF). We propose a recovery method which can detect and estimate targets' delay and Doppler, using a linear, non-adaptive sampling technique at a rate significantly lower than the radar signal's Nyquist frequency, assuming the number of targets L is small.

Current state-of-the-art radar systems sample at the signal's Nyquist rate, which can be hundreds of MHz and higher. Similarly to the ultrasound application, the goal of our approach, breaking the link between the signal bandwidth and sampling rate, is achieved by using FRI signal model and the Xampling method. The latter yields compressed samples ("Xamples"), containing the information needed to recover the desired signal parameters. This work expands [5], adding Doppler to the target model and proposing a new digital recovery method to estimate it by relying on beamforming ideas operating on sub-Nyquist samples, as we showed in the context of ultrasound imaging.

A. Signal Model

We consider a radar transceiver that transmits a pulse train

$$x_T(t) = \sum_{p=0}^{P-1} h(t - p\tau), \quad 0 \le t \le P\tau \tag{8}$$

consisting of P equally spaced pulses $h(t)$. The pulse-to-pulse delay τ is referred to as the PRI. The pulse $h(t)$ is a known time-limited baseband function with continuous-time Fourier transform (CTFT) $H(\omega) = \int_{-\infty}^{\infty} h(t)e^{-j\omega t}dt$. We assume that $H(\omega)$ has negligible energy at frequencies beyond $B_h/2$ and we refer to B_h as the bandwidth of $h(t)$. The target scene is composed of L non-fluctuating point targets, where we assume that L is known, although this assumption can easily be relaxed. The pulses reflect off the L targets and propagate back to the transceiver. Each target l is defined by three parameters: a delay τ_l, a Doppler frequency ν_l and a complex amplitude α_l, proportional to the target's radar cross section (RCS) and all propagation factors.

Under several assumptions [9], we can write the received signal as

$$x(t) = \sum_{p=0}^{P-1} \sum_{l=0}^{L-1} \alpha_l h(t - \tau_l - p\tau) e^{-j\nu_l p\tau}. \tag{9}$$

It will be convenient to express the signal as a sum of single frames

$$x(t) = \sum_{p=0}^{P-1} x_p(t), \tag{10}$$

where

$$x_p(t) = \sum_{l=0}^{L-1} \alpha_l h(t - \tau_l - p\tau) e^{-j\nu_l p\tau}. \tag{11}$$

It is evident from (9) that we are dealing with an FRI signal, since it can be described by $3L$ parameters spanning an interval of duration $P\tau$, yielding a rate of innovation of $3L/P\tau$. Our goal is to accurately detect the L targets, i.e. to estimate the $3L$ parameters $\{\alpha_l, \tau_l, \nu_l\}_{l=0}^{L-1}$ in (9), using the least possible number of digital samples.

B. Doppler Focusing

The Doppler Focusing processing technique uses target echoes from different pulses to create a single superimposed pulse, improving SNR for robustness against noise and implicitly estimating targets' Doppler in the process. Using (11), we define the following time shift and modulation operation on the received signal:

$$\begin{aligned}\Phi(t;\nu) &= \sum_{p=0}^{P-1} x_p(t + p\tau) e^{j\nu p\tau} \\ &= \sum_{p=0}^{P-1} \sum_{l=0}^{L-1} \alpha_l h(t - \tau_l) e^{j(\nu - \nu_l)p\tau} \\ &= \sum_{l=0}^{L-1} \alpha_l h(t - \tau_l) \sum_{p=0}^{P-1} e^{j(\nu - \nu_l)p\tau}.\end{aligned} \tag{12}$$

We now analyze the sum of exponents in (12). For any given ν, targets with Doppler frequency ν_l in a band of width $2\pi/P\tau$ around ν, i.e. in $\Phi(t;\nu)'s$ "focus zone", will achieve

coherent integration and an SNR boost of approximately

$$g(\nu|\nu_l) = \sum_{p=0}^{P-1} e^{j(\nu-\nu_l)p\tau} \overset{|\nu-\nu_l|<2\pi/P\tau}{\cong} P \qquad (13)$$

compared with a single pulse. On the other hand, since the sum of P equally spaced points covering the unit circle is generally close to zero, targets with ν_l not "in focus" will approximately cancel out. Thus $g(\nu|\nu_l) \cong 0$ for $|\nu - \nu_l| > 2\pi/P\tau$. Hence we can approximate (12) by

$$\Phi(t;\nu) \cong P \sum_{l:|\nu-\nu_l|<2\pi/P\tau} \alpha_l h(t-\tau_l). \qquad (14)$$

Instead of trying to estimate delay and Doppler together, we have reduced our problem to delay only estimation for a small range of Doppler frequencies, with increased amplitude for improved performance against noise.

C. Delay-Doppler Recovery Using Doppler Focusing

Calculating the DFT of each of the pulses $x_p(t)$ of the multi-pulse signal (9), and since $x_p(t)$ is confined to the interval $t \in [p\tau, (p+1)\tau]$, we obtain

$$c_p[k] = \frac{1}{\tau} H(2\pi k/\tau) \sum_{l=0}^{L-1} \alpha_l e^{-j\nu_l p\tau} e^{-j2\pi k\tau_l/\tau}, \qquad (15)$$

where we used the fact that since both $k, p \in \mathbb{Z}$ we have $e^{-j2\pi kp} \equiv 1$. From (15) we see that all $3L$ unknown parameters $\{\alpha_l, \tau_l, \nu_l\}_{l=0}^{L-1}$ are embodied in the Fourier coefficients $c_p[k]$ in the form of a complex sinusoid problem.

Having acquired $c_p[k]$ using a framework similar to one introduced in section II-C, we now perform the Doppler focusing operation for a specific frequency ν

$$\begin{aligned}\Psi_\nu[k] &= \sum_{p=0}^{P-1} c_p[k] e^{j\nu p\tau} \\ &= \frac{1}{\tau} H(2\pi k/\tau) \sum_{l=0}^{L-1} \alpha_l e^{-j2\pi k\tau_l/\tau} \sum_{p=0}^{P-1} e^{j(\nu-\nu_l)p\tau}. \qquad (16)\end{aligned}$$

Following the same arguments as in (13), for any target l satisfying $|\nu - \nu_l| < 2\pi/P\tau$ we have

$$\sum_{p=0}^{P-1} e^{j(\nu-\nu_l)p\tau} \cong P. \qquad (17)$$

Therefore, Doppler focusing can be performed on the low rate sub-Nyquist samples:

$$\Psi_\nu[k] \cong \frac{P}{\tau} H(2\pi k/\tau) \sum_{l:|\nu-\nu_l|<2\pi/P\tau} \alpha_l e^{-j2\pi k\tau_l/\tau}. \qquad (18)$$

Equation (18) is scaled by P compared with a single pulse, increasing SNR for improved performance with noise. Furthermore, we reduced the number of active delays. For each ν we now have a delay estimation problem, which can be written in vector form as

$$\mathbf{\Psi}_\nu = \frac{P}{\tau} \mathbf{H} \mathbf{V} \mathbf{x}_\nu, \qquad (19)$$

where

$$\mathbf{\Psi}_\nu = [\Psi_\nu[k_0] \ ... \ \Psi_\nu[k_{|\kappa|-1}]]^T \in \mathbb{C}^{|\kappa|}. \qquad (20)$$

This is a CS problem which has already been solved [3], [9], [10]. We emphasize that the Doppler focusing technique is a continuous operation on ν, and can be performed for any Doppler frequency. Since the focus zone for each ν is of width $2\pi/P\tau$, we can find various finite sets of ν's spanning $[0, 2\pi/\tau]$. For any such set, define its size as N_ν. For each ν in the set, we solve (19) assuming $\mathbf{x}_\nu$'s support is of size L. This problem can be solved using an abundance of CS algorithms [11]–[13]. After solving N_ν separate CS problems with dictionary of size $|\kappa| \times N_\tau$, we hold at most LN_ν estimated amplitudes. Since the absolute value of amplitudes recovered in the support is indicative of true target existence as opposed to noise, we take the L strongest ones as true target locations.

REFERENCES

[1] M. Vetterli, P. Marziliano, and T. Blu, "Sampling signals with finite rate of innovation," *IEEE Transactions on Signal Processing*, vol. 50, no. 6, pp. 1417–1428, 2002.

[2] R. Tur, Y.C. Eldar, and Z. Friedman, "Innovation rate sampling of pulse streams with application to ultrasound imaging," *IEEE Transactions on Signal Processing*, vol. 59, no. 4, pp. 1827–1842, 2011.

[3] K. Gedalyahu, R. Tur, and Y. C. Eldar, "Multichannel sampling of pulse streams at the rate of innovation," *IEEE Transactions on Signal Processing*, vol. 59, no. 4, pp. 1491–1504, 2011.

[4] N. Wagner, Y. C. Eldar, and Z. Friedman, "Compressed beamforming in ultrasound imaging," *IEEE Transactions on Signal Processing*, vol. 60, no. 9, pp. 4643–4657, 2012.

[5] E. Baransky, G. Itzhak, I. Shmuel, N. Wagner, E. Shoshan, and Y. C. Eldar, "A Sub-Nyquist Radar Prototype: Hardware and Algorithms," *submitted to IEEE Transactions on Aerospace and Electronic Systems, special issue on Compressed Sensing for Radar*, Aug. 2012.

[6] E. J. Candes, Y. C. Eldar, D. Needell, and P. Randall, "Compressed sensing with coherent and redundant dictionaries," *Applied and Computational Harmonic Analysis*, vol. 31, no. 1, pp. 59–73, 2011.

[7] T. Chernyakova, Y.C. Eldar, and R. Amit, "Fourier domain beamforming for medical ultrasound," *arXiv preprint arXiv:1212.4940*, 2012.

[8] M. Mishali, Y. C. Eldar, O. Dounaevsky, and E. Shoshan, "Xampling: Analog to Digital at Sub-Nyquist Rates," *IET Circuits, Devices and Systems*, vol. 5, no. 1, pp. 8–20, Jan. 2011.

[9] O. Bar-Ilan and Y.C. Eldar, "Sub-Nyquist Radar via Doppler Focusing," *arXiv preprint arXiv:1211.0722*, 2012.

[10] W. U. Bajwa, K. Gedalyahu, and Y. C. Eldar, "Identification of Parametric Underspread Linear Systems and Super-Resolution Radar," *IEEE Transactions on Signal Processing*, vol. 59, no. 6, pp. 2548–2561, June 2011.

[11] S. G. Mallat and Z. Zhang, "Matching Pursuits with Time-Frequency Dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3397–3415, Dec. 1993.

[12] T. Blumensath and M. Davies, "Iterative Hard Thresholding for Compressive Sensing," *Applied Computational Harmonic Analysis*, vol. 27, no. 3, pp. 265–274, 2009.

[13] Y. C. Eldar and G. Kutyniok, "Compressed sensing: Theory and applications," *New York: Cambridge Univ. Press*, 2012.

Robust Spike Train Recovery from Noisy Data by Structured Low Rank Approximation

Laurent Condat
GIPSA-lab, University of Grenoble, Grenoble, France
Contact: see http://www.gipsa-lab.grenoble-inp.fr/∼laurent.condat/

Akira Hirabayashi
Yamaguchi University, Yamaguchi, Japan

***Abstract*—We consider the recovery of a finite stream of Dirac pulses at nonuniform locations, from noisy lowpass-filtered samples. We show that maximum-likelihood estimation of the unknown parameters amounts to solve a difficult, even believed NP-hard, matrix problem of structured low rank approximation. We propose a new heuristic iterative optimization algorithm to solve it. Although it comes, in absence of convexity, with no convergence proof, it converges in practice to a local solution, and even to the global solution of the problem, when the noise level is not too high. Thus, our method improves upon the classical Cadzow denoising method, for same implementation ease and speed.**

I. Introduction and Problem Formulation

Reconstruction of signals lying in linear spaces, including bandlimited signals and splines, has long been the dominant paradigm in sampling theory, rooted in Shannon's work. Recently, analog reconstruction from discrete samples has been enlarged to a broader class of signals, with so-called finite rate of innovation, i.e. ruled by parsimonious models [1]–[3]. This theory predates and parallels the emerging framework of sparse recovery and compressed sensing [4]. The most studied problem in this context, on which we focus in this paper, is the recovery of a finite stream of Dirac pulses, a.k.a. a spike train, from uniform, noisy, lowpass-filtered samples [1], [5]–[8].

More precisely, the sought-after unknown signal s consists of K Dirac pulses in the finite interval $[0, \tau[$, where the real $\tau > 0$ and the integer $K \geq 1$ are known; that is

$$s(t) = \sum_{k=1}^{K} a_k \delta(t - t_k), \quad \forall t \in [0, \tau[, \tag{1}$$

where $\delta(t)$ is the Dirac mass distribution, $\{t_k\}_{k=1}^K$ are the unknown distinct locations in $[0, \tau[$, and $\{a_k\}_{k=1}^K$ are the unknown real nonzero amplitudes. The goal is to obtain estimates of these $2K$ values, which forms a deterministic (non-Bayesian) parametric estimation problem. The available data are, classically, linear uniform noisy measurements $\{v_n\}_{n=0}^{N-1}$ on s, of the form

$$v_n = \int_0^\tau s(t)\varphi\Big(\frac{n\tau}{N} - t\Big)dt + \varepsilon_n \tag{2}$$

$$= \sum_{k=1}^{K} a_k \varphi\Big(\frac{n\tau}{N} - t_k\Big) + \varepsilon_n, \quad \forall n = 0, \ldots, N-1, \tag{3}$$

where $\varphi(t)$ is the sampling function and the $\varepsilon_n \sim \mathcal{N}(0, \sigma^2)$ are independent random realizations of Gaussian noise. Note that other noise models could be considered as well, by changing the cost function in eqns. (5), (7), (9) below.

The questions of the choice of the function φ and of the number N of measurements allowing perfect reconstruction, in absence of noise, has been addressed in the literature [6], [7], [9]. In a nutshell, the condition $N \geq 2K + 1$, which we hereafter assume to be true, is necessary and sufficient, provided that φ satisfies some constraints in Fourier domain. Additionally, we assume, without loss of generality and only to simplify the notations, that N is odd, of the form $N = 2M + 1$. Since our emphasis here is on appropriately handling the presence of noise and not on being the most general, we adopt the simplest choice of the Dirichlet sampling function [6], which amounts to periodizing the signal s on the real line before sampling it with the sinc kernel:

$$\varphi(t) = \frac{\sin(N\pi t/\tau)}{N \sin(\pi t/\tau)} = \frac{1}{N} \sum_{m=-M}^{M} e^{j2\pi m t/\tau}, \quad \forall t \in \mathbb{R}. \tag{4}$$

The extension of the setting to the reconstruction of pulses with real shape, instead of the ideal Dirac distribution, is of obvious practical interest in ultrawideband communications [2] or to detect impulsive signals in biomedical applications [6]. This generalization, or equivalently the choice of another sampling function φ, can be done without difficulty, as shown in [6], and will not be addressed here.

The paper is organized as follows. In Sect. II, we formulate the maximum likelihood estimation problem and in Sect. III, we show that it amounts to a low rank matrix approximation problem. The new algorithm to solve it is presented in Sect. IV.

II. Maximum Likelihood Parameter Estimation

A natural approach to solve parametric estimation problems is maximum likelihood (ML) estimation; it consists in selecting the model which is the most likely to explain the observed noisy data. In our case, as we have assumed Gaussian noise, this corresponds to solving the nonlinear least-squares problem [10]:

$$\underset{\substack{\{t'_k\}_{k=1}^K \in [0,\tau[^K \\ \{a'_k\}_{k=1}^K \in \mathbb{R}^K}}{\text{minimize}} \sum_{n=0}^{N-1} \left| v_n - \sum_{k=1}^{K} a'_k \varphi\Big(\frac{n\tau}{N} - t'_k\Big) \right|^2. \tag{5}$$

Now, applying the discrete Fourier transform to the vector of samples $\{v_n\}_{n=0}^{N-1}$ yields the Fourier coefficients defined by

$\hat{v}_m = \sum_{n=0}^{N-1} v_n e^{-j2\pi mn/N}, \quad \forall m = -M, \dots, M$. We define the Fourier coefficients $\{\hat{\varepsilon}_m\}_{m=-M}^{M}$ similarly. Then, it is easy to show that

$$\hat{v}_m = \sum_{k=1}^{K} a_k e^{-j2\pi m t_k/\tau} + \hat{\varepsilon}_m, \quad \forall m = -M, \dots, M. \quad (6)$$

Since the inverse discrete Fourier transform is unitary, up to a constant, the problem (5) can be rewritten as [10]:

$$\underset{\substack{\{t'_k\}_{k=1}^{K} \in [0,\tau[^K \\ \{a'_k\}_{k=1}^{K} \in \mathbb{R}^K}}{\text{minimize}} \quad \sum_{m=-M}^{M} \left| \hat{v}_m - \sum_{k=1}^{K} a'_k e^{-j2\pi m t'_k/\tau} \right|^2. \quad (7)$$

Thus, (7) takes the form of a spectral estimation problem, which consists in retrieving the parameters of a sum of complex exponentials from noisy samples [11]. However, solving (7) is very difficult task, as the function to minimize is oscillating, with many local minima [12]. Numerous methods have been proposed to find a local minimum of the cost function in (7). They mostly proceed by iteratively refining an initial estimate of the solution, which has to be already of good quality. Also, when $N \gg K$ and the locations t_k are not too close to each other, classical spectral estimation techniques like MUSIC and ESPRIT can be used; they are fast but statistically suboptimal. The main advantage of the proposed approach is that it gets rid of such limitations, without any simplifying assumption.

III. Prony's Annihilation Property: Reformulation as Matrix Approximation Problem

Let us assume temporarily that there is no noise, i.e. $\hat{\varepsilon}_m = 0$ in (6). Then, the sequence of Fourier coefficients $\{\hat{v}_m\}_{m=-M}^{M}$ can be *annihilated*, a known property which dates back to Prony's work in the eighteenth century [13]. That is, its convolution with the sequence $\{h_k\}_{k=0}^{K}$ is identically zero: $\sum_{k=0}^{K} h_k \hat{v}_{m-k} = 0, \quad \forall m = -M+K, \dots, M$, where the annihilating filter h is defined, up to a constant, by $\sum_{k=0}^{K} h_k z^k = \prod_{k=1}^{K} (z - e^{j2\pi t_k/\tau})$. In matrix form, the annihilation property is

$$\underbrace{\begin{pmatrix} \hat{v}_{-M+K} & \cdots & \hat{v}_{-M} \\ \vdots & \ddots & \vdots \\ \vdots & \ddots & \vdots \\ \hat{v}_{M} & \cdots & \hat{v}_{M-K} \end{pmatrix}}_{\mathbf{T}_K} \begin{pmatrix} h_0 \\ \vdots \\ h_K \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}. \quad (8)$$

Let us choose an integer P in $K, \dots, M$ and define the Toeplitz—i.e. with constant values along its diagonals—matrix $\mathbf{T}_P$, of size $N - P \times P + 1$, obtained by arranging the values $\{\hat{v}_m\}_{m=-M}^{M}$ in its first row and column; $\mathbf{T}_K$ is depicted in (8). Then, the existence of an annihilating filter of size $K+1$ for the sequence $\{\hat{v}_m\}_{m=-M}^{M}$ is completely equivalent to the property that $\mathbf{T}_P$ has rank at most K.

Hence, turning back to the case when noise is present in the data, we can rewrite (7) as the following *structured low rank approximation* (SLRA) matrix problem:

$$\text{Find } \widetilde{\mathbf{T}}_P \in \underset{\mathbf{T}' \in \mathbb{C}^{N-P\times P+1}}{\arg\min} \|\mathbf{T}' - \mathbf{T}_P\|_w^2 \quad \text{s. t.} \quad \mathbf{T}' \text{ is Toeplitz} \quad \text{and} \quad \operatorname{rank}(\mathbf{T}') \le K, \quad (9)$$

where the weighted Frobenius norm of a matrix $\mathbf{A} = \{a_{i,j}\} \in \mathbb{C}^{N-P\times P+1}$ is defined by $\|\mathbf{A}\|_w^2 = \sum_{i=1}^{N-P} \sum_{j=1}^{P+1} w_{i,j} |a_{i,j}|^2$ and $w_{i,j}$ is the inverse of the size of the diagonal going through the position (i,j), see formula in [14, eq. (16)].

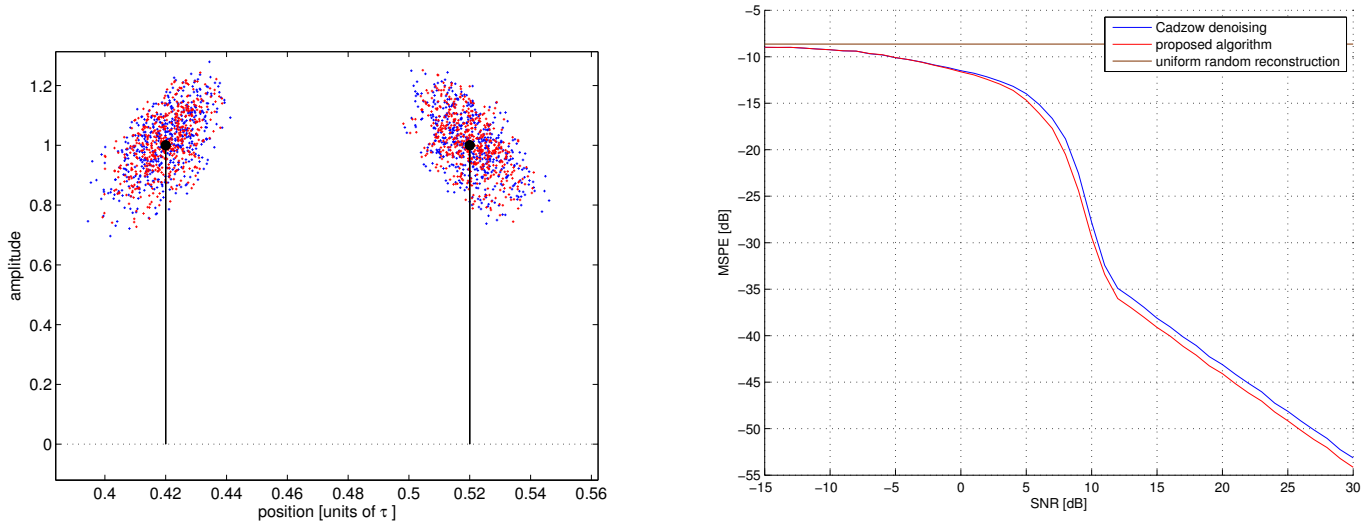
After the SLRA problem (9) has been solved, the procedure to recover the estimates of the parameters is the following [1]. First, reshape the obtained Toeplitz matrix $\widetilde{\mathbf{T}}_P$ to a Toeplitz matrix $\widetilde{\mathbf{T}}_K$ of size $N - K \times K + 1$. Second, compute the right singular vector $\tilde{\mathbf{h}} = \{\tilde{h}_k\}_{k=0}^{K}$ of $\widetilde{\mathbf{T}}_K$, corresponding to the singular value 0. Third, compute the roots $\{\tilde{z}_k\}_{k=1}^{K}$ of the polynomial $\sum_{k=0}^{K} \tilde{h}_k z^k$; the estimates $\{\tilde{t}_k\}_{k=1}^{K}$ of the locations are given by $\tilde{t}_k = \frac{\tau}{2\pi} \arg_{[0,2\pi[}(\tilde{z}_k)$. Fourth, the estimates $\{\tilde{a}_k\}_{k=1}^{K}$ of the amplitudes are obtained by solving the linear system $\widetilde{\mathbf{U}}^{\mathrm{H}} \widetilde{\mathbf{U}} \tilde{\mathbf{a}} = \widetilde{\mathbf{U}}^{\mathrm{H}} \hat{\mathbf{v}}$, where $\hat{\mathbf{v}} = [\hat{v}_{-M} \ \cdots \ \hat{v}_M]^{\mathrm{T}}$, $.^{\mathrm{H}}$ denotes the Hermitian transpose, and

$$\widetilde{\mathbf{U}} = \begin{pmatrix} e^{j2\pi M \tilde{t}_1/\tau} & \dots & e^{j2\pi M \tilde{t}_K/\tau} \\ \vdots & \vdots & \vdots \\ e^{-j2\pi M \tilde{t}_1/\tau} & \dots & e^{-j2\pi M \tilde{t}_K/\tau} \end{pmatrix}. \quad (10)$$

We note that this procedure yields the ML estimates solution to (7), only if the roots $\{\tilde{z}_k\}_{k=1}^{K}$ are all on the complex unit circle. This is the case, by centro-Hermitian symmetry of the matrices, except if the noise level is too high; in this case, two roots could merge and then split in a pair $(\tilde{z}_k, \tilde{z}_{k'} = 1/\tilde{z}_k^*)$ on both sides of the unit circle, yielding $\tilde{t}_k = \tilde{t}_{k'}$.

Thus, the proposed process consists in *denoising* the matrix $\mathbf{T}_P$, or equivalently the measurements $\{v_n\}_{n=0}^{N-1}$, by finding the closest data consistent with the model's structure, from which the parameters are estimated by Prony's method. In absence of noise, the parameters are perfectly recovered. However, the SLRA problem (9) at the heart of the procedure, which consists in projecting a matrix on a nonconvex manifold, is believed to be NP-hard [15]. Yet, the main advantage of the SLRA formulation, compared to (7), is that there is no initialization problem: an iterative algorithm to solve (9) proceeds directly, with the noisy matrix $\mathbf{T}_P$ as initial estimate of the solution $\widetilde{\mathbf{T}}_P$. Moreover, for a low noise level, an algorithm converging to a local solution will actually find the global solution $\widetilde{\mathbf{T}}_P$, as we observe in practice.

We now tackle the state-of-the-art to solve SLRA problems, which have a wide range of applications [15]. A few algorithms, able to find a local solution of the SLRA problem (9), have been proposed in the community of numerical algebra [16]–[18]. For instance, the iterative approach in [16] is based on a BFGS quasi-Newton solver. Besides the difficulty of implementation, the algorithm is very costly, as it requires computing many singular value decompositions (SVD) at each iteration. To our knowledge, the only publicly available software package for SLRA is the one currently in development by Ivan Markovsky [19]. However, it only handles real-valued,



(a) (b)

Fig. 1. The signal to estimate from $N = 11$ noisy measurements, consists in $K = 2$ Dirac pulses. The true parameters are $(t_1, t_2) = (0.42, 0.52)$ and $(a_1, a_2) = (1, 1)$, with $\tau = 1$ and $P = M = 5$. (a) In black, the true pulses. In blue and red, the locations and amplitudes reconstructed by Cadzow denoising and the proposed algorithm, respectively, for 500 different noise realizations. The signal-to-noise-ratio (SNR) was 15dB and the computation time for every reconstruction, with 50 iterations, was 14ms. The proposed method yields lower errors, with a points cloud slightly less dispersed. (b) Plot in log-log scale of the mean squared periodic error (MSPE) on the locations $\min\left((\tilde{t}_1 - t_1)_\tau^2 + (\tilde{t}_2 - t_2)_\tau^2, (\tilde{t}_1 - t_2)_\tau^2 + (\tilde{t}_2 - t_1)_\tau^2\right)$, where $(x)_\tau = \left((x + \frac{\tau}{2}) \mod \tau\right) - \frac{\tau}{2}$, averaged over 10,000 noise realizations for every SNR value. An upper bound of the error is given by the naive estimator, which sets the locations randomly and uniformly in $[0, \tau[$.

and not complex-valued, matrices. We note that replacing in the problem the rank by its convex surrogate, the nuclear norm, does not perform well in our setting, where two close pulses yield highly coherent measurements [20]. Thus, practitioners rely on a popular heuristic method, called *Cadzow denoising* [21], which is used in [1], [6] for the recovery of Dirac pulses. This algorithm consists in denoising the matrix $\mathbf{T}_P$ by alternating projections: at each iteration, the matrix is replaced by its closest, in Frobenius norm, matrix of rank at most K, and then the obtained matrix is replaced by its closest Toeplitz matrix. Although Cadzow denoising seems to always converge in practice to a Toeplitz matrix of rank at most K, there exists no global proof of convergence to date, contrary to a common belief [22]. Anyways, the obtained matrix is not a local minimizer of the cost function $\| \cdot - \mathbf{T}_P\|_w^2$ [12], [16]. In the next section, we propose a new algorithm to compute a local solution of the SLRA problem (9), thus improving theoretically upon Cadzow denoising.

IV. A New Optimization Method for SLRA

Let us consider the generic optimization problem:

$$\text{Find } \tilde{x} \in \operatorname*{arg\,min}_{x \in \mathcal{H}} F(x) \quad \text{s.t.} \quad x \in \Omega_1 \cap \Omega_2, \qquad (11)$$

where $\mathcal{H}$ is a real Hilbert space of finite dimension, Ω_1 and Ω_2 are two closed subsets of $\mathcal{H}$, and $F : \mathcal{H} \to \mathbb{R}$ is a differentiable function with Lipschitz-continuous gradient; that is, there exists some $\beta > 0$ such that $\|\nabla F(x') - \nabla F(x)\| \leq \beta \|x - x'\|$, $\forall x, x' \in \mathcal{H}$. Recently [23], the first author proposed a new algorithm to solve (11):

Optimization algorithm. Choose the parameters $\mu > 0$, $\gamma \in]0, 1[$, and the initial elements $x^{(0)}, s^{(0)} \in \mathcal{H}$. Then iterate, for every $i \geq 0$,

$$\left| \begin{array}{l} x^{(i+1)} = P_{\Omega_1}\left(s^{(i)} + \gamma(x^{(i)} - s^{(i)}) - \mu \nabla F(x^{(i)})\right) \\ s^{(i+1)} = s^{(i)} - x^{(i+1)} + P_{\Omega_2}(2x^{(i+1)} - s^{(i)}) \end{array} \right. ,$$

where P_Ω denotes the closest-point projection onto $\Omega \subset \mathcal{H}$. It has been proved in [23] that if Ω_1 and Ω_2 are convex and $2\gamma > \beta\mu$, the sequence $(x^{(i)})_{i \in \mathbb{N}}$ converges to some element $\tilde{x}$ solution to the problem (11).

In absence of convexity, this result does not apply, so that we will use the method as a heuristic, without guarantee of convergence. The SLRA problem (9) can be recast as an instance of (11) as follows: $\mathcal{H} = \mathbb{C}^{N-P \times P+1}$ is the real Hilbert space of complex-valued matrices of size $N-P \times P+1$ with centro-Hermitian symmetry, endowed with Frobenius inner product $\langle \mathbf{X}, \mathbf{X}' \rangle = \sum_{i,j} x_{i,j} x'^*_{i,j}$; Ω_1 is the closed nonconvex subset of $\mathcal{H}$ of matrices with rank at most K; Ω_2 is the linear subspace of $\mathcal{H}$ of Toeplitz matrices. The operations involved in the algorithm are the following:

- P_{Ω_1} corresponds to SVD truncation, according to the Schmidt-Eckart-Young theorem: if a matrix $\mathbf{X}$ has SVD $\mathbf{X} = \mathbf{L}\mathbf{\Sigma}\mathbf{R}^{\mathrm{H}}$, then $P_{\Omega_1}(\mathbf{X})$ is obtained by setting to zero the singular values in $\mathbf{\Sigma}$, except the K largest.
- The "Toeplitzation" operation P_{Ω_2} simply consists in averaging along the diagonals of the matrix.
- The cost function is $F(\mathbf{X}) = \frac{1}{2}\|\mathbf{X} - \mathbf{T}_P\|_w^2$, so that $\nabla F(\mathbf{X}) = \mathbf{W} \circ (\mathbf{X} - \mathbf{T}_P)$, where $\circ$ is the entrywise (Hadamard) product and the matrix $\mathbf{W}$ has entries $\{w_{i,j}\}$.

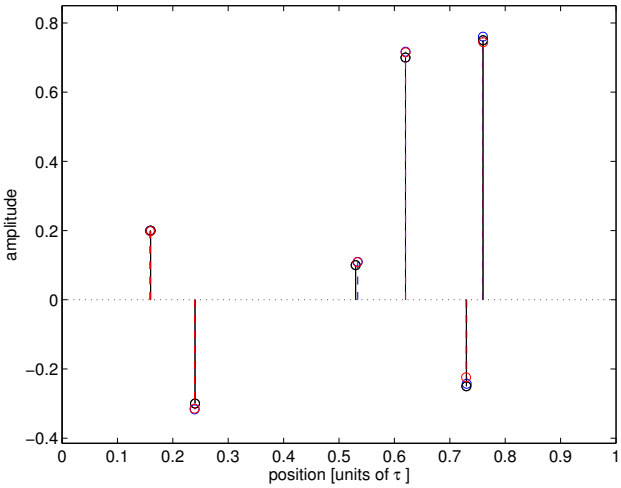


Fig. 2. The signal consists of $K = 6$ Dirac pulses. We have $N = 25$ noisy measurements with SNR=25dB. In black: true pulses. In blue and red: reconstructed positions and amplitudes of the pulses with Cadzow denoising and the proposed algorithm, respectively. The computation time, with 50 iterations, was 19ms in both cases.

The Lipschitz constant of ∇F is $\beta = \max(\{w_{i,j}\}) = 1$.

We observed empirically that the proposed algorithm always converges, for an appropriate choice of μ and γ. Moreover, the matrix obtained at convergence is always Toeplitz, of rank at most K, and a local solution to (9); see more details in [14].

We show in Fig. 1 a comparison with Cadzow denoising for the recovery of $K = 2$ Dirac pulses from $N = 11$ measurements. We observe that the estimation error on the pulses' locations is about 10% lower in average with our method. We recognize that this improvement is small for the simple setting considered here, with ideal Dirac pulses and a sinc sampling kernel. Our ongoing work is to investigate more general scenarios, with pulses having real shape and noise which is not white and Gaussian. We expect the improvement of our method over Cadzow denoising to be more significant in such cases. Yet, we emphasize that both methods have essentially the same complexity and convergence speed, dominated by one SVD per iteration. Another example is given in Fig.2 and experiments with larger size are shown in the extended version of this paper [14].

V. Conclusion

We proposed a new heuristic optimization algorithm to solve structured low rank approximation problems. For the recovery of Dirac pulses, this efficient matrix denoising procedure, combined with Prony's extraction method, yields the maximum-likelihood parameter estimates, up to some threshold SNR. Many theoretical questions related to the performances of the approach are open and currently investigated by the authors. Especially, stability guarantees similar to the ones recently developed for a convex relaxation of the problem [8], [24], are sought after. A Matlab implementation of the proposed method is available on the webpage of the first author.

Acknowledgment

The first author started this work at Yamaguchi University, invited by the second author, thanks to a three-months fellowship of the Japanese Society for the Promotion of Science.

References

[1] T. Blu, P.-L. Dragotti, M. Vetterli, P. Marziliano, and L. Coulot, "Sparse sampling of signal innovations," *IEEE Signal Processing Mag.*, vol. 25, no. 2, pp. 31–40, Mar. 2008, Special issue on Compressive Sampling.

[2] M. Mishali and Y. C. Eldar, "Sub-Nyquist sampling: Bridging theory and practice," *IEEE Signal Processing Mag.*, vol. 28, no. 6, pp. 98–124, Nov. 2011.

[3] J. Urigüen, Y. C. Eldar, P. L. Dragotti, and Z. Ben-Haim, "Sampling at the rate of innovation: Theory and applications," in *Compressed Sensing: Theory and Applications*, Y. C. Eldar and G. Kutyniok, Eds. Cambridge University Press, 2012.

[4] T. Strohmer, "Measure what should be measured: Progress and challenges in compressive sensing," *IEEE Signal Processing Lett.*, vol. 19, no. 12, pp. 887–893, Dec. 2012.

[5] M. Vetterli, P. Marziliano, and T. Blu, "Sampling signals with finite rate of innovation," *IEEE Trans. Signal Processing*, vol. 50, no. 6, pp. 1417–1428, June 2002.

[6] R. Tur, Y. C. Eldar, and Z. Friedman, "Innovation rate sampling of pulse streams with application to ultrasound imaging," *IEEE Trans. Signal Processing*, vol. 59, no. 4, pp. 1827–1842, Apr. 2011.

[7] K. Gedalyahu, R. Tur, and Y. C. Eldar, "Multichannel sampling of pulse streams at the rate of innovation," *IEEE Trans. Signal Processing*, vol. 59, no. 4, pp. 1491–1504, Apr. 2011.

[8] E. J. Candès and C. Fernandez-Granda, "Super-resolution from noisy data," preprint arXiv:1211.0290, 2012.

[9] Z. Ben-Haim, T. Michaeli, and Y. C. Eldar, "Performance bounds and design criteria for estimating finite rate of innovation signals," *IEEE Trans. Inform. Theory*, vol. 58, no. 8, pp. 4993–5015, Aug. 2012.

[10] A Hirabayashi, T. Iwami, S. Maeda, and Y. Hironaga, "Reconstruction of the sequence of Diracs from noisy samples via maximum likelihood estimation," in *Proc. of IEEE ICASSP*, 2012, pp. 3805–3808.

[11] P. Stoica and R. Moses, *Spectral Analysis of Signals*, Prentice Hall, NJ, 2005.

[12] J. Gillard and A. Zhigljavsky, "Analysis of structured low rank approximation as an optimization problem," *Informatica*, vol. 22, no. 4, pp. 489–505, 2011.

[13] T. Blu, "The generalized annihilation property—A tool for solving finite rate of innovation problems," in *Proc. of Int. Workshop on Sampling Theory and Appl. (SampTA)*, Marseille, France, May 2009.

[14] L. Condat and A. Hirabayashi, "Cadzow denoising upgraded: A new projection method for the recovery of Dirac pulses from noisy linear measurements," preprint hal-00759253, 2012.

[15] I. Markovsky, *Low Rank Approximation: Algorithms, Implementation, Applications*, Springer, 2012.

[16] M. T. Chu, R. E. Funderlic, and R. J. Plemmons, "Structured low rank approximation," *Linear Algebra Appl.*, vol. 366, pp. 157–172, 2003.

[17] M. Schuermans, *Weighted low rank approximation: Algorithms and applications*, Ph.D. thesis, Katholieke Universiteit Leuven, 2006.

[18] R. Borsdorf, *Structured Matrix Nearness Problems: Theory and Algorithms*, Ph.D. thesis, The University of Manchester, UK, June 2012.

[19] I. Markovsky and K. Usevich, "Software for weighted structured low-rank approximation," Tech. Rep. 339974, ECS, Univ. of Southampton, 2012, documention of a software package, see `https://github.com/slra/slra`.

[20] I. Markovsky, "How effective is the nuclear norm heuristic in solving data approximation problems?," in *Proc. of IFAC Symposium on System Identification (SYSID)*, Brussels, Belgium, July 2012.

[21] J. A. Cadzow, "Signal enhancement—A composite property mapping algorithm," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, no. 1, pp. 49–62, Jan. 1988.

[22] D. R. Luke, "Prox-regularity of rank constraint sets and implications for algorithms," preprint arXiv:1112.0526, 2011.

[23] L. Condat, "A primal-dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms," *J. Optim. Theory and Appl.*, 2013, to appear.

[24] B. N. Bhaskar, G. Tang, and B. Recht, "Atomic norm denoising with applications to line spectral estimation," preprint arXiv:1204.0562, 2012.

Multichannel ECG Analysis Using VPW-FRI

Amrish Nair*, Pina Marziliano*, R. Frank Quick†, Ronald. E. Crochiere† and Gilles Baechler‡
*School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
†Qualcomm Inc., San Diego, CA 92121, USA
‡School of Computer and Communication Sciences, EPFL, Lausanne, Switzerland

***Abstract*—In this paper, we present an application of Variable Pulse Width Finite Rate of Innovation (VPW-FRI) in dealing with multichannel Electrocardiogram (ECG) data using a common annihilator. By extending the conventional FRI model to include additional parameters such as pulse width and asymmetry, VPW-FRI has been able to deal with a more general class of pulses. The common annihilator, which is introduced in the annihilating filter step, shows a common support in multichannel ECG data, which provides interesting possibilities in compression. A model based de-noising method will be presented which is fast and non-iterative. Also, an application to detect QRS complexes in ECG signals will be demonstrated. The results will show the robustness of the common annihilator and the QRS detection even in the presence of noise.**

I. Introduction

The concept of sampling and reconstructing signals at the rate of innovation was first presented by *Vetterli et al.* [1]. They showed that non band-limited classes of signals such as streams of Diracs had a finite number of degrees of freedom and could be completely defined by their location and amplitude parameters. These classes of signals were termed Finite Rate of Innovation (FRI) signals. These FRI signals could be sampled minimally at the rate of innovation and perfectly reconstructed.

Variable Pulse Width FRI (VPW-FRI) was developed by *Quick et al.* [2] as an extension of the traditional FRI method in that it added two additional parameters, namely the pulse width and asymmetry, to the model. This allows it some flexibility in dealing with pulses of various forms and widens the scope of its application. It does this by considering roots which fall inside the unit circle as compared to traditional FRI where the roots lie on the unit circle.

The generalisation of Diracs in VPW-FRI allowed it to be used successfully in compression of Electrocardiogram (ECG) signals [2], [3] where the P, QRS and T waveforms could be represented by pulses of varying amplitude, width and asymmetry. This allowed for a compression scheme which only requires 7 pulses per beat, with 4 parameters per pulse, which is far below the Nyquist rate of around $200-250$Hz at which most devices record ECG signals.

Other methods have also been used for compression such as compressed sensing [7], wavelet methods [8] and finite rate of innovation [4]. The FRI method in [4] divides the ECG signal into two parts. The QRS is modelled as a non-uniform linear spline while the remainder of the signal is considered a residual signal which is sampled at a low rate of $15Hz$. The difference here is that VPW-FRI considers each waveform, P, QRS and T, as a pulse and parameterizes them accordingly. Using VPW-FRI allows for a much lower number of samples and higher compression ratio.

In this paper, we demonstrate a multichannel approach to calculating the location parameter. To achieve this, a common annihilator is used in the reconstruction step to derive the locations. This aids in the compression of the multichannel signal and has potential applications such as QRS detection which we will also present.

Also, VPW-FRI has de-noising capabilities. This is achieved through a model based de-noising method [5] which is fast and non-iterative. This is its main advantage especially when compared to Cadzow [2], [6] denoising which is iterative and requires oversampling. Most importantly, de-noising is done without affecting the morphology of the pulses which is especially important when clinicians examine an ECG recording.

This paper is organised as follows. Section II will present some background on FRI theory followed by an explanation of VPW-FRI. Section III will demonstrate the multichannel VPW-FRI approach. This will be followed by Section IV where an application of VPW FRI in ECG wave detection will be shown. The 12 lead ECG data used and the results will be presented in Section V. Finally, conclusions will be drawn and some thoughts on future work will constitute Section VI.

II. Variable Pulse Width Finite Rate of Innovation

Since VPW-FRI is an extension of the original FRI theory, we will present a short description of FRI theory followed by the changes in the VPW-FRI algorithm.

A. FRI

A stream of K Diracs with period τ is defined by

$$x(t) = \sum_{k=0}^{K-1} b_k \delta(t - t_k) \quad (1)$$

$$= \underbrace{\sum_{m\in\mathbb{Z}} \frac{1}{\tau} \sum_{k=0}^{K-1} b_k \underbrace{e^{-i(2\pi m t_k)/\tau}}_{u_k^m} e^{i(2\pi m t)/\tau}}_{X[m]} \quad (2)$$

where Eq. (2) is the Poisson Summation Formula derivation of Eq. (1). The signal is then sampled uniformly. The samples, y_n are defined by

$$y_n = \langle h_b(t-nT), x(t)\rangle, \quad n=0,...,N-1 \tag{3}$$
$$= \sum_{m=-M}^{M} X[m]e^{i(2\pi mnT/\tau)}, \tag{4}$$

where T represents the sampling period, N is the number of samples, $B \geq \frac{2K}{\tau}$, $M = \lfloor B\tau/2 \rfloor$ and the sampling kernel $h_b(t) = Bsinc(Bt)$.

In the reconstruction step, the annihilating filter [1] in Eq. (5) is used to retrieve the u_k values

$$\begin{bmatrix} X[-1] & \dots & X[-K] \\ X[0] & \dots & X[-K+1] \\ \vdots & \ddots & \vdots \\ X[K-2] & \dots & X[-1] \end{bmatrix} \cdot \begin{bmatrix} A[1] \\ A[2] \\ \vdots \\ A[K] \end{bmatrix} = 0, \tag{5}$$

where $A[k]$ represents the annihilating filter coefficients.

A common way of solving for A would be to find the minimal right singular vector of the Toeplitz matrix in Eq. (5). Since the filter coefficients are of the form

$$A(z) = \sum_{k=0}^{K} A[k]z^{-k} = \prod_{k=0}^{K-1}(1-u_k z^{-1}), \tag{6}$$

the roots of the filter coefficients would correspond to u_k, defined in Eq. (2), and the locations, t_k can be calculated directly from u_k.

The amplitudes, b_k, can be resolved using a Vandermonde system of equations [1].

B. Sampling and Reconstruction of VPW-FRI signals

The Dirac model can be generalised with the addition of width and asymmetry parameters. This can be used to expand FRI theory by interpreting the u_k and $X[m]$ coefficients differently. The u_k values are defined as

$$u_k = e^{-2\pi(a_k+it_k)/\tau}, \quad a_k \geq 0 \tag{7}$$

where a_k is the width parameter. The $X[m]$ coefficients are defined as

$$X[m] = X^{(1)}[m] + X^{(2)}[m], \tag{8}$$

where

$$X^{(1)}[m] = \sum_{k=0}^{K-1} c_k e^{-2\pi(a_k|m|+it_k m)/\tau} \tag{9}$$

and

$$X^{(2)}[m] = -\sum_{k=0}^{K-1} d_k sgn(m) e^{-2\pi(a_k|m|+it_k m)/\tau}. \tag{10}$$

The $X^{(2)}[m]$ coefficients are the Hibert transform of $X^{(1)}[m]$ and the spectra of $X^{(1)}[m]$ and $X^{(2)}[m]$ are symmetric.

The same annihilating filter in Eq. (5) can be used. For stability, the annihilating filter roots which lie within the unit circle are admitted and those which lie outside are rejected. The t_k and a_k parameters can be retrieved from the roots of the annihilating filter coefficients.

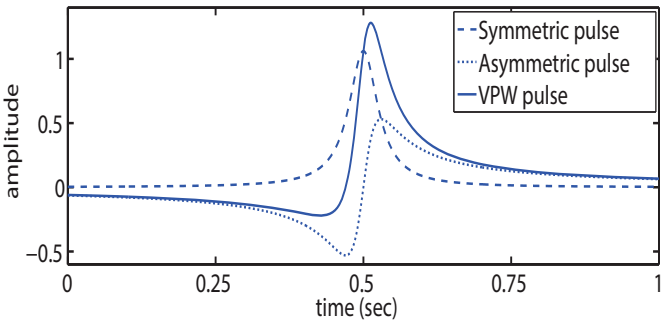


Fig. 1. Symmetric and asymmetric components of a VPW-FRI pulse

The $\{c_k\}_{k=0}^{K-1}$ and $\{d_k\}_{k=0}^{K-1}$ coefficients, which are the real and imaginary part of b_k respectively, can be solved using the Vandermonde system [1] over the complex numbers as compared to the original FRI theory where it is solved over the real numbers.

The continuous-time signal, $x(t)$ can be recovered by applying the inverse Fourier Transform,

$$x(t) = \sum_{k=0}^{K=1} x_k(t) \tag{11}$$
$$= \sum_{k=0}^{K-1}\sum_{n\in\mathbb{Z}} c_k \frac{a_k}{\pi(a_k^2+(t-t_k-n\tau)^2)}$$
$$+ \sum_{k=0}^{K-1}\sum_{n\in\mathbb{Z}} d_k \frac{t-t_k-n\pi}{\pi(a_k^2+(t-t_k-n\tau)^2)}.$$

An alternate formula for $x_k(t)$ that avoids the infinite sum is given by:

$$x_k(t) = \frac{c_k}{\tau}\frac{1-|z_t|^2}{(1-z_t)(1-z_t^*)} + \frac{d_k}{\tau}\frac{2\Im\{z_t\}}{(1-z_t)(1-z_t^*)} \tag{12}$$

where $z_t = e^{2\pi(-a_k+i(t-t_k))/\tau}$. As can be seen in Equation (11) and in Fig. 1, the VPW pulse consists of a symmetric and asymmetric pulse. The symmetric pulse is a Cauchy-Lorentz function and the asymmetric pulse is the Hilbert Transform of the symmetric pulse.

III. VPW-FRI ON MULTICHANNEL DATA

When dealing with multichannel data where the pulses occur at the same locations across all the channels, multi lead ECG for example, it would make sense to compute the locations for all the channels simultaneously rather than for each channel individually. This is achieved using a common annihilator which is the main mechanism that allows VPW-FRI to handle multichannel data.

In FRI theory, the annihilating filter would be where the u_k values are determined. However, it only deals with single channel information. Therefore by modifying the input to the annihilating filter, we can create a common annihilator for all the input channels. This can be achieved [11] by stacking the Toeplitz matrices of each channel vertically and applying the annihilating filter,

$$\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_M \end{bmatrix} \cdot \begin{bmatrix} A[1] \\ A[2] \\ \vdots \\ A[K] \end{bmatrix} = 0. \quad (13)$$

where $\{X_m\}_{m=1}^{M}$ represents the Toeplitz matrices of the M channels and $\{A[k]\}_{k=1}^{K}$ represents the annihilating filter coefficients similar to Eq. (6). The roots of the annihilating filter would yield the common locations of the pulses across all the channels.

A model based de-noising technique was implemented in this paper which is based on the subspace based approach presented in [5]. From Eq. (13), V is used to estimate the noiseless signal. Therefore,

$$\overline{V} = \underline{V} \cdot \Phi_H, \quad (14)$$

where $\overline{(\cdot)}$ and $\underline{(\cdot)}$ denote the operation of omitting the first and last row of $(\cdot)$, respectively. The conjugates of the eigenvalues of Φ_H will yield the roots of the annihilating filter and not the filter coefficients as seen in Eq. (5). For a detailed proof, please refer to [5].

The u_k values retrieved from the roots of the annihilating filter can be used to calculate the locations, t_k, for the pulses in all the channels.

This offers an interesting perspective especially when considering the physiology and the way the heart's electrical signals are recorded. The denominator of the filter describes the common activities such as time of arrival of the electrical vectors at the electrodes while the numerator captures the morphological information of the pulse. This could be studied further especially when developing automated diagnostic or wave detection tools.

IV. QRS Detection

One application of VPW-FRI, besides sampling and reconstruction, is QRS detection in ECG signals. Paired with the common annihilator method presented in Section III, this method of QRS detection is workable even in noisy signals.

The QRS complexes present the sharpest transition out of all the ECG waveforms. Hence, in Eq. (13), the highest values of $diag(S)$ would correspond to the QRS complexes due to the fact that it has the highest energy out of all the pulses. This can also be seen in the roots of the annihilating filter, as the roots closest to the unit circle would represent the QRS complexes though the distinction is not as clear. If we represent $A_n = \{S_{1,1}, S_{2,2}, \ldots, S_{N-1,N-1}\}$ and $B_n = \{S_{2,2}, S_{3,3}, \ldots, S_{N,N}\}$,

$$E_n = A_n / B_n, \quad n = 1, \ldots, N-1. \quad (15)$$

Then the number of QRS complexes can be found by thresholding E_n. Empirically, this threshold was found to be 1.2.

By only keeping the subspace associated with these QRS complexes, the QRS pulses can be accurately identified.

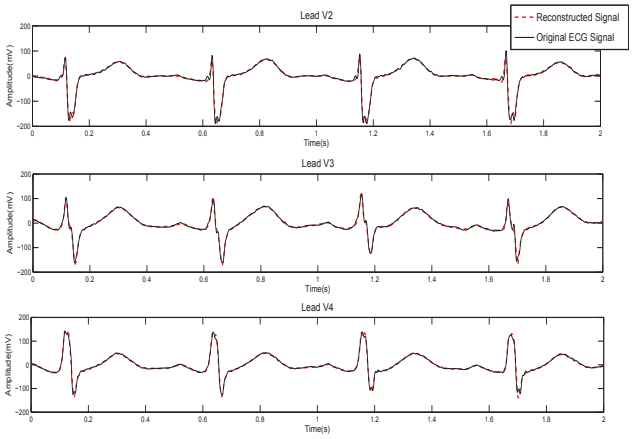


Fig. 2. Reconstructed signal for ECG leads V2-V4

V. Results

In this section, the ECG data that was used to generate results will be introduced. This will be followed by results from the VPW-FRI, common annihilator and QRS detection.

A. Data

The data used was 12 lead Stress ECG data recordings from Tan Tock Seng Hospital, Singapore. The subjects were patients who were undergoing treadmill ECG tests as recommended by their physician. All subjects voluntarily signed an agreement to have their anonymised data used for research purposes. The test conducted were under the conditions of the BRUCE protocol [9] which is a stress ECG protocol where the incline and speed of the treadmill are increased at intervals of 3mins.

The data was collected using the GE Marquette CASE Stress System with the T2100 treadmill. This data, collected from 6 patients, varied in length from 12 mins to 20 mins long depending on the patient's fitness level , cardiac health and the discretion of the physician. The leads recorded are I, II, III, aVR, aVL, aVF, V1, V2, V3, V4, V5 and V6. A simple and concise write up about the leads and their significance can be found in [10]. The data is sampled at $200Hz$.

B. Results

The reconstruction error of VPW-FRI used in this paper is the Signal to Residue Ratio (SRR) which is defined as

$$SRR = 10log\left(\frac{\sum_{n=0}^{N-1} x[n]^2}{\sum_{n=0}^{N-1}(x[n]-\hat{x}[n])^2}\right). \quad (16)$$

A hundred segments of data, each $2s$ long, were used to evaluate the performance of the reconstruction. Section V-A. One segment can be viewed in Fig. 2.

For the VPW-FRI with the common annihilator, the algorithm tested with a mean SRR of $\mu = 19.41dB$ with a standard deviation of $\sigma = 2.28dB$ as can be seen in Table I. The low standard deviation shows consistency in reconstructing all the channels using the common annihilator. The high SRR coupled with the low standard deviation also

TABLE I
SRR VALUES ACROSS ALL 12 ECG LEADS

Mean SRR	19.41
Standard Deviation	2.28
Minimum SRR	14.81
Maximum SRR	22.25

proves the theoretical prediction that the common annihilator would provide information on the common parameters of all the channels in ECG.

The segments run in this test were good quality signals as they were relatively free of noise. The purpose of this was to demonstrate the sampling and reconstruction ability of VPW-FRI. Seven pulses were used per QRS complex.

The de-noising capability of the algorithm is significant as can be seen in Fig. 3. Noise in the form of Additive Gaussian White Noise (AWGN) was added at an SNR of $10dB$ to simulate Electromyogram (EMG) or muscle noise. It should be noted that the Cadzow de-noising [2] performs similarly but is iterative and therefore more computationally intensive.

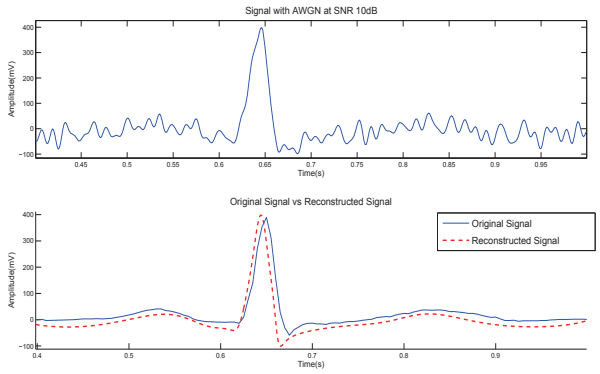


Fig. 3. Denoising on an ECG signal with AWGN at SNR 10dB

The QRS detection also tested well. Again, AWGN at SNR 0dB was added to test the robustness of the detection algorithm. The AWGN was added to all 12 channels. The QRS detector was then applied with only the pulse associated with the QRS being reconstructed. This was tested on the same 100 sets of signal used earlier in this section. A one second segment from lead II can be seen in Fig. 4. It was able to detect the number of QRS complexes and the locations perfectly on 97 of those segments. On the other 3 segments, it missed one QRS. However, when the SNR is raised to 5dB, it was able to detect all the QRS complexes in all the segments perfectly.

VI. CONCLUSION

The results demonstrate the robustness of the VPW-FRI method in compressing signals, in de-noising and also in wave detection. They also demonstrate that for the case of multichannel data, the ECG signals share a common support which translates to having a common denominator in the VPW-FRI model. This leads to additional opportunities for compression in the case of multichannel data. Future work

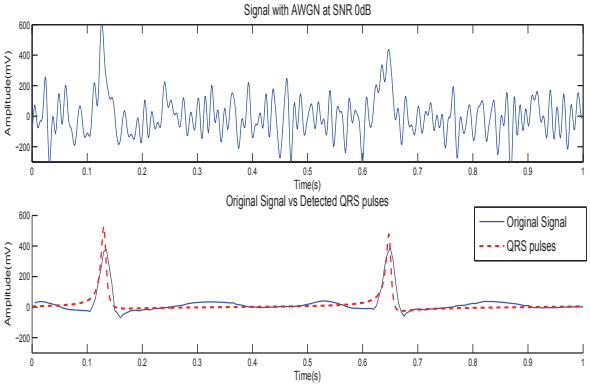


Fig. 4. QRS detection on ECG signal with AWGN at SNR 0dB

can be in the direction of application of VPW-FRI for feature detection in ECG as well as testing for compatibility with other biomedical signals.

ACKNOWLEDGMENT

The authors would like to thank Assistant Professor Dr. David Foo Chee Guan, Head of Cardiology Department, Tan Tock Seng Hospital, Singapore for his help and guidance in collecting the multi-lead ECG data.

REFERENCES

[1] M. Vetterli, P. Marziliano and T. Blu, "Sampling Signals With Finite Rate of Innovation", *IEEE Transactions on Signal Processing*, vol. 50, no. 6, pp. 1417-1428, June 2002.

[2] R. F. Quick, R. E. Crochiere, J. H. Hong, A. Hormati and G. Baechler, "Application of FRI to Modeling of Electrocardiogram Signals", *IEEE International Conference on Engineering in MEdicine and Biology 2012, San Diego*, pp. 2909-2912, August 2012.

[3] G. Baechler, N. Freris, R. F. Quick and R. E. Crochiere, "Finite Rate of Innovation Based Modeling and Compression of ECG Signals", *Submitted to IEEE International Conference on Acoustics, Speech and Signal Processing 2013*.

[4] Y. Hao, P. Marziliano, M. Vetterli, T. Blu, "Compression of ECG as a Signal with Finite Rate of Innovation", *27th IEEE International Conference on Engineering in Medicine and Biology*, pp. 7564-7567, January 2005

[5] I. Maravic and M. Vetterli, "Sampling and Reconstruction of Signals with Finite Rate of Innovation in the Presence Noise", *IEEE Transactions on Signal Processing*, vol.53, no.8, pp. 2788-2805, August 2005.

[6] T. Blu, P. L. Dragotti, M. Vetterli, P. Marziliano and L. Coulot, "Sparse Sampling of Signal Innovations", *IEEE Signal Processing Magazine*, vol. 25, no. 2, 2008.

[7] L. F. Polania, R. E. Carrillo, M. B. Velasco and K. E. Barner, "Compressed Sensing Based Method for ECG Compression", *in IEEE International Conference on Acoustics, Speech and Signal Processing*, pp.761-764, May 2011

[8] M. Nakashizuka, H. Kikuchi, H. Makino and I. Ishii, "ECG Data Compression by Multiscale Peak Analysis", *in IEEE International Conference on Acoustics, Speech and Signal Processing*, vol.2, pp.1105-1108, May 1995.

[9] Robert A. Bruce, Frank W. Lovejoy, Raymond Pearson, Paul N. G. Yu, George B. Brothers, Tulio Velasquez, "Normal respiratory and circulatory pathways of adaptation in exercise", *J. Clin. Invest 28(6 Pt2)*, pp. 1423-1430, November 1949

[10] "Basics-ECGpedia",*http://en.ecgpedia.org/wiki/Basics*, Viewed on 15th January 2013

[11] A. Hormati and M. Vetterli, "Compressive Sampling of Multiple SparseSignals Having Common Support Using Finite Rate of Innovation Principles", *IEEE Signal Processing Letters*, vol. 18, no. 5, pp. 331-334, May 2011

Recovery of bilevel causal signals with finite rate of innovation using positive sampling kernels

Gayatri Ramesh
Department of Mathematics
University of Central Florida
Orlando, FL 32816
Email: gayatriramesh@knights.ucf.edu

Elie Atallah
Department of Mathematics
University of Central Florida
Orlando, FL 32816
Email: elieatallah@knights.ucf.edu

Qiyu Sun
Department of Mathematics
University of Central Florida
Orlando, FL 32816
Email: qiyu.sun@ucf.edu

***Abstract*—Bilevel signal x with maximal local rate of innovation R is a continuous-time signal that takes only two values 0 and 1 and that there is at most one transition position in any time period of $1/R$. In this note, we introduce a recovery method for bilevel causal signals x with maximal local rate of innovation R from their uniform samples $x*h(nT), n \geq 1$, where the sampling kernel h is causal and positive on $(0, T)$, and the sampling rate $\tau := 1/T$ is at (or above) the maximal local rate of innovation R. We also discuss stability of the bilevel signal recovery procedure in the presence of bounded noises.**

I. Introduction

Let $T > 0$ and N be a nonnegative integer or infinity, and denote by χ_E the indicator function on a set E. In this note, we consider *bilevel causal signals*

$$x(t) := \sum_{i=1}^{N} \chi_{[t_{2i-1}, t_{2i})}(t) \tag{1}$$

with unknown transition values (positions) $t_i, 1 \leq i \leq 2N$, satisfying

$$t_i < t_{i+1}, \ 1 \leq i < 2N; \tag{2}$$

and also a uniform generalized sampling process

$$x(t) \longmapsto x * h(t) \longmapsto \{x * h(nT)\}_{n \geq 1} \tag{3}$$

with sampling kernel h being causal and uniform sampling taken every T seconds. For the bilevel causal signal x in (1), define its *maximal local rate of innovation* R by reciprocal of the maximal positive number σ_0 such that there is at most one transition position $t_i, 1 \leq i \leq 2N$, in any time period $[t, t + \sigma_0), t \geq 0$, that is,

$$R = \sup_{1 \leq i < 2N} \frac{1}{t_{i+1} - t_i}. \tag{4}$$

The concept of signals with finite rate of innovation was introduced by Vetterli, Marziliano and Blu [1]. Examples of signals with finite rate of innovation include streams of Diracs, piecewise polynomials, band-limited signals, and signals in a finitely-generated shift-invariant space [1]–[4]. In the past ten years, the paradigm for reconstructing signals with finite rate of innovation from their samples has been developed, see for instance [1], [2] and [4]–[13] and references therein.

Precise identification of transition positions is important to reach meaningful conclusions in many applications. Vetterli, Marziliano and Blu show in [1] that a bilevel signal x defined in (1) can be reconstructed from its samples (3) when the sampling kernel h is the box spline $\chi_{[0,T)}$ (or the hat spline $(T - |t|)\chi_{[-T,T)}(t)$) and the sample rate $\tau := 1/T$ is at (or above) the maximal local rate of innovation R of the signal x. In this note, we show that bilevel causal signals x defined in (1) are uniquely determined from their samples $x * h(nT), n \geq 1$, in (3) if the sampling kernel h is causal and positive on $(0, T)$, and the sample rate τ is at (or above) the maximal local rate of innovation R, see Theorem 1. Our numerical simulations indicate that the bilevel signal recovery procedure from noisy samples $x*h(nT)+\epsilon_n, n \geq 1$, is stable when there are limited numbers of transition positions for the bilevel signal x.

II. Recovery of bilevel causal signals

In this section, we provide a necessary condition on the sampling kernel h such that bilevel signals x in (1) are uniquely determined from their samples $\{x * h(nT)\}$ in (3). Also in this section, we propose an algorithm for the bilevel signal recovery.

The main theorem of this note is as follows:

Theorem 1: Let $T > 0$ and set $\tau = 1/T$. If h is a causal sampling kernel with $h(t) > 0$ on $(0, T)$, then any bilevel causal signal x in (1) with maximal local rate of innovation R being less than or equal to the sampling rate τ can be recovered from its samples $x * h(nT), n \geq 1$.

Proof: Let

$$H(t) = \int_0^t h(s)ds, \ 0 \leq t \leq T. \tag{5}$$

Then $H(0) = 0$ and H is a strictly increasing function on $[0, T)$ as h is strictly positive on $(0, T)$. Denote its inverse function on $[0, T]$ by $H^{-1} : [0, H(T)] \longmapsto [0, T]$.

Let x be a bilevel causal signal in (1) with transition positions $t_i, 1 \leq i \leq 2N$, satisfying (2). Then its first sample $y_1 = x * h(T)$ is given by

$$\begin{aligned} y_1 &= \int_0^\infty x(t)h\big(T - t\big)dt = \int_0^T x(t)h\big(T - t\big)dt \\ &= \int_0^T \chi_{[t_1,t_2)}(t)h\big(T - t\big)dt = H\big(\max\{T - t_1, 0\}\big), \end{aligned}$$

where the first two equalities hold by the causality of the signal x and the sampling kernel h, and the fourth equality follows from (1) and the observation that

$$t_i \geq t_2 = (t_2 - t_1) + t_1 \geq 1/R + 0 \geq 1/\tau = T, \ i \geq 2$$

by (2), (4) and the assumption that $R \leq \tau$. Recall that H is strictly increasing on $[0, T)$. Then there exists a transition position in the time range $[0, T)$ if and only if $y_1 = x*h(T) > 0$. Moreover, if it exists, it is given by

$$t_1 = T - H^{-1}(y_1). \tag{6}$$

Thus for a bilevel causal signal, we may determine from its first sample $x * h(T)$ the (non-)existence of its transition position in the time period $[0, T)$ and further its transition position in that time period if there is one.

Inductively, we assume that all transition positions of the bilevel signal x in the time range $[0, nT)$ have been determined from its samples $y_k = x * h(kT), 1 \leq k \leq n$. We examine four cases to determine its transition position in the time period $[nT, (n+1)T)$ from the sample $y_{n+1} = x * h((n+1)T)$.

Case 1: There is no transition position in $[0, nT)$.

In this case, following the above argument to determine transition positions in the time range $[0, T)$, we have that there exists a transition position in $[nT, (n+1)T)$ if and only if $y_{n+1} > 0$. If there is, the transition position is the first transition position t_1 of the bilevel causal signal x, and

$$t_1 = (n+1)T - H^{-1}(y_{n+1}). \tag{7}$$

Case 2: The last transition position in $[0, nT)$ is t_{2i_0-1} for some $i_0 \geq 1$.

In this case, $t_{2i_0} \geq nT$ and $t_i \geq (n+1)T$ for all $i > 2i_0$. Thus

$$\begin{aligned} y_{n+1} &= \int_0^{(n+1)T} x(t) h\big((n+1)T - t\big) dt \\ &= \int_0^{(n+1)T} h\big((n+1)T - t\big) \\ &\quad \times \Big(\sum_{i=1}^{i_0-1} \chi_{[t_{2i-1}, t_{2i})}(t) + \chi_{[t_{2i_0-1}, (n+1)T)}(t) \Big) dt \\ &\quad - \int_{nT}^{(n+1)T} h\big((n+1)T - t\big) \\ &\quad \times \chi_{[\min(t_{2i_0}, (n+1)T), (n+1)T)}(t) dt. \end{aligned}$$

Hence there exists a transition position t_{2i_0} in the time range $[nT, (n+1)T)$ if and only if

$$\begin{aligned} \tilde{y}_{n+1} &:= -y_{n+1} + \int_0^{(n+1)T} h\big((n+1)T - t\big) \\ &\quad \times \Big(\sum_{i=1}^{i_0-1} \chi_{[t_{2i-1}, t_{2i})}(t) + \chi_{[t_{2i_0-1}, (n+1)T)}(t) \Big) dt \end{aligned} \tag{8}$$

is positive. Moreover if $\tilde{y}_{n+1} > 0$, the transition position t_{2i_0} in the time range $[nT, (n+1)T)$ is determined by

$$t_{2i_0} = (n+1)T - H^{-1}(\tilde{y}_{k+1}). \tag{9}$$

Case 3: The last transition position in $[0, nT)$ is t_{2i_0} for some $1 \leq i_0 < N$.

In this case, the $(n+1)$-th sample $y_{n+1} = x * h((n+1)T)$ is given by

$$\begin{aligned} y_{n+1} &= \int_0^{nT} \Big(\sum_{i=1}^{i_0} \chi_{[t_{2i-1}, t_{2i})(t)} \Big) h\big((n+1)T - t\big) dt \\ &\quad + \int_{\min(t_{2i_0+1}, (n+1)T)}^{(n+1)T} h\big((n+1)T - t\big) dt. \end{aligned} \tag{10}$$

Thus there exists a transition value $t_{2i_0+1} \in [nT, (n+1)T)$ if and only if

$$\tilde{y}_{n+1} := y_{n+1} - \int_0^{nT} \Big(\sum_{i=1}^{i_0} \chi_{[t_{2i-1}, t_{2i})(t)} \Big) h\big((n+1)T - t\big) dt \tag{11}$$

is positive. Also we see that if $\tilde{y}_{n+1} > 0$, then the transition value t_{2i_0+1} can be obtained by

$$t_{2i_0+1} = (n+1)T - H^{-1}(\tilde{y}_{n+1}). \tag{12}$$

Case 4: The last transition position in $[0, nT)$ is t_{2N}.

In this case, all transition positions of the bilevel signal x have been recovered already. Hence the bilevel signal x is fully recovered.

This completes our inductive proof. ■

From the above argument of Theorem 1, we can use the following algorithm to recover a bilevel causal signal x in (1) from its samples $x * h(nT), 1 \leq n \leq K$,where $K > t_{2N}\tau$:

Bilevel Signal Recovery Algorithm:

Step 1: If all samples $y_n = x * h(nT), 1 \leq n \leq K$, are zero, then set $x = 0$ and stop; else find the first nonzero sample, say $y_{n_0} > 0$, the first transition position of the bilevel signal x is located at $t_1 := n_0 - H^{-1}(y_{n_0})$, and set $n = n_0$.

Step 2: Do Step 2a if the last transition position in the time range $[0, nT)$ is t_{2i_0-1} for some $i_0 \geq 1$; do Step 2b elseif the last transition position in the time range $[0, nT)$ is t_{2i_0} for some $1 \leq i_0 < N$; and do Step 4 else.

- Step 2a: Define t_{2i_0} as in (9) if $\tilde{y}_{n+1}$ in (8) is positive, else do Step 3.
- Step 2b: Define t_{2i_0+1} as in (12) if $\tilde{y}_{n+1}$ in (11) is positive, else do Step 3.

Step 3: Set $n = n + 1$. Do Step 2 if $n < K$, and Step 4 if $n = K$.

Step 4: Stop as all transition positions $t_i, 1 \leq i \leq 2N$, of the bilevel signal x are recovered.

We finish this section with a remark that the requirement $R \leq \tau$ in Theorem 1 can be relaxed to the following: There is at most one transition position $t_i, 1 \leq i \leq 2N$, in each sampling range $[nT, (n+1)T), n \geq 1$.

III. Stable recovery of bilevel causal signals

In this section, we consider the maximal sampling error $\sup_n |x * h(nT) - \tilde{x} * h(nT)|$ of two bilevel signals x and $\tilde{x}$ when maximal error of their transition positions are small. We then present some numerical simulations on recovery of a bilevel signal x in (1) from its noisy samples $\{x*h(nT)+\epsilon_n\}$ in (3), where $\epsilon_n, n \geq 1$, are bounded noises of low levels.

First we notice that sampling procedure from bilevel signals x to their samples $\{x * h(nT)\}$ are stable in bounded norm.

Theorem 2: Let $T > 0$, h be a bounded filter supported in $[0, MT)$, $x(t) = \sum_{i=1}^{N} \chi_{[t_{2i-1},t_{2i})}(t)$ be a bilevel causal signal with maximal local innovation rate $R \leq \tau := 1/T$, and $\tilde{x}(t) = \sum_{i=1}^{N} \chi_{[t_{2i-1}+\delta_{2i-1},t_{2i}+\delta_{2i})}$ be a perturbation of the bilevel signal x with perturbed transition positions $\{t_i+\delta_i\}_{i=1}^{2N}$ satisfying

$$\delta := \sup_{1\leq i\leq 2N} |\tilde{t}_i - t_i| < \frac{1}{2R}.$$

Then the sample errors between $x*h(nT)$ and $\tilde{x}*h(nT), n \geq 1$, are dominated by $(\lfloor MRT \rfloor + 2)\|h\|_\infty\delta$, i.e.,

$$|x*h(nT)-\tilde{x}*h(nT)| \leq \big(\lfloor MRT \rfloor+2\big)\|h\|_\infty\delta,\ n \geq 1, \quad (13)$$

where $\|h\|_\infty$ is the L^∞ norm of the sampling kernel h.

Proof: By the assumption on maximal local innovation rate R of the bilevel signal x and the maximal transition position perturbation δ between bilevel signals x and $\tilde{x}$, we have that

$$|x(t) - \tilde{x}(t)| = \sum_{i=1}^{2N} \chi_{t_i+[\min(\delta_i,0),\max(\delta_i,0))}(t).$$

This together with the support assumption for the sampling kernel h gives that

$$\begin{aligned} & |x * h(nT) - \tilde{x} * h(nT)| \\ = & \Big| \int_0^{nT} (x(t) - \tilde{x}(t))h(kT - t)dt \Big| \\ \leq & \|h\|_\infty \int_{(n-M)T}^{nT} \sum_{i=1}^{2N} \chi_{t_i+[\min(\delta_i,0),\max(\delta_i,0))}(t)dt. \end{aligned}$$

Therefore

$$\begin{aligned} & |x * h(nT) - \tilde{x} * h(nT)| \\ \leq & \delta\|h\|_\infty \#\{t_i : t_i \in [(n-M)T - \delta, nT + \delta)\} \\ \leq & \delta\|h\|_\infty(\lfloor (MT + 2\delta)/(1/R) \rfloor + 1) \\ \leq & \delta\|h\|_\infty(\lfloor MRT \rfloor + 2), \end{aligned}$$

where the first inequality holds as $t_i \in [(n-M)T - \delta, nT + \delta)$ if $t_i + [\min(\delta_i, 0), \max(\delta_i, 0))$ and $[(n-M)T, nT)$ have nonempty intersection, the second inequality is true as $t_{i+1} - t_i \geq 1/R$ for all $1 \leq i < 2N$, and the last inequality follows from the assumptions that $\delta < 1/(2R)$ and $R \leq \tau$. This proves the sampling error estimate (13) between the bilevel causal signals x and $\tilde{x}$. ∎

Now we consider the corresponding nonlinear inverse problem how to recover a bilevel signal x from its noisy samples $\{x*h(nT)+\epsilon_n\}$ in (3), where $\epsilon_n, n \geq 1$, are bounded noises. Let us start by looking at two examples.

Example 1: Take $x_1(t) = \sum_{i=1}^{\infty} \chi_{[2i-1,2i)}(t)$ as the original bilevel signal and $h_1(t) = \chi_{[0,2)}(t)$ as the sampling kernel. For sufficiently small $\epsilon > 0$, define $x_{1,\epsilon} = \sum_{i=1}^{\infty} \chi_{[(1+\epsilon)(2i-1),2(1+\epsilon)i)}(t)$. Then for every $i \geq 1$, the i-th transition positions of bilevel signals x_1 and $x_{1,\epsilon}$ are i and $i(1+\epsilon)$ respectively (hence their difference is $i\epsilon$ that could be arbitrary large for sufficiently large i), but on the other hand, maximal sampling errors for those two bilevel signals x_1 and $x_{1,\epsilon}$ are bounded by ϵ,

$$|x_{1,\epsilon} * h_1(n) - x_1 * h_1(n)| = |x_{1,\epsilon} * h_1(n) - 1| \leq \epsilon,\ n \geq 1.$$

This leads to instability of the recovery procedure from samples $\{x_1 * h_1(n)\}$ to the bilevel signal x_1 in the presence of bounded noises.

Example 2: Take x_1 and h_1 in Example 1 as the original bilevel signal and the sampling kernel respectively. Define $x_{2,\epsilon} = \sum_{i=1}^{\infty} \chi_{[2i-1+\epsilon,2i+\epsilon)}(t)$ for sufficiently small $\epsilon > 0$. Then for every $i \geq 1$ the difference between i-th transition positions of bilevel signals x_1 and $x_{2,\epsilon}$ is always ϵ, and there is no difference between their n-th samples except for $n = 1$. This suggests that the recovery procedure from samples $\{x_1 * h_1(n)\}$ to the bilevel signal x_1 is not locally-behaved and the reconstruction error on transition positions could disseminate.

From the above two examples, we see that the nonlinear recovery procedure from samples $\{x*h(n)\}$ to bilevel signals x is *unstable* in the presence of bounded noises and that it is *globally-behaved* in general. In this note, we present some initial numerical simulations with small numbers of transition positions, sampling rate over maximal local rate of innovation and very low levels of noise. Detailed noise performance analysis and stable recovery in the presence of other types of noises will be discussed in the coming paper.

Take a sampling kernel $h_0(t) = \frac{t+1}{2}\chi_{[0,1)}(t)+(2t-1)\chi_{[1,2)}$, and a bilevel signal

$$\begin{aligned} x_0(t) = & \chi_{[0.3791,1.9885)}(t) + \chi_{[3.1306,4.3440)}(t) \\ & +\chi_{[5.7552,7.1820)}(t) + \chi_{[8.7423,10.1052)}(t) \\ & +\chi_{[11.4200,12.6884)}(t) \end{aligned} \quad (14)$$

containing 10 transition positions, see Figure 1. Here transition

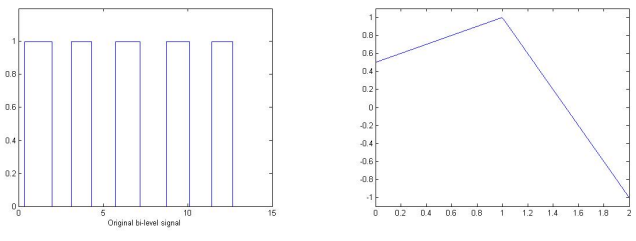


Fig. 1. Bi-level signal x_0 (left) and sampling kernel h_0 (right)

positions $t_i^0, 1 \leq i \leq 10$, of the bilevel signal x_0 are randomly

selected so that $t_i^0 - t_{i-1}^0 \in [1.1, 1.9], 2 \le i \le 10$. The bilevel signal x_0 in (14) has 0.8756 as its maximal local rate of innovation. We sample the convolution $x_0 * h_1$ between x_0 and h_1 every second, which generates the sampling vector $Y_0 = (x_0 * h(1), \dots, x_0 * h(14))$, and then we add bounded random noise to the sampling vector

$$Y_\delta = Y_0 + \delta(\epsilon_1, \dots, \epsilon_{14})$$

with noise level $\delta \ge 0$, where $\epsilon_i \in [-1, 1], 1 \le i \le 14$, are random noises. We apply the bilevel signal recovery algorithm in Section II with some technical adjustment when the reconstructed transition position is approximately located at some sampling positions, and denote the first ten transition positions of the reconstructed bilevel signal x_δ by $t_{1,\delta}, \dots, t_{10,\delta}$. Define maximal error of first ten transition positions by

$$P(\delta) = \max_{1 \le i \le 10} |t_{i,\delta} - t_i^0|.$$

We perform the bilevel signal recovery algorithm in Section II 50 times for every noise level $\delta \in [0, 0.03]$. The maximal value of $P(\delta)$ after performing the algorithm 50 times is plotted in Figure 2 with solid line, while the average value of $P(\delta)$ plotted with dashed line. Notice that $\max_{1 \le n \le 14} |x_0 * h_1(n)| = 0.9796$. Thus the maximal error $P(\delta)$ of transition positions is less than 10% when the noise level $\epsilon = \max_{n \ge 1} |\epsilon_n|$ is at (or below) 2% of the maximal sample value $\max_{n \ge 1} |x_0 * h_0(nT)|$, while some transition positions could not be recovered approximately when the noise level is above 3%. This indicates that our algorithm to recover the bilevel signal from its noisy samples is "reliable" only for low level of bounded noises. We doubt that it is because of the instability of the nonlinear recovery procedure in the presence of bounded noises. We will do the detailed noise performance analysis in the coming paper.

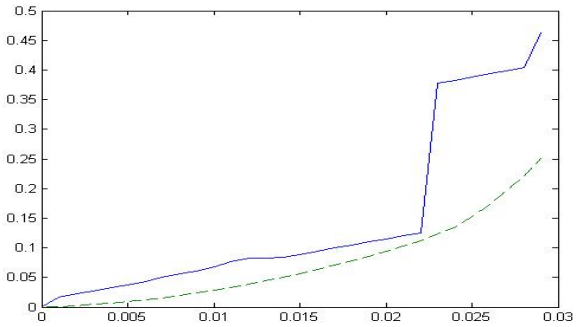


Fig. 2. Maximal transition position error

IV. Conclusion

In this note, we show that bilevel causal signals x could be reconstructed from their samples $x * h(nT), n \ge 1$, if the sampling kernel h is causal and positive on $(0, T)$ and if the sample rate is at (or above) the maximal local rate of innovation of the bilevel signal x. We also propose a stable bilevel signal recovery algorithm in the presence of bounded noise if the number of transition positions of bilevel signals is not large. We remark that the bilevel signal recovery algorithm proposed in this note is applicable when uniform sampling $x*h$ every T second replaced by nonuniform sampling $\{x*h(s_n)\}$ with sampling density $\sup_{n \ge 1} |s_{n+1} - s_n| \le T$, and bilevel causal signal $x = \sum_{i=1}^N \chi_{[t_{2i-1}, t_{2i})}(t)$ with maximal local rate of innovation $R \le 1/T$ replaced by box causal signals $x = \sum_{i=1}^N c_i \chi_{[t_{2i-1}, t_{2i})}(t)$ with maximal local rate of innovation $R \le 1/(2T)$, where for every $1 \le i \le N$, c_i is height of the box located on the time period $[t_{2i-1}, t_{2i})$.

Acknowledgment

The authors would like to thank Professor Ram Mohapatra for his help in the preparation of this note. This work is supported in part by the National Science Foundation (DMS-1109063).

References

[1] M. Vetterli, P. Marziliano, and T. Blu, Sampling signals with finite rate of innovation, *IEEE Trans. Signal Proc.*, **50**(2002), 1417–1428.
[2] R. J.-M. Cramer, R. A. Scholtz, and M. Z. Win, Evaluation of an ultra wide-band propagation channel, *IEEE Trans. Antennas and Propagation*, **50**(2002), 561–569.
[3] D. Donoho, Compressive sampling, *IEEE Trans. Inform. Theory*, **52**(2006), 1289–1306.
[4] Q. Sun, Frames in spaces with finite rate of innovations, *Adv. Comput. Math.*, **28**(2008), 301–329.
[5] I. Maravic and M. Vetterli, Sampling and reconstruction of signals with finite rate of innovation in the presence of noise, *IEEE Trans. Signal Processing*, **53**(2005), 2788–2805.
[6] P. Marziliano, M. Vetterli, and T. Blu, Sampling and exact reconstruction of bandlimited signals with shot noise, IEEE Trans. Inform. Theory, **52**(2006), 2230–2233.
[7] Q. Sun, Non-uniform sampling and reconstruction for signals with finite rate of innovations, *SIAM J. Math. Anal.*, **38**(2006), 1389–1422.
[8] P. L. Dragotti, M. Vetterli and T. Blu, Sampling moments and reconstructing signals of finite rate of innovation: Shannon meets Strang-Fix, *IEEE Trans. Signal Processing*, **55**(2007), 1741–1757.
[9] P. Shukla and P. L. Dragotti, Sampling schemes for multidimensional signals with finite rate of innovation, IEEE Trans. Signal Process., **55**(2007), 3670–3686.
[10] T. Blu, P. L. Dragotti, M. Veterli, P. Marziliano, and L. Coulot, Sparse sampling of signal innovations: theory, algorithms and performace bounds, IEEE Signal Proc. Mag., **31**(2008), 31–40.
[11] N. Bi, M. Z. Nashed and Q. Sun, Reconstructing signals with finite rate of innovation from noisy samples, *Acta Appl. Math.*, **107**(2009), 339–372.
[12] T. Michaeli and Y. C. Eldar, Xampling at the rate of innovation, *IEEE Trans. Signal Processing*, **60**(2012), 1121–1133.
[13] Q. Sun, Localized nonlinear functional equations and two sampling problems in signal processing, arXiv:1304.2664

Approximate FRI with Arbitrary Kernels

Jose Antonio Urigüen
Imperial College of London
jau08@imperial.ac.uk

Pier Luigi Dragotti
Imperial College of London
p.dragotti@imperial.ac.uk

Thierry Blu
The Chinese University of Hong Kong
thierry.blu@m4x.org

***Abstract*—In recent years, several methods have been developed for sampling and exact reconstruction of specific classes of non-bandlimited signals known as signals with finite rate of innovation (FRI). This is achieved by using adequate sampling kernels and reconstruction schemes, for example the exponential reproducing kernels of [1]. Proper linear combinations of this type of kernel with its shifted versions may reproduce polynomials or exponentials exactly.**

In this paper we briefly review the ideal FRI sampling and reconstruction scheme and some of the existing techniques to combat noise. We then present an alternative perspective of the FRI retrieval step, based on moments [1] and approximate reproduction of exponentials. Allowing for a controlled model mismatch, we propose a unified reconstruction stage that addresses two current limitations in FRI: the number of degrees of freedom and the stability of the retrieval. Moreover, the approach is universal in that it can be used with any sampling kernel from which enough information is available.

***Index Terms*—FRI, Sampling, Noise, Matrix Pencil, Approximation**

I. Introduction

Sampling, or the conversion of signals from analog to digital, provides the connection between the continuous-time and discrete-time worlds. The acquisition process is usually modelled as a filtering stage of the input $x(t)$ with a smoothing function $\varphi(t)$ (or sampling kernel), followed by uniform sampling at a rate $f_{\mathrm{s}} = \frac{1}{T}$[Hz]. According to this setup, the measurements are given by

$$y_n = \int_{-\infty}^{\infty} x(t)\varphi\left(\frac{t}{T} - n\right) \mathrm{d}t = \left\langle x(t), \varphi\left(\frac{t}{T} - n\right)\right\rangle.$$

The fundamental problem of sampling is to recover the original waveform $x(t)$ using the samples y_n. When the signal is bandlimited, the answer due to the Nyquist-Shannon theorem is well known. Recently, however, it has been shown [2], [1], [3] that it is possible to sample and perfectly reconstruct specific classes of non-bandlimited signals, known as signals with finite rate of innovation (FRI). Perfect reconstruction is achieved by using a variation of Prony's method, called the annihilating filter method [3], [4].

In this paper we introduce the approximate recovery of FRI signals, from noisy samples taken by an arbitrary kernel. Our analysis follows the setup of [1], where the key to FRI reconstruction is exact reproduction of exponentials. We introduce the new property of approximate reproduction of exponentials by finding proper linear combinations of the sampling kernel. The main advantages of our method are that we can increase the number of measurements, improve the stability of the recovery and generalise the reconstruction stage.

The outline of the paper is as follows. In Section II we review the noiseless scenario of [1] and then give an overview of existing denoising techniques [3], [5]. In Section III we introduce the approximate FRI scenario. We first study the approximate reproduction of exponentials, and then apply this property to the recovery of FRI signals. We also propose an iterative algorithm to refine the accuracy of the reconstruction. Finally, in Section IV we show simulation results, to then conclude in Section V.

This work is supported by the European Research Council (ERC) starting investigator award Nr. 277800 (RecoSamp).

II. Sampling signals with FRI

A. Perfect reconstruction of a stream of Diracs

We first summarise the main steps needed to sample and perfectly recontruct a train of K Diracs

$$x(t) = \sum_{k=0}^{K-1} a_k \delta(t - t_k), \tag{1}$$

where $t_k \in [0, \tau)$, from the samples

$$y_n = \left\langle x(t), \varphi\left(\frac{t}{T} - n\right)\right\rangle = \sum_{k=0}^{K-1} a_k \varphi\left(\frac{t_k}{T} - n\right), \tag{2}$$

for $n = 0, 1, \ldots, N-1$. Here we assume that the sampling period T is such that $\tau = NT$. Moreover, $\varphi(t)$ is an exponential reproducing kernel [1], [6] of compact support that satisfies

$$\sum_{n\in\mathbb{Z}} c_{m,n}\varphi(t - n) = \mathrm{e}^{\alpha_m t}, \tag{3}$$

for proper coefficients $c_{m,n}$, with $m = 0, \ldots, P$ and $\alpha_m \in \mathbb{C}$.

To begin, we linearly combine the samples y_n with the coefficients $c_{m,n}$ of (3) and obtain the new measurements (exponential moments):

$$s_m = \sum_{n=0}^{N-1} c_{m,n} y_n, \tag{4}$$

for $m = 0, \ldots, P$. Then, given that the signal $x(t)$ is a stream of Diracs (1) and combining (4) with (2) we have [1]:

$$s_m = \left\langle x(t), \sum_{n=0}^{N-1} c_{m,n}\varphi\left(\frac{t}{T} - n\right)\right\rangle = \sum_{k=0}^{K-1} x_k u_k^m, \tag{5}$$

with $x_k = a_k \mathrm{e}^{\alpha_0 \frac{t_k}{T}}$ and $u_k = \mathrm{e}^{\lambda \frac{t_k}{T}}$. In order for (5) to hold, we have restricted our analysis to parameters of the form $\alpha_m = \alpha_0 + m\lambda$, where $m = 0, \ldots, P$ and $\alpha_0, \lambda \in \mathbb{C}$. The reason we use these parameters is that they are needed for the values s_m to have a power sum series form (5), which is key to the recovery stage.

The new pairs of unknowns (u_k, x_k) for $k = 0, \ldots, K-1$ can then be retrieved from the measurements s_m using the annihilating filter method [2], [1], [3]. Let h_m with $m = 0, \ldots, K$ denote the filter with z-transform $\hat{h}(z) = \sum_{m=0}^{K} h_m z^{-m} = \prod_{k=0}^{K-1}\left(1 - u_k z^{-1}\right)$. Then, h_m annihilates the series s_m:

$$h_m \star s_m = \sum_{i=0}^{K} h_i s_{m-i} = \sum_{k=0}^{K-1} x_k u_k^m \underbrace{\sum_{i=0}^{K} h_i u_k^{-i}}_{\hat{h}(u_k)} = 0. \tag{6}$$

The zeros of this filter uniquely define the values u_k provided the locations t_k are different. Interestingly, identity (6) can be written in matrix-vector form as:

$$\mathbf{S}\mathbf{h} = 0 \tag{7}$$

which reveals that the Toeplitz matrix $\mathbf{S}$ is rank deficient. The annihilating filter is therefore in the null space of $\mathbf{S}$. By solving the above system, we find the coefficients h_m, and then retrieve u_k from

the roots of $\hat{h}(z)$. Finally, we determine the weights x_k by solving the first K consecutive equations in (5). Notice that the problem can be solved only when $N \geqslant P+1 \geqslant 2K$.

An exponential reproducing kernel is any function $\varphi(t)$ that, together with its shifted versions, can reproduce exponentials, that is, it satisfies (3). The coefficients $c_{m,n}$ are given by

$$c_{m,n} = \int_{-\infty}^{\infty} \mathrm{e}^{\alpha_m t} \tilde{\varphi}(t-n)\,\mathrm{d}t = c_{m,n}\mathrm{e}^{\alpha_m n},$$

where the function $\tilde{\varphi}(t)$ is such that $\langle \tilde{\varphi}(t-n), \varphi(t-m)\rangle = \delta_{m-n}$, and with $c_{m,0} = \int_{-\infty}^{\infty} \mathrm{e}^{\alpha_m x}\tilde{\varphi}(x)\mathrm{d}x$.

Any exponential reproducing kernel can be written as $\varphi(t) = \gamma(t) * \beta_{\vec{\alpha}}(t)$ [1], [6], where $\gamma(t)$ is an arbitrary function, even a distribution, $\beta_{\vec{\alpha}}(t)$ is an E-Spline and $\vec{\alpha} = \{\alpha_m\}_{m=0}^{P}$. The Fourier domain representation of an E-Spline of order $P+1$ is:

$$\hat{\beta}_{\vec{\alpha}}(\omega) = \prod_{m=0}^{P} \frac{1-\mathrm{e}^{\alpha_m - j\omega}}{j\omega - \alpha_m}.$$

In this paper we work with real valued sampling kernels characterised by $\gamma(t)$ and $\beta_{\vec{\alpha}}(t)$ being real. Since $\alpha_m = \alpha_0 + m\lambda$ with $m = 0,\ldots,P$ this implies $\lambda = (\alpha_0^* - \alpha_0)/P$. Note that this condition makes α_m and s_m exist in complex conjugate pairs.

B. Sampling signals with FRI in the presence of noise

Noise is generally present in data acquisition, making the solution explained so far ideal. Assume the noiseless samples y_n are corrupted by additive noise such that we have access to the measurements $\tilde{y}_n = y_n + \epsilon_n$ for $n = 0,\ldots,N-1$. In this situation, the moments, given by the linear combination of samples (4), become noisy:

$$\tilde{s}_m = \underbrace{\sum_{k=0}^{K-1} x_k u_k^m}_{s_m} + \underbrace{\sum_{n=0}^{N-1} c_{m,n}\epsilon_n}_{b_m}, \tag{8}$$

and perfect reconstruction is no longer possible. If ϵ_n are i.i.d. Gaussian, then b_m are samples of Gaussian noise, but not necessarily white.

Note that now (7) becomes approximate due to $\tilde{\mathbf{S}} = \mathbf{S} + \mathbf{B}$, where $\mathbf{B}$ is a Toeplitz matrix formed from the values b_m of (8). Thus, we need alternative ways of solving (7), for instance by using total least squares and Cadzow [3] or the matrix pencil method [7], [5]. The latter can be summarised as follows: obtain the SVD decomposition $\tilde{\mathbf{S}} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^H$, keep the K columns of $\mathbf{U}$ corresponding to the dominant singular values and estimate u_k as the eigenvalues of $\underline{\mathbf{U}}_K^{+}\overline{\mathbf{U}}_K$. Here, $\underline{(\cdot)}$ and $\overline{(\cdot)}$ are operations to omit the last and first rows of $(\cdot)$.

In addition, note that the covariance matrix of the noise $\mathbf{R}_\mathbf{B} = \mathrm{E}\{\mathbf{B}^H\mathbf{B}\}$ may not be a multiple of the identity. In order for SVD to operate properly it is necessary to pre-whiten the noise [8], for instance by using a linear transform $\mathbf{W} = \mathbf{R}_\mathbf{B}^{\dagger/2}$ [9] for $\mathbf{A} = \mathbf{B}\mathbf{W}$ to satisfy that $\mathbf{R}_\mathbf{A} = \mathrm{E}\{\mathbf{A}^H\mathbf{A}\} = \mathbf{I}$. Here, $(\cdot)^{\dagger/2}$ denotes the square root of the pseudoinverse of $(\cdot)$. In our simulations, we apply pre-whitening on $\tilde{\mathbf{S}}$ such that $\tilde{\mathbf{S}}\mathbf{W}$ is now contaminated by white noise $\mathbf{A}$. We then directly use matrix pencil on $\tilde{\mathbf{S}}\mathbf{W}$.

In order to analyse the effect of noise on the accuracy with which FRI signals can be recovered we use the Cramér–Rao lower bound (CRB). This is a lower bound on the mean square error (MSE) that applies to any unbiased estimator [4]. A stream of K Diracs is completely characterised by the vector $\Theta = (t_0,\ldots,t_{K-1},a_0,\ldots,a_{K-1})^T$, of K locations and amplitudes. And the goal of FRI reconstruction is to estimate Θ either from the vector of N samples $\tilde{\mathbf{y}} = (\tilde{y}_0,\ldots,\tilde{y}_{N-1})^T$ or the vector of $P+1$ noisy moments $\tilde{\mathbf{s}} = (\tilde{s}_0,\ldots,\tilde{s}_P)^T$.

The analysis of the CRB for the estimation problem given $\tilde{\mathbf{y}}$ is detailed in [3]. In our simulations, we compare the estimation accuracy with this bound, but we consider the CRB for the estimation from $\tilde{\mathbf{s}}$. Given values s_m that exist in complex conjugate pairs, then any unbiased estimate $\hat{\Theta}(\tilde{\mathbf{s}}) = (\hat{t}_0,\ldots,\hat{t}_{K-1},\hat{a}_0,\ldots,\hat{a}_{K-1})^T$ has a covariance matrix that is lower bounded by [10]

$$\mathrm{cov}(\hat{\Theta}(\tilde{\mathbf{s}})) \geqslant (\Phi^H\mathbf{R}^{-1}\Phi)^{-1}. \tag{9}$$

Here, $(\cdot)^H$ denotes Hermitian transpose. Moreover, provided ϵ_n are samples of additive white Gaussian noise, of zero mean and variance σ^2, then $\mathbf{R} = \mathrm{E}\{\mathbf{b}\mathbf{b}^H\} = \sigma^2\mathbf{C}\mathbf{C}^H$, because $\mathbf{b}$ is the vector of noise values b_m of (8). The matrix Φ in (9) takes the form:

$$\left(\begin{array}{ccc|ccc} a_0\alpha_0\mathrm{e}^{\alpha_0 t_0} & \ldots & a_{K-1}\alpha_0\mathrm{e}^{\alpha_0 t_{K-1}} & \mathrm{e}^{\alpha_0 t_0} & \ldots & \mathrm{e}^{\alpha_0 t_{K-1}} \\ a_0\alpha_1\mathrm{e}^{\alpha_1 t_0} & \ldots & a_{K-1}\alpha_1\mathrm{e}^{\alpha_1 t_{K-1}} & \mathrm{e}^{\alpha_1 t_0} & \ldots & \mathrm{e}^{\alpha_1 t_{K-1}} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ a_0\alpha_P\mathrm{e}^{\alpha_P t_0} & \ldots & a_{K-1}\alpha_P\mathrm{e}^{\alpha_P t_{K-1}} & \mathrm{e}^{\alpha_P t_0} & \ldots & \mathrm{e}^{\alpha_P t_{K-1}} \end{array}\right).$$

III. Universal sampling of signals with FRI

In many practical circumstances we may not be able to choose the sampling kernel $\varphi(t)$, or even know its exact shape. In such cases there might not be an easy way of finding coefficients $c_{m,n}$ for the linear combination of samples (4) to yield a power sum series (5). And this is key in the FRI setting to map the signal reconstruction problem to Prony's method in spectral-line estimation theory.

In this section we consider any function $\varphi(t)$ for which the exponential reproduction property (3) is only approximate. We propose to use the coefficients $c_{m,n}$ for approximate reproduction to build (4) such that they yield a power sum series (5) from which the FRI parameters can be retrieved.

A. Approximate reproduction of exponentials

Assume we want to use a function $\varphi(t)$ and its integer shifts to approximate the exponential $\mathrm{e}^{\alpha t}$. In other words, we seek the coefficients c_n that best fit:

$$\sum_{n\in\mathbb{Z}} c_n\varphi(t-n) \cong \mathrm{e}^{\alpha t}. \tag{10}$$

In order to do so, we directly use $c_n = c_0\mathrm{e}^{\alpha n}$. Then, equation (10) is equivalent to approximating $g_\alpha(t) = c_0\sum_{n\in\mathbb{Z}}\mathrm{e}^{-\alpha(t-n)}\varphi(t-n)$ by the constant value 1. We also note that $g_\alpha(t)$ is a 1-periodic function, because $g_\alpha(t) = g_\alpha(t+1)$. It can therefore be decomposed using the Fourier series as

$$g_\alpha(t) = \sum_{l\in\mathbb{Z}} g_l\mathrm{e}^{j2\pi lt}, \tag{11}$$

where

$$\begin{aligned} g_l &= \int_0^1 g_\alpha(t)\mathrm{e}^{-j2\pi lt}\mathrm{d}t = c_0\sum_{k\in\mathbb{Z}}\int_0^1 \mathrm{e}^{-\alpha(t-k)}\varphi(t-k)\mathrm{e}^{-j2\pi lt}\mathrm{d}t \\ &\overset{(a)}{=} c_0\int_{-\infty}^{\infty}\mathrm{e}^{-\alpha x}\varphi(x)\mathrm{e}^{-j2\pi lx}\mathrm{d}x = c_0\hat{\varphi}(\alpha + j2\pi l). \end{aligned}$$

Here, (a) is due to using $x = t-k$ and combining the sum over $k\in\mathbb{Z}$ and the integral dependent on k. Also $\hat{\varphi}(s) = \int_{-\infty}^{\infty}\varphi(x)\mathrm{e}^{-sx}\mathrm{d}x$ denotes the Laplace transform of $\varphi(x)$.

In general $\varphi(t)$ may be any function and we can find different sets of coefficients c_n for (10) to hold. The accuracy of our approximation is given by:

$$\varepsilon(t) = \mathrm{e}^{\alpha t}\left[1 - c_0 \sum_{l\in\mathbb{Z}} \hat{\varphi}(\alpha + j2\pi l)\mathrm{e}^{j2\pi l t}\right]. \qquad (12)$$

Note that if the Laplace transform of $\varphi(t)$ decays sufficiently quickly, very few terms are needed to have an accurate bound for the error.

A natural choice of the coefficients $c_n = c_0\mathrm{e}^{\alpha n}$ is obtained by discarding every term in (11) for $l \neq 0$ and making $g_0 = 1$, hence $c_0 = \hat{\varphi}(\alpha)^{-1}$. Interestingly, this is a simplified version of the least-squares coefficients [11] for the approximation in (10). The main advantage of using coefficients with $c_0 = \hat{\varphi}(\alpha)^{-1}$ is that they are very easy to compute, because they only require the knowledge of the Laplace transform of $\varphi(t)$ at α.

We conclude with an example. Consider a linear spline that reproduces polynomials of orders 0 and 1 exactly, as shown in Figure 1 (a). We want to approximate the complex exponentials $\mathrm{e}^{j-\frac{\pi}{16}(2m-7)t}$ for $m = 3$ and $m = 0$ by using linear combinations of the spline. This can be done by selecting coefficients $c_{m,n} = \hat{\varphi}(\alpha_m)^{-1}\mathrm{e}^{\alpha_m n}$ where $\alpha_m = j\frac{\pi}{16}(2m-7)$. We illustrate the reproduction of the real part of the complex exponentials in Figure 1 (b-c). Note how the one with lower frequency is better approximated. Moreover, we have seen experimentally that higher order splines tend to improve the quality of the approximation. Also note there is no fixed number of exponentials that may be well approximated.

B. Approximate FRI recovery

Consider again the stream of Diracs (1) and samples of the form (2), now taken by an arbitrary sampling kernel $\varphi(t)$. In order to retrieve the locations t_k and amplitudes a_k for $k = 0, \ldots, K-1$, we first obtain the coefficients $c_{m,n} = \hat{\varphi}(\alpha_m)^{-1}\mathrm{e}^{\alpha_m n}$ for $m = 0, \ldots, P$. We only need to know the Laplace transform of $\varphi(t)$ at α_m. Note that P is a free parameter, subject to $P + 1 \geqslant 2K$.

We proceed in the same way as in the case of exact reproduction of exponentials, but now the exponential moments take the form

$$s_m = \Big\langle x(t), \underbrace{\sum_{n=0}^{N-1} c_{m,n}\varphi(t-n)}_{\mathrm{e}^{\alpha_m t} - \varepsilon_m(t)} \Big\rangle = \sum_{k=0}^{K-1} x_k u_k^m - \zeta_m$$

where $x_k = a_k\mathrm{e}^{\alpha_0 t_k}$ and $u_k = \mathrm{e}^{\lambda t_k}$. Here we have used $T = 1$ and $\alpha_m = \alpha_0 + m\lambda$, with $m = 0, \ldots, P$, and $\alpha_0, \lambda \in \mathbb{C}$. There is a model mismatch due to the approximation error $\varepsilon_m(t)$ of (12), equal to $\zeta_m = \sum_{k=0}^{K-1} a_k\varepsilon_m(t_k)$.

The model mismatch depends on the quality of the approximation, and depends on the coefficients $c_{m,n}$ and the values α_m and P. We treat this error as noise, and retrieve the parameters of the signal using the methods of Section II-B. In close-to-noiseless settings, the estimation of the Diracs can be refined using the iterative procedure shown in Algorithm 1.

C. How to select the approximation parameters α_m

In order to simplify the problem, we restrict the exponential parameters to be of the form:

$$\alpha_m = j\omega_m = j\frac{\pi}{L}(2m - P) \quad m = 0, \ldots, P. \qquad (13)$$

Purely imaginary parameters allow for a more stable retrieval of the pairs (t_k, a_k) from (5). The values to be determined are, therefore, P and L. We choose the values that minimise the first diagonal term of (9) when $K = 1$, which corresponds to the error in the recovery of the location t_0. In most cases we have analysed, the best P is greater or equal than the support of the sampling kernel $\varphi(t)$ and L is in the range $P + 1 \leqslant L \leqslant 4(P + 1)$.

Algorithm 1 Recovery of a train of K Diracs using approximation of exponentials

1: Compute the moments $s_m^0 = \sum_n c_{m,n}y_n$ and set $s_m^i = s_m^0$, for $m = 0, \ldots, P$.
2: Build the system of equations (6) using s_m^i and retrieve the annihilating filter h_m.
3: Calculate u_k^i from the roots of h_m, and $t_k^i = \frac{1}{\lambda}\ln u_k^i$, for the ith iteration.
4: Find x_k^i from the first K consecutive equations in (5), and the amplitudes $a_k^i = x_k^i\mathrm{e}^{-\alpha_0 t_k}$.
5: Recalculate the moments for the next iteration by removing the model mismatch:

$$s_m^{i+1} = s_m^0 + \sum_{k=0}^{K-1} a_k^i\varepsilon_m(t_k^i),$$

where $\varepsilon_m(t)$ is given by (12).
6: Repeat steps 2 to 5 until convergence of the values (a_k^i, t_k^i).

IV. Simulations

We take $N = 31$ samples of a train of K Diracs using a B-Spline kernel, and we corrupt the measurements with additive white Gaussian noise of variance σ^2. This is chosen according to the required signal-to-noise ratio $\mathrm{SNR(dB)} = 10\log(\|\mathbf{y}\|^2/N\sigma^2)$. We then obtain the approximation coefficients $c_{m,n} = \hat{\varphi}(\alpha_m)^{-1}\mathrm{e}^{\alpha_m n}$, where α_m is as in (13) with $L = 2(P + 1)$ and $m = 0, \ldots, P$. Finally, we compute the noisy $P + 1$ moments and retrieve the innovation parameters (t_k, a_k), for $k = 0, \ldots, K-1$, using the matrix pencil method. We calculate the standard deviation of the error in the estimation of the location, over 1000 realisations of the noise, and compare it to the sample-based and moment-based CRBs of Section II-B.

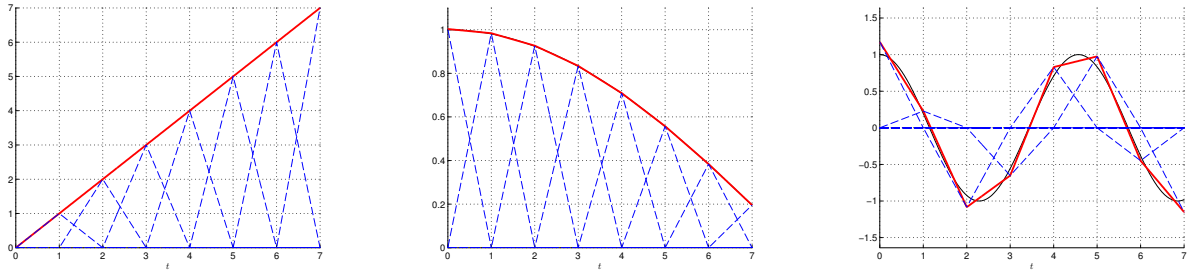
Figure 2 shows the deviation in the location of $K = 6$ Diracs. We compare the performance (a) when we sample with a B-Spline of order 26 and use the default retrieval based on the reproduction of polynomials [1], with (b) when we sample with a B-Spline of order 6 and apply the retrieval based on approximation of exponentials, with 26 moments; both aided with pre-whitening. The SNR is 20dB. It is only in the latter case that we can recover all the Diracs. Moreover, the accuracy with which the Diracs are recovered is one order of magnitude better for the approximated FRI.

We show further results when we use the approximate method to retrieve $K = 2$ Diracs from the samples taken by a B-Spline kernel of order 6. Even when we fix the order of the kernel, we can use any number of moments $P + 1$ to improve the performance. Figures 3 (a-d) are for parameters (13) with $L = \frac{3}{2}(P+1)$ and $m = 0, \ldots, P$. As the number of moments $P+1$ increases, the performance is better and eventually reaches the sample-based CRB.

V. Conclusions

We have presented an alternative FRI retrieval approach, based on the approximate reproduction of exponentials. Allowing for a controlled model mismatch, we propose a standard reconstruction stage that is able to increase the stability of existing FRI schemes.

Moreover, in many practical circumstances we may not be able to choose the sampling kernel or even know its exact shape. However,

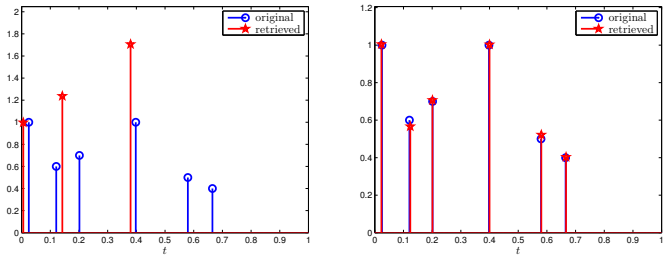


(a) Reproduction of t (b) Approximation of $\mathrm{Re}\{\mathrm{e}^{-j\frac{\pi}{16}t}\}$ (c) Approximation of $\mathrm{Re}\{\mathrm{e}^{-j\frac{7\pi}{16}t}\}$

Figure 1. *B-Spline kernel reproduction capabilities*. Figure (a) shows the exact reconstruction of a polynomial of order 1, by using a proper combination of shifted versions of a linear spline. Figures (b-c) show the approximation of the real parts of 2 complex exponentials: $\mathrm{e}^{j\frac{\pi}{16}(2m-7)t}$ for $m = 3, 0$, by using a proper combination of shifted versions of the linear spline. The coefficients are $c_{m,n} = \hat{\varphi}(\alpha_m)^{-1}\mathrm{e}^{\alpha_m n}$ where $\alpha_m = j\frac{\pi}{16}(2m-7)$. We plot the weighted and shifted versions of the splines with dashed blue lines, the reproduced polynomial and exponentials with red solid lines, and the exact functions with solid black lines.

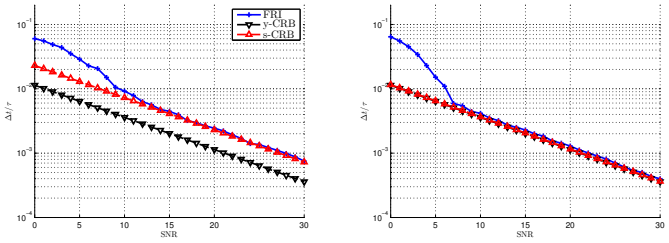
we have seen that if we know the Laplace transform of the kernel at values α_m, we can find coefficients for the linear combination of shifted versions of the sampling kernel to approximate exponentials $\mathrm{e}^{\alpha_m t}$. Equipped with this property we can sample a stream of K Diracs and retrieve it from $2K$ measurements. The accuracy of the reconstruction depends on the quality of the the approximation and the level of noise.

Future work includes FRI retrieval with partial information on the sampling kernel, with more challenging existing FRI kernels (such as the Gaussian), and extensions to more dimensions and non-uniform sampling. In addition, approximate reconstruction may also be generalised when we have access to measurements taken by different kernels, each of which is capable of approximating certain exponentials.



(a) Default (b) Approximation

Figure 2. *B-Spline kernel behaviour*. We retrieve $K = 6$ Diracs from $N = 31$ noisy samples: (a) using the polynomial recovery of [1], with a kernel of order 26 and also $P+1 = 26$ moments; (b) using the approximated recovery with parameters (13) where $L = 2(P+1)$ and $m = 0, \ldots, P$, with a kernel of order 6 and $P + 1 = 26$ moments. The SNR in both cases is 20dB.



(a) $P + 1 = 6$ (b) $P + 1 = 31$

Figure 3. *Approximated retrieval using a B-Spline*. These figures show the error in the estimation of the first Dirac out of $K = 2$ retrieved using the approximated FRI recovery. We show how, even when we fix the order of the kernel to 6, we can reconstruct any number of moments $P + 1$ and improve the performance. In fact, with the appropriate choice $L = \frac{3}{2}(P + 1)$ the performance improves until the sample-based CRB is reached.

BIBLIOGRAPHY

[1] P. L. Dragotti, M. Vetterli, and T. Blu, "Sampling Moments and Reconstructing Signals of Finite Rate of Innovation: Shannon Meets Strang-Fix," *IEEE Transactions on Signal Processing*, vol. 55, pp. 1741–1757, May 2007.

[2] M. Vetterli, P. Marziliano, and T. Blu, "Sampling signals with finite rate of innovation," *IEEE Transactions on Signal Processing*, vol. 50, pp. 1417–1428, June 2002.

[3] T. Blu, P. L. Dragotti, M. Vetterli, P. Marziliano, and L. Coulot, "Sparse Sampling of Signal Innovations," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 31–40, 2008.

[4] P. Stoica and R. L. Moses, *Introduction to Spectral Analysis*. Englewood Cliffs, NJ: Prentice-Hall, 2000.

[5] I. Maravic and M. Vetterli, "Sampling and reconstruction of signals with finite rate of innovation in the presence of noise," *IEEE Transactions on Signal Processing*, vol. 53, pp. 2788–2805, August 2005.

[6] M. Unser and T. Blu, "Cardinal Exponential Splines: Part I—Theory and Filtering Algorithms," *IEEE Transactions on Signal Processing*, vol. 53, pp. 1425–1438, April 2005.

[7] Y. Hua and T. K. Sakar, "Matrix Pencil Method for Estimating Parameters of Exponentially Damped Undamped Sinusoids in Noise," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 38, pp. 814–824, May 1990.

[8] B. De Moor, "The Singular Value Decomposition and Long and Short Spaces of Noisy Matrices," *IEEE Transactions on Signal Processing*, vol. 41, pp. 2826–2838, September 1993.

[9] Y. C. Eldar and A. V. Oppenheim, "MMSE Whitening and Subspace Whitening," *IEEE Trans. Signal Processing*, vol. 49, pp. 1846–1851, July 2003.

[10] E. Ollila, "On the Cramér-Rao bound for the constrained and unconstrained complex parameters," *Sensor Array and Multichannel Signal Processing Workshop*, pp. 414–418, July 2008.

[11] M. Unser, A. Aldroubi, and M. Eden, "Polynomial Spline Signal Approximations: Filter Design and Asymptotic Equivalence with Shannon's Sampling Theorem," *IEEE Transactions on Information Theory*, vol. 38, pp. 95–103, January 1992.

Algebraic signal sampling, Gibbs phenomenon and Prony-type systems

Dmitry Batenkov*† and Yosef Yomdin*‡
*Department of Mathematics,Weizmann Institute of Science, Rehovot 76100, Israel
†Email: dima.batenkov@weizmann.ac.il
‡Email: yosef.yomdin@weizmann.ac.il

***Abstract*—Systems of Prony type appear in various signal reconstruction problems such as finite rate of innovation, superresolution and Fourier inversion of piecewise smooth functions. We propose a novel approach for solving Prony-type systems, which requires sampling the signal at arithmetic progressions. By keeping the number of equations small and fixed, we demonstrate that such "decimation" can lead to practical improvements in the reconstruction accuracy. As an application, we provide a solution to the so-called Eckhoff's conjecture, which asked for reconstructing jump positions and magnitudes of a piecewise-smooth function from its Fourier coefficients with maximal possible asymptotic accuracy – thus eliminating the Gibbs phenomenon.**

I. Introduction

The "Prony system" of equations

$$m_k = \sum_{j=1}^{K} c_j z_j^k, \qquad c_j, z_j \in \mathbb{C},\ k \in \mathbb{N} \tag{1}$$

appeared originally in the work of R.Prony [18] in the context of fitting a sum of exponentials to observed data samples. He showed that the unknowns $\{c_j, z_j\}_{j=1}^{K}$ can be recovered explicitly from $\{m_0, \dots, m_{2K-1}\}$ by what is known today as "Prony's method". The system (1) appears in areas such as frequency estimation, Padé approximation, array processing, statistics, interpolation, quadrature, radar signal detection, error correction codes, and many more. In modern signal processing, (1) is of fundamental importance in the field of sub-Nyquist sampling (related terms are superresolution [9], [10] and finite rate of innovation [12]). A basic problem there is to recover an unknown "spike train", a linear combination of δ-functions

$$f(x) = \sum_{j=1}^{K} b_j \delta(x - x_j), \qquad c_j \in \mathbb{R},\ x_j \in [-\pi, \pi]$$

from its Fourier samples

$$\widehat{f}(k) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t)\, \mathrm{e}^{-\imath k t}\, \mathrm{d}\, t. \tag{2}$$

This research was supported by the Adams Fellowship Program of the Israeli Academy of Sciences and Humanities, ISF Grant No. 639/09 and by the Minerva foundation.

The resulting system is of course a special case of (1). If a more general model is considered,

$$f(x) = \sum_{j=1}^{K} \sum_{\ell=0}^{\ell_j - 1} b_{\ell,j} \delta^{(\ell)}(x - x_j), \quad b_{\ell,j} \in \mathbb{R},\ x_j \in [-\pi, \pi], \tag{3}$$

then (2) becomes, after a change of variables,

$$m_k = \sum_{j=1}^{K} z_j^k \sum_{\ell=0}^{\ell_j - 1} c_{\ell,j} k^\ell, \qquad c_{\ell,j} \in \mathbb{C},\ |z_j| = 1. \tag{4}$$

Many research efforts are devoted to stable solution of Prony-type systems (see e.g. [2], [8], [11], [17], [19] and references therein). We propose a novel approach to this problem, which requires sampling the signal at arithmetic progressions. By keeping the number of equations small and fixed, we demonstrate (in Section II) that such "decimation" can lead to practical improvements in the reconstruction accuracy, to a certain extent avoiding a well-known numerical instability of these systems.

In Section III we consider the problem of recovering a piecewise-smooth function, including the positions of its discontinuities, from its Fourier samples. The algebraic reconstruction method due to K.Eckhoff in essense required a solution of a particular instance of the system (4) with the error in the left-hand side having a certain asymptotic decay rate. Previously it was shown in [6], [7] that this approach yields a nonlinear approximation which is "half as accurate" compared to the best possible bound. As we elaborate in Section III, applying the decimation technique to the Prony-type system results in full asymptotic accuracy, thus completely eliminating the Gibbs phenomenon.

In Section IV we discuss several promising directions for future research.

II. Decimated Prony-type systems

Suppose that the "polynomial Prony model" (4) is to be fitted to the noisy measurements $\widetilde{m}_0, \dots \widetilde{m}_{M-1}$. We denote the number of unknowns by $R = \sum_{j=1}^{K} (\ell_j + 1)$. At first sight, using all the M measurements for fitting should improve reconstruction accuracy. While this is certainly justified in the case where the noise statistics are known (as demonstrated in e.g. [2], [19]), this might backfire if the noise is "adversary",

or "worst-case". Potts & Tasche [17] show that when Prony system (1) is solved by least squares minimization for all M equations at once, then even if the nodes $\{z_j\}$ are detected very accurately, the error for magnitudes is amplified by a factor of $\sqrt{MR}$. This shows that it might actually be productive to stay with small number of measurements. We are therefore justified in making a simplifying assumption that the number of equations used for reconstruction equals the number of unknowns R. In this case the solution to the reconstruction problem can be characterized as the exact inversion of the measurement mapping $\mathcal{P}_I : \mathbb{C}^R \to \mathbb{C}^R$ which associates to any parameter vector $\mathbf{x} = \left\{\{c_{ij}\}, \{x_i\}\right\} \in \mathbb{C}^R$ its corresponding exact measurement vector $\mathbf{y} = \left(m_{i_0}, \ldots, m_{i_{R-1}}\right) \in \mathbb{C}^R$ where $I = \{i_0 < i_1 < \cdots < i_{R-1}\} \subset [0, M-1]$ is a given index set. Perhaps the most natural choice for the index sets I is given by arithmetic progressions

$$I_{t,p} = \{t, t+p, \ldots, t+(R-1)\,p\}\,, \qquad t \geqslant 0,\ p \geqslant 1.$$

Following [8], we estimate for such $I = I_{t,p}$ the (local) stability of inversion by the *Lipschitz constant of* $\mathcal{P}_I^{-1}$ at the regular points of $\mathcal{P}_I$, which in turn are given by the following proposition.

Proposition 1. *The vector* $\mathbf{x} = (\{z_j, c_{i,j}\}) \in \mathbb{C}^R$ *is a regular point of* $\mathcal{P}_I$ *with* $I = I_{t,p}$ *if and only if* $z_j^p \neq z_i^p$ *for* $i \neq j$, *and* $c_{\ell_j - 1, j} \neq 0$ *for all* $j = 1, \ldots, K$.

We have the following upper bound on the accuracy of *any solution method*.

Theorem 2. *Consider the polynomial Prony system* (4) *with a fixed structure* $\left\{K, \{\ell_j\}_{j=1}^K\right\}$ *on* $I = I_{t,p}$, *and let* $\mathbf{x} = (\{z_j, c_{i,j}\}) \in \mathbb{C}^R$ *be a regular point of* $\mathcal{P}_I$. *If the error in each measurement is bounded in absolute value by* $\varepsilon \ll 1$, *then the errors in recovering the components of the original parameter vector* $\mathbf{x}$ *satisify*

$$|\Delta c_{i,j}| \quad \leq \quad C\,(i, \ell_j) \left(\frac{2}{\delta_p}\right)^R \left(\frac{1}{2} + \frac{R}{\delta_p}\right)^{\ell_j} \frac{t^{\ell_j - i}}{p^i} \left(1 + \frac{|c_{i-1,j}|}{\left|c_{\ell_j - 1, j}\right|}\right) \varepsilon,$$

$$|\Delta z_j| \quad \leq \quad \frac{2}{\ell_j!} \left(\frac{2}{\delta_p}\right)^R \frac{1}{\left|c_{\ell_j - 1, j}\right|} p^{-\ell_j} \varepsilon,$$

where $\delta_p \stackrel{def}{=} \min_{i \neq j} \left|z_j^p - z_i^p\right|$ *and* $C\,(i, \ell_j)$ *is an explicit constant (for consistency we take* $c_{-1,j} = 0$ *in the above formula).*

This result directly generalizes earlier stability estimates of [8] for the special case $I = I_{0,1}$. The proofs of both Proposition 1 and Theorem 2 are based on factorizing the Jacobian matrix of the map $\mathcal{P}_I$ along the same lines as in [8], while adding the analysis of the Jacobian's dependence on t and p.

Now suppose that the number of available measurements $M \to \infty$, while the noise ε remains bounded. It is easy to see that for the index set $I = I_{0, \lfloor \frac{M}{R} \rfloor}$ we obtain an improvement in accuracy of recovering the jump z_j of the order $\sim M^{\ell_j}$, compared with the non-decimated measurement set $I_{0,1}$.

Remark 3. If initially two nodes are close (say by δ), the decimation improves accuracy up to a certain limit. To see this, just substitute $\delta_p \sim p\delta$ into Theorem 2 and get an improvement by factor of $p^{-R-\ell_j}$.

Turning to particular solution methods, the decimation is fairly straightforward to implement. Indeed, taking any algorithm for the standard Prony-type system, one just needs to make the following modifications (for simplicity we consider only the recovery of the nodes $\{z_j\}$).

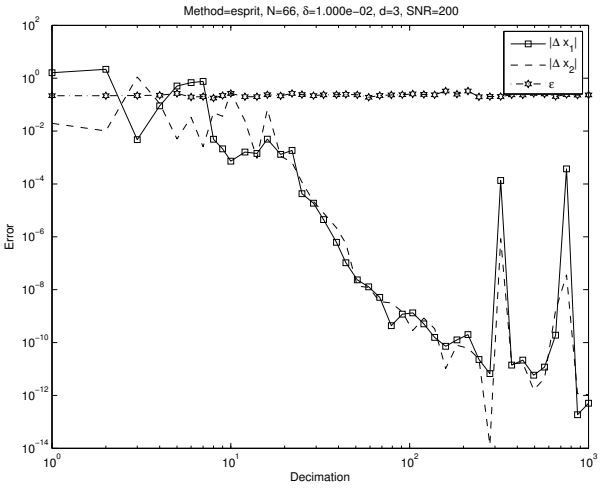
1) Choose the decimation parameter p.
2) Feed the original algorithm with the decimated measurements $m_0, m_p, m_{2p} \ldots$, and obtain the estimated node w_j.
3) Take $z_j = \sqrt[p]{w_j}$.

We have tested the decimation technique according to the above procedure on two well-known algorithms for Prony systems - ESPRIT [2] and nonlinear least squares (LS, implemented by MATLAB's `lsqnonlin`). In the first experiment, we fixed the number of measurements to be 66, and changed the decimation parameter p, while keeping the noise level constant. The accuracy of recovery increased with p – see Figure 1 on page 3. In the second experiment, we fixed the highest available measurement to be $M = 2200$, and changed the decimation from $p = 1$ to $p = 100$ (thereby reducing the number of measurements from 2200 to just 22). The accuracy of recovery stayed relatively constant – see Figure 2 on page 3. Note that such a reduction leads to a corresponding decrease in the running time (calculating singular-value decomposition of large matrices, as in ESPRIT, is a time-consuming operation).

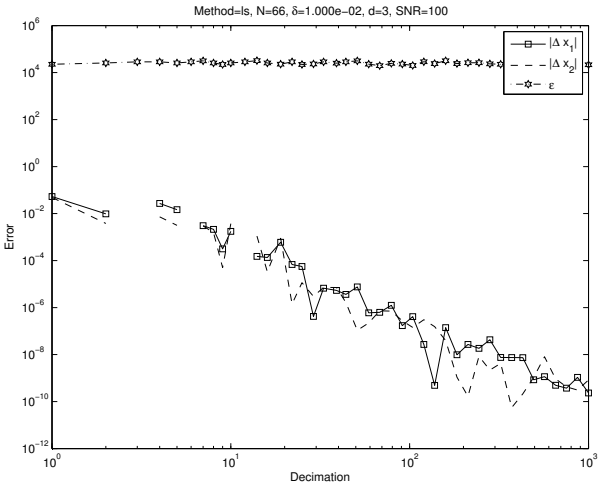
III. Piecewise-smooth Fourier reconstruction

Consider the problem of reconstructing an integrable function $f : [-\pi, \pi] \to \mathbb{R}$ from a finite number of its Fourier coefficients (2). If f is C^d and periodic, then the truncated Fourier series $\mathfrak{F}_M(f) \stackrel{\text{def}}{=} \sum_{|k|=0}^{M} c_k(f)\, \mathrm{e}^{\imath k x}$ approximates f with error at most $C \cdot M^{-d-1}$, which is optimal. If, however, f is not smooth even at a single point, the rate of accuracy drops to only M^{-1}. Still, one can hope to restore the best accuracy by using the a-priori information to produce some non-standard summation method. This accuracy problem, also known as the Gibbs phenomenon, is very important in applications, such as calculation of shock waves in PDEs. It has received much attention especially in the last few decades - see e.g. a recent book [16].

The so-called "algebraic approach" to this problem, first suggested by K.Eckhoff [13], is as follows. Assume that f has $K > 0$ jump discontinuities $\{x_j\}_{j=1}^K$, and $f \in C^d$ in every segment (x_{j-1}, x_j). We say that in this case f belongs to the class $PC\,(d, K)$. Denote the associated jump magnitudes at x_j by $a_{\ell,j} \stackrel{\text{def}}{=} f^{(\ell)}(x_j^+) - f^{(\ell)}(x_j^-)$. Then write the piecewise smooth f as the sum $f = \Psi + \Phi$, where $\Psi(x)$ is smooth and periodic and $\Phi(x)$ is a piecewise polynomial of degree d, uniquely determined by $\{x_j\}, \{a_{\ell,j}\}$ such that it "absorbs" all the discontinuities of f and its first d derivatives. In particular,

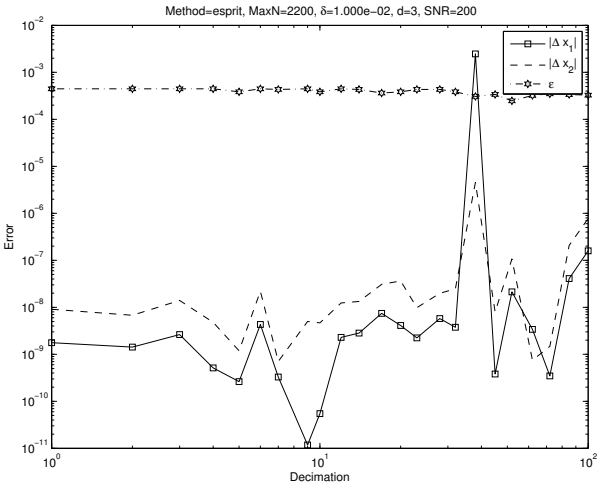


(a) ESPRIT, $d = 3$

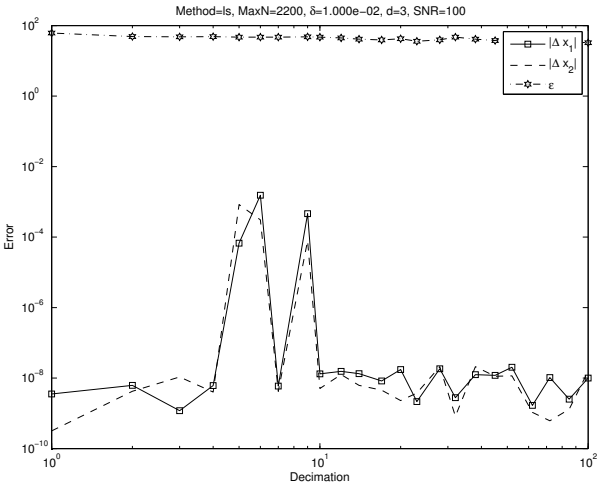


(b) LS, $d = 3$

Figure 1: Reconstruction error as a function of the decimation with fixed number of measurements ($M = 66$). The signal has two nodes with distance $\delta = 10^{-2}$ between each other. Notice that ESPRIT requires significantly higher Signal-to-Noise Ratio in order to achieve the same performance as LS.



(a) ESPRIT, $d = 3$



(b) LS, $d = 3$

Figure 2: Reconstruction error as a function of the decimation, reducing number of measurements from $M = 2200$ to $M = 22$. The signal has two nodes with distance $\delta = 10^{-2}$ between each other. The reconstruction accuracy remains almost constant.

the Fourier coefficients of Φ have the explicit form

$$c_k(\Phi) = \frac{1}{2\pi} \sum_{j=1}^{K} \mathrm{e}^{-\imath k x_j} \sum_{\ell=0}^{d} (\imath k)^{-\ell-1} a_{\ell,j}, \quad k = 1, 2, \ldots. \qquad (5)$$

For $k \gg 1$, we have $|c_k(\Phi)| \sim k^{-1}$, while $|c_k(\Psi)| \sim k^{-d-2}$. Consequently, Eckhoff suggested to pick large enough k and solve the approximate system of equations (4) where $m_k = 2\pi (\imath k)^{d+1} c_k(f)$, $z_j = \mathrm{e}^{-\imath x_j}$ and $c_{\ell,j} = \imath^{\ell} a_{d-\ell,j}$. His proposed method of solution was to use the known values $\{m_k\}_{k \in I}$ where

$$I = \{M - (d+1)K + 1, M - (d+1)K + 2, \ldots, M\} \quad (6)$$

to construct an algebraic equation satisfied by the unknowns $\{z_1, \ldots, z_K\}$, and solve this equation numerically. Based on some explicit computations for $d = 1, 2$; $K = 1$ and large number of numerical experiments, he conjectured that his method would reconstruct the jump locations with accuracy M^{-d-1}.

Let us consider the problem in the framework of Prony type system (4). The error term is of magnitude $|\varepsilon| \sim M^{-1}$. The index set (6) is just $I_{t,p}$ with $t \sim M$, $p = 1$ (i.e. no decimation). Therefore, by Theorem 2 we get accuracy only of order $|\Delta x_j| \sim M^{-1}$.

Now consider the decimated setting for this problem. By the above, we can approximate eash jump x_j up to accuracy M^{-1}. Set

$$N = \left\lfloor \frac{M}{(d+2)K} \right\rfloor.$$

Now take the index set $I_{t,p}$ where $t = p = N$, i.e. $I_{N,N} =$

Algorithm 1 Full accuracy Fourier reconstruction of piecewise smooth functions

Let $f \in PC(d, K)$, and assume that $f = \Phi^{(d)} + \Psi$ where $\Phi^{(d)}$ is the piecewise polynomial absorbing all discontinuities of f, and $\Psi \in C^d$.

1) Obtain initial approximations for $\{x_1, \ldots, x_K\}$ by any standard method (i.e. Eckhoff's method of order zero).
2) Localize each x_j by multiplying with a mollifier (convolution in Fourier domain).
3) Solve resulting Prony system with $K = 1$ and $t = p = \left\lfloor \frac{M}{d+2} \right\rfloor$ (decimation).
4) Take the final approximation to be

$$\widetilde{f} = \widetilde{\Phi}\left(\{\widetilde{a}_{\ell,j}, \widetilde{x}_j\}\right) + \sum_{|k| \leq M} \left\{ c_k(f) - \frac{1}{2\pi} \sum_{j=1}^{K} \mathrm{e}^{-\imath \widetilde{x}_j k} \sum_{\ell=0}^{d} \frac{\widetilde{a}_{\ell,j}}{(\imath k)^{\ell+1}} \right\} \mathrm{e}^{\imath k x}.$$

$\{N, 2N, \ldots, M\}$. As before, $|\epsilon| \sim M^{-1}$, but now due to decimation we get accuracy $|\Delta x_j| \sim N^{-d-1} N^{-1} \sim M^{-d-2}$. In [3], [7] we develop an algorithm (see Algorithm 1) which in fact attains this accuracy. This result can be summarized as follows.

Theorem 4. *Let $f \in PC(d, K)$, so that $f = \Phi^{(d)} + \Psi$ where $\Phi^{(d)}$ is the piecewise polynomial with Fourier coefficients* (5)*, and $\Psi \in C^d$. Assume that there exist constants J, A, B, R such that*

$$\begin{aligned} \min_{i \neq j} |x_i - x_j| &\geq J > 0, \quad |c_k(\Psi)| \leq R \cdot k^{-d-2}, \\ |a_{\ell,j}| &\leq A < \infty, \quad |a_{0,j}| \geq B > 0. \end{aligned}$$

Then the approximation $\widetilde{f}$ obtained by Algorithm 1 satisfies for $M \gg 1$

$$\begin{aligned} |\widetilde{x}_j - x_j| &\leq C_1(d, K, J, A, B, R) \cdot M^{-d-2}; \\ |\widetilde{a}_{\ell,j} - a_{\ell,j}| &\leq C_2(d, K, J, A, B, R) \cdot M^{\ell-d-1}, \quad 0 \leqslant \ell \leqslant d; \\ \left|\widetilde{f}(x) - f(x)\right| &\leq C_3(d, K, J, A, B, R) \cdot M^{-d-1}. \end{aligned}$$

Note that the pointwise bound $\left|f(x) - \widetilde{f}(x)\right|$ is valid "away from discontinuities". Some numerical experiments, elaborated in [3], [7], confirm these theoretical accuracy predictions.

IV. Future work

Stable solution of Prony-type systems in the most general setting must take into account the possibility of colliding nodes. We believe that a reparametrization of the equations in the basis of finite differences is a promising approach to this problem. We have obtained initial results in [5], [20], and plan to continue in this direction.

The Fourier inversion problem for piecewise-analytic functions is still widely open (see e.g. [1]). While our results provide spectral convergence in this setting, it is still unknown if the algebraic method can be pushed to exponential or at least root-exponential accuracy.

Edge detection from spectral data is a well-researched problem, see e.g. [14], [15] and references therein. We expect that the 1D procedure can be generalized to treat the general case via some form of a "separation", or "slice reconstruction" (see e.g. [4] for an example of such a method, dealing with reconstruction from moments).

References

[1] B. Adcock, A.C. Hansen, and A. Shadrin. A stability barrier for reconstructions from Fourier samples. *arXiv preprint arXiv:1210.7831*, 2012.

[2] R. Badeau, B. David, and G. Richard. Performance of ESPRIT for estimating mixtures of complex exponentials modulated by polynomials. *Signal Processing, IEEE Transactions on*, 56(2):492–504, 2008.

[3] D. Batenkov. Complete Algebraic Reconstruction of Piecewise-Smooth Functions from Fourier Data. *arXiv preprint arXiv:1211.0680*.

[4] D. Batenkov, V. Golubyatnikov, and Y. Yomdin. Reconstruction of Planar Domains from Partial Integral Measurements. In *Proc. Complex Analysis & Dynamical Systems V*, 2011.

[5] D. Batenkov and Y. Yomdin. Geometry and Singularities of the Prony Mapping. *preprint*.

[6] D. Batenkov and Y. Yomdin. Algebraic Reconstruction of piecewise-smooth functions from Fourier data. In *Proceedings of the 9th International Conference on Sampling Theory and Applications (SAMPTA)*, 2011.

[7] D. Batenkov and Y. Yomdin. Algebraic Fourier reconstruction of piecewise smooth functions. *Mathematics of Computation*, 81:277–318, 2012.

[8] D. Batenkov and Y. Yomdin. On the accuracy of solving confluent Prony systems. *SIAM J. Appl. Math.*, 73(1):134–154, 2013.

[9] E. Candes and C. Fernandez-Granda. Towards a mathematical theory of super-resolution. *To appear in Communications on Pure and Applied Mathematics*, 2012.

[10] D.L. Donoho. Superresolution via sparsity constraints. *SIAM Journal on Mathematical Analysis*, 23(5):1309–1331, 1992.

[11] D.L. Donoho, M. Elad, and V.N. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *Information Theory, IEEE Transactions on*, 52(1):6–18, 2006.

[12] P.L. Dragotti, M. Vetterli, and T. Blu. Sampling Moments and Reconstructing Signals of Finite Rate of Innovation: Shannon meets Strang-Fix. *IEEE Transactions on Signal Processing*, 55(5):1741, 2007.

[13] K.S. Eckhoff. Accurate reconstructions of functions of finite regularity from truncated Fourier series expansions. *Mathematics of Computation*, 64(210):671–690, 1995.

[14] S. Engelberg and E. Tadmor. Recovery of edges from spectral data with noise - a new perspective. *SIAM Journal on Numerical Analysis*, 46(5):2620–2635, 2008.

[15] L. Greengard and C. Stucchio. Spectral edge detection in two dimensions using wavefronts. *Applied and Computational Harmonic Analysis*, 30(1):69–95, 2011.

[16] Abdul J. Jerri, editor. *Advances in the Gibbs Phenomenon*. Σ Sampling Publishing, 2011.

[17] D. Potts and M. Tasche. Parameter estimation for exponential sums by approximate Prony method. *Signal Processing*, 90(5):1631–1642, 2010.

[18] R. Prony. Essai experimental et analytique. *J. Ec. Polytech.(Paris)*, 2:24–76, 1795.

[19] P. Stoica and N. Arye. MUSIC, maximum likelihood, and Cramer-Rao bound. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 37(5):720–741, 1989.

[20] Y. Yomdin. Singularities in Algebraic Data Acquisition. In M. Manoel, M.C.R. Fuster, and C.T.C. Wall, editors, *Real and Complex Singularities*. Cambridge University Press, 2010.

Super-resolution via superset selection and pruning

Laurent Demanet
Department of Mathematics
Massachusetts Institute of Technology
Cambridge, MA 02139
Email: laurent@math.mit.edu

Deanna Needell
Department of Mathematics
Claremont McKenna College
Claremont, CA 91711
Email: dneedell@cmc.edu

Nam Nguyen
Department of Mathematics
Massachusetts Institute of Technology
Cambridge, MA 02139
Email: namnguyen@math.mit.edu

***Abstract*—We present a pursuit-like algorithm that we call the "superset method" for recovery of sparse vectors from consecutive Fourier measurements in the super-resolution regime. The algorithm has a subspace identification step that hinges on the translation invariance of the Fourier transform, followed by a removal step to estimate the solution's support. The superset method is always successful in the noiseless regime (unlike ℓ_1 minimization) and generalizes to higher dimensions (unlike the matrix pencil method). Relative robustness to noise is demonstrated numerically.**

Acknowledgments. LD acknowledges funding from the Air Force Office of Scientific Research, the National Science Foundation, and the Alfred P. Sloan Foundation. LD is grateful to Jean-Francois Mercier and George Papanicolaou for early discussions on super-resolution.

I. Introduction

We consider the problem of recovering a sparse vector $x_0 \in \mathbb{R}^n$, or an approximation thereof, from $m \leq n$ contiguous Fourier measurements

$$y = Ax_0 + e, \tag{1}$$

where A is the partial, short and wide Fourier matrix $A_{jk} = e^{2\pi ijk/n}$, $0 \leq j < m$, $-n/2 \leq k < n/2$, n even, and, say, $e \sim N(0, \sigma^2 I_m)$.

When recovery is successful in this scenario of contiguous measurements, we may speak of super-resolution: the spacing between neighboring nonzero components in x_0 can be much smaller than the Rayleigh limit n/m suggested by Shannon-Nyquist theory. But in contrast to the compressed sensing scenario, where the m values of j are drawn at random from $\{0, \ldots, n-1\}$, super-resolution can be arbitrarily ill-posed. Open questions concern not only recovery bounds, but the very algorithms needed to define good estimators.

Various techniques have been proposed in the literature to tackle super-resolution, such as MUSIC [11], Prony's method / finite rate of innovation [8] [1] [13], the matrix pencil method [9], ℓ_1 minimization [7] [5] [3] [2], and greedy pursuits [6].

Prony and matrix pencil methods are based on eigenvalue computations: they work well with exact measurements, but their performance is poorly understood in the presence of noise, and they are not obviously set up in higher dimensions. As for ℓ_1 minimization, there is good evidence that k-sparse *nonnegative* signals can be recovered from only $2k+1$ noiseless Fourier coefficients by imposing the positivity constraint with or without ℓ_1 minimization, see [4] [7] and [5]. The work of [3] extends this result to the continuous setting by using total variation minimization. Recently, Candès and Fernandez-Granda showed that the solution to an ℓ_1-minimization problem with a $\|A^*(y - Ax)\|_1$ misfit will be close to the true signal, assuming that locations of any two consecutive nonzero coefficients are separated by at least four times the super-resolution factor n/m [2]. Such optimization ideas have the advantage of being easily generalizable to higher dimensions. On the flip side, ℓ_1 minimization super-resolution is known to fail on sparse signals with nearby components that alternate signs.

In this paper, we discuss a simple algorithm for solving (1) based on

- subspace identification as in the matrix pencil method, but without the subsequent eigenvalue computation; and
- a removal procedure for tightening the active set, reminding of a step in certain greedy pursuits.

This algorithm can outperform the well-known matrix pencil method, as we show in the numerical section, and it is generalizable to higher dimensions. It is a one-pass procedure that does not suffer from slow convergence in situations of high coherence. We also show that the algorithm provides perfect recovery for the (not combinatorially hard in the Fourier case) noiseless ℓ_0 problem

$$\min_x |\operatorname{supp} x| \qquad \text{s.t.} \qquad Ax = y. \tag{2}$$

II. Noiseless subspace identification

For completeness we start by recalling the classical uniqueness result for (2).

Lemma 1. *Let $x_0 \in \mathbb{R}^n$ with support T such that $m \geq 2|T|$, and let $y = Ax_0$. Then the unique minimizer of (2) is x_0.*

We make use of the following notations. Denote $\operatorname{supp} x_0$ by T, and write A_T for the restriction of A to its columns in T. Let T^c for the complement of T. Let a_k for the k-th column of A. The superscript L is used to denote a restriction of a matrix to its first L rows, as in A_T^L.

The "superset method" hinges on a special property that the partial Fourier matrix A does not share with arbitrary dictionaries: each column a_k is *translation-invariant* in the sense that any restriction of a_k to $s \leq m$ consecutive elements gives rise to the same sequence, up to an overall scalar. In other words, exponentials are eigenfunctions of the translation operator. This structure is important. There is an opportunity cost in ignoring it and treating (1) as a generic compressed sensing problem.

A way to leverage translation invariance is to recognize that it gives access to the *subspace* spanned by the atoms a_k for $k \in T$, such that $y = \sum_{k\in T}(x_0)_k a_k$. Algorithmically, one picks a number $1 < L < m$ and juxtaposes translated copies of (restrictions of) y into the Hankel matrix $Y = \text{Hankel}(y)$, defined as

$$Y = \begin{pmatrix} y_0 & y_1 & \cdots & y_{m-L-1} \\ y_1 & y_2 & \cdots & y_{m-L} \\ \vdots & \vdots & \vdots & \vdots \\ y_{L-1} & y_L & \cdots & y_m \end{pmatrix}.$$

The range of Y is the subspace we seek.

Lemma 2. *If $L \geq |T|$, then the rank of Y is $|T|$, and*

$$\textit{Ran}\, Y = \textit{Ran}\, A_T^L.$$

The lemma suggests a simple recovery procedure in the noiseless case: loop over all the candidate atoms a_k for $-n/2 \leq k < n/2$ and select those for which the angle

$$\angle(a_k^L, \text{Ran}\, Y) = 0. \tag{3}$$

Once the set T is identified, the solution is obtained by solving the determined system

$$A_T x_T = y, \qquad x_{T^c} = 0. \tag{4}$$

This procedure (unsurprisingly) provides a solution to the noise-free ℓ_0 sparse recovery problem (2).

Theorem 3. *Let $x_0 \in \mathbb{R}^n$ with support T such that $m > 2|T|$, and let $y = Ax_0$. Consider x defined by (3) and (4), where the Hankel matrix Y is built with $|T|+1 \leq L \leq m-|T|-1$. Then $x = x_0$.*

The proofs of lemma 2 and theorem 3 hinge on the fact that A has full spark.

The idea of subspace identification is at the heart of a different method, the matrix pencil, which seeks the rank-reducing numbers z of the pencil

$$\overline{Y} - z\underline{Y},$$

where $\overline{Y}$ is Y with its first row removed, and $\underline{Y}$ is Y with its last row removed. These numbers z are computed as the generalized eigenvalues of the couple $(\underline{Y}^*\overline{Y}, \underline{Y}^*\underline{Y})$. z can also be found via solving the eigenvalues of the matrix $\underline{Y}^\dagger\overline{Y}$. When $|T| \leq L \leq m-|T|$, the collection of these generalized eigenvalues includes $e^{2\pi ijk/n}$ for $k \in T$, as well as $m-L-|T|$ zeros. There exist variants that consider a Toeplitz matrix instead of a Hankel matrix, with slightly better numerical stability properties. When $L = |T|$, the matrix pencil method reduces to Prony's method, a numerically inferior choice that should be avoided in practice if possible.

III. Noisy subspace identification

The problem becomes more difficult when the observations are contaminated by noise. In this situation $\text{Ran}\, A_T^L \neq \text{Ran}\, Y$, though in low-noise situations we may still be able to recover T from the indices of the smallest angles $\angle(a_k^L, \text{Ran}\, Y)$.

Proposition 4. *Let $y = y_0 + e$ with $e \sim N(0, \sigma^2 I_m)$, and form the corresponding $L \times (m-L)$ matrices Y and Y_0 as previously. Denote the singular values of Y_0^{m-L} by $s_{n,0}$. Then there exists positive c_1, C_1 and c, such that with probability at least $1 - c_1 m^{-C_1}$,*

$$\sin\angle(a_k^L, \textit{Ran}\, Y) \leq c\,\varepsilon_1 \tag{5}$$

for all indices k in the support set and

$$\varepsilon_1 = \frac{|T|}{\left\|a_k^L\right\|_2} \frac{\sigma\sqrt{L\log m}}{|x_{0_{\min}}|} \sqrt{\frac{|x_{0_{\max}}|}{s_{|T|,0}}}. \tag{6}$$

Proof: Here we sketch the proof of this proposition. We note that $a_k^L \in \text{Ran}\, Y_0$ when k is in the true support. Thus

$$\sin\angle(a_k^L, \text{Ran}\, Y) = \frac{\left\|(I - \mathcal{P}_Y)a_k^L\right\|_2}{\left\|a_k^L\right\|_2} = \frac{\left\|\mathcal{P}_{Y^\perp} a_k^L\right\|_2}{\left\|a_k^L\right\|_2}.$$

Denote the compact singular value decomposition of $A_T^L = US^LV^*$. Recalling that $a_k^L \in \text{Ran}\, Y_0$ and a well-known fact that $Y_0 = A_T^L D(A_T^{m-L})^*$ where $D = \text{diag}((x_0)_T)$, we can write $a_k^L = U\alpha = \sum_{i=1}^{|T|} \alpha_i u_i$. Thus,

$$\sin\angle(a_k^L, \text{Ran}\, Y) \leq \sum_{i=1}^{|T|} |\alpha_i| \frac{\|\mathcal{P}_{Y^\perp} u_i\|_2}{\left\|a_k^L\right\|_2}. \tag{7}$$

Next, since $Y = Y_0 + E = A_T^L D(A_T^{m-L})^* + E$, we have $Y[D(A_T^{m-L})^*]^\dagger = A_T^L + E[D(A_T^{m-L})^*]^\dagger$ where $A^\dagger$ is the pseudo-inverse matrix of A. By multiplying both sides by $(\mathcal{P}_{Y^\perp} u_i)^*$, we get

$$(\mathcal{P}_{Y^\perp} u_i)^* Y[D(A_T^{m-L})^*]^\dagger = (\mathcal{P}_{Y^\perp} u_i)^* \left(A_T^L + E[D(A_T^{m-L})^*]^\dagger\right).$$

Since the vector $\mathcal{P}_{Y^\perp} u_i$ is orthogonal to $\text{Ran}\, Y$, the left hand side is zero. Thus multiplying both sides by v_i, the i-th right singular vector of A_T^L, we have

$$0 = (\mathcal{P}_{Y^\perp} u_i)^* A_T^L v_i + (\mathcal{P}_{Y^\perp} u_i)^* E[D(A_T^{m-L})^*]^\dagger v_i.$$

We can see that $(\mathcal{P}_{Y^\perp} u_i)^* A_T^L v_i = (\mathcal{P}_{Y^\perp} u_i)^* s_i^L u_i = s_i^L \|\mathcal{P}_{Y^\perp} u_i\|_2^2$ where s_i^L is the i-th singular value of A_T^L. We therefore obtain

$$\begin{aligned} s_i^L \|\mathcal{P}_{Y^\perp} u_i\|_2^2 &= -(\mathcal{P}_{Y^\perp} u_i)^* E[D(A_T^{m-L})^*]^\dagger v_i \\ &\leq \|(\mathcal{P}_{Y^\perp} u_i\|_2 \|E\| \left\|D^\dagger\right\| \left\|[(A_T^{m-L})^*]^\dagger\right\|. \end{aligned}$$

This leads to the upper bound

$$\begin{aligned} \|\mathcal{P}_{Y^\perp} u_i\|_2 &\leq \frac{1}{s_i^L} \|E\| \left\|D^\dagger\right\| \left\|[(A_T^{m-L})^*]^\dagger\right\| \\ &= \frac{\|E\|}{s_i^L} \frac{1}{|x_{0_{\min}}|} \frac{1}{s_{|T|}^{m-L}}, \end{aligned} \tag{8}$$

where $s_{|T|}^{m-L}$ is the smallest singular value of A_T^{m-L}.

Recalling that $a_k^L = U\alpha$, we have $\alpha_i = u_i^* a_k^L$. From the SVD of A_T^L, we see that $A_T^L(A_T^L)^* = U(S^L)^2U^*$, so that

$$U^* A_T^L (A_T^L)^* U^* = (S^L)^2.$$

This identity implies that $\left\|u_i^* A_T^L\right\|_2 = s_i^L$, and thus, $|\alpha_i| \le s_i^L$. Combining this result with (8) and (7) yields

$$\sin\angle(a_k^L, \operatorname{Ran} Y) \le |T| \frac{\|E\|}{s_{|T|}^{m-L}|x_{0_{\min}}|} \frac{1}{\left\|a_k^L\right\|_2}. \qquad (9)$$

Using the matrix Bernstein inequality of [12] one obtains that $\|E\| \le \sigma\sqrt{cL\log m}$ with high probability. Finally, writing Y_T^{m-L} as $Y_T^{m-L} = A_T^{m-L} D^{1/2} (D^{1/2})^* (A_T^{m-L})^*$, we have

$$s_{|T|,0} = \min_z \frac{\left\|A_T^{m-L} D^{1/2} z\right\|_2^2}{\|z\|_2^2} = \min_h \frac{\left\|A_T^{m-L} h\right\|_2^2}{\left\|D^{-1/2} h\right\|_2^2}$$
$$\le \min_h \frac{\left\|A_T^{m-L} h\right\|_2^2}{\|h\|_2^2 \, s_{\min}(D^{-1})} \le (s_{|T|}^{m-L})^2 |x_{0_{\max}}|,$$

which completes the proof. ∎

There are a few unknown quantities involving ϵ_1, which can empirically be controlled. The support size T can be estimated by a reasonably large constant, say $m/2$. The dynamic range of the signal can presumably be known if we know in prior the type of underlying signal of interest. The singular value $s_{|T|,0}$ of Y_0^{m-L} can be replaced by that of Y^{m-L} via the simple Weyl's inequality $|s_i - s_{i,0}| \le \|\operatorname{Hankel}(e)\|$, which can in turn be controlled as $O(\sigma\sqrt{L\log m})$ with high probability.

The subspace identification step now gathers all the values of k such that

$$\sin\angle(a_k^L, \operatorname{Ran} Y) \le c\,\varepsilon_1.$$

The resulting set Ω of indices is only expected to be a *superset* of the true support T, with high probability.

A second step is now needed to prune Ω in order to extract T. For this purpose, a loop over k is set up where we test the membership of y in $\operatorname{Ran} A_{\Omega\setminus k}$, the range of A_Ω with the k-th column removed. We are now considering a new set of angles where the roles of y and A are reversed: in a noiseless situation, $k \in T$ if and only if

$$\angle(y, \operatorname{Ran} A_{\Omega\setminus k}) \neq 0.$$

When noise is present, we first filter out the noise off Ω by projecting y onto the range of A_Ω, then estimate $k \in T$ only when the angle is above a certain threshold. It is easier to work directly with projections Π:

$$\|\Pi_\Omega y - \Pi_{\Omega\setminus k} y\| = \sin\angle(\Pi_\Omega y, \operatorname{Ran} A_{\Omega\setminus k}) \, \|\Pi_\Omega y\|.$$

The effect of noise on the left-hand side is as follows.

Proposition 5. *Let $y = y_0 + e$ with $e \sim N(0, \sigma^2 I_m)$. Let $\Pi_\Omega y$ be the projection of y onto Ran A_Ω, and let $\Delta\Pi = \Pi_\Omega - \Pi_{\Omega\setminus k}$. Then there exists $c > 0$ such that, with high probability,*

$$|\,\|\Delta\Pi y\| - \|\Delta\Pi y_0\|\,| \le c\,\varepsilon_2,$$

with $\varepsilon_2 = \sigma$.

Algorithm 1 for the superset method implements the removal step in an iterative fashion, one atom at a time.

Algorithm 1 Superset selection and pruning

input: Partial Fourier matrix $A \in C^{m\times n}$, $y = Ax_0 + e$, parameter L, thresholds ε_1 and ε_2.
initialization: $Y = \operatorname{Hankel}(y) \in \mathbb{C}^{L\times(m-L)}$
support identification
  decompose: $\widetilde{Q}\widetilde{R} = Y\widetilde{E}$, $\widetilde{Q} \in \mathbb{C}^{L\times r}$
  project: $a_k \leftarrow A_{\{k\}}$ (for all k)
  $\gamma_k \leftarrow \left\|a_k - \widetilde{Q}\widetilde{Q}^* a_k\right\| / \|a_k\|$
  $\Omega = \{k : \gamma_k \le \varepsilon_1\}$
while true **do**
  decompose: $QR = A_\Omega E$, $Q \in \mathbb{C}^{m\times|\Omega|}$
  remove: $\forall k \in \Omega$: $Q_{(k)}R_{(k)} = A_{\Omega\setminus k}E_{(k)}$
  $\delta_k \leftarrow \|(Q_{(k)}Q_{(k)}^* - QQ^*)y\|_2$
  $k_0 \leftarrow \operatorname{argmin}_k \delta_k$
  if $\delta_{k_0} < \varepsilon_2$, $\Omega \leftarrow \Omega\setminus k_0$
  else break
end while
output: $\widehat{x} = \operatorname{argmin}_x \|y - A_\Omega x\|$

IV. Experimental Results

In the first simulation, we fix $n = 1000$ and $m = 120$ and construct an n-dimensional signal x_0 whose nonzero components are well separated by at least $4n/m$, a distance equivalent to four times the super-resolution factor n/m. The spike magnitudes are independently set to $\pm 1/\sqrt{29}$ with probability $1/2$. The noise vector e is drawn from $N(0, \sigma^2 I_m)$ with $\sigma = 10^{-3}$. We fix the thresholds ε_1 via (6) with $c = 1$ and $\varepsilon_2 = 10\sigma$. Throughout our simulations, we set $L = \lfloor m/3 \rfloor$. As can be seen from Fig. 1, top row, the recovered signal from the superset method is reasonable, with $\|\widehat{x} - x_0\|_2 = 0.075$, while the reconstruction via ℓ_1-minimization tends to exhibit incorrect clusters around the true spikes.

Our next simulation considers a more challenging signal model with a strongly coherent matrix A. For example, with $n = 1000$ and $m = 120$, the coherence of the matrix A with normalized columns a_i is $\mu = \max_{i\neq j} |\langle a_i, a_j\rangle| = 0.9765$. The signal in this simulation is shown in Fig. 1, bottom row. It consists of five spike clusters: each of the first two clusters consists of a single spike, and each of the last four clusters contains two neighboring spikes. The signs of these neighboring spikes either agree or differ. We set m, σ and ε_2 as in the previous simulation, and we let the constant c in the equation (6) of ε_1 equal to 5. Recovery via the superset method is accurate, while ℓ_1 minimization fails at least with clusters of opposite-sign spikes.

In the next simulation, we consider a signal of size $n = 1000$ which contains two nearby spikes at locations $[100, 101]$ and has magnitudes $1/\sqrt{2}$ and $-1/\sqrt{2}$. We empirically investigate the algorithm's ability to recover the signal from varying measurements $m = \{10, 20, ..., 220\}$ and noise levels

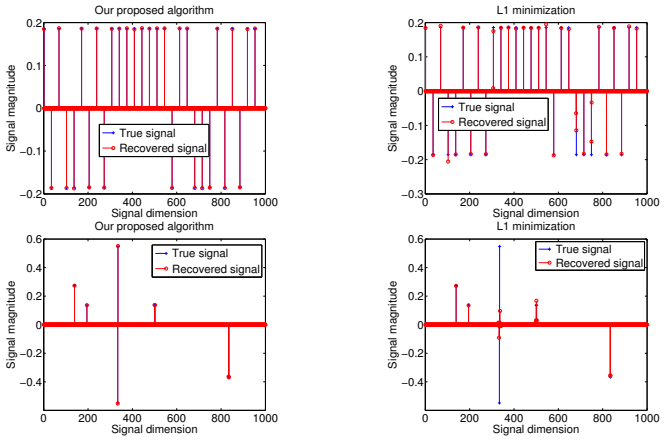


Fig. 1. Original (blue) and recovered (red) signals. Left column: the superset method. Right column: ℓ_1-minimization. Top row: a signal with well-separated spikes. Bottom row: spike spacing below the Rayleigh length.

$log_{10}\sigma = \{-3.5, -3.4, ..., -2\}$. For each pair (m, σ), we report the frequency of success over 100 random realizations of e. The greyscale goes from white (100 successes) to black (100 failures). A trial is declared successful if the recovered $\widehat{x}$ satisfies $\|\widehat{x} - x_0\|_2 / \|x_0\|_2 < 10^{-3}$. The horizontal axis indicates the noise level σ in log scale, and the vertical axis indicates $\log_{10}(1-\mu)$ where μ is the coherence as earlier.

We note that the coherence is inversely proportional to the amount of measurements m and proportional to the super-resolution factor n/m: increasing m (decreasing the super-resolution factor) will reduce the coherence μ. On the vertical axis, smaller values imply higher coherence, or equivalently smaller amount of measurements. As shown in Fig. 2, for reasonably small noise, the algorithm is able to recover the signal exactly even the coherence is nearly 1.

For reference, we also compare the superset method with the matrix pencil method as set up in [10]. The noise is filtered out by preparing low-rank approximations of $\underline{Y}$ and $\overline{Y}$ where only the singular values above $c\sigma\sqrt{L \log L}$ are kept, for some heuristically optimized constant c. Two more signals are considered: (1) a 3-sparse signal consisting of three neighboring spikes, each of magnitude $1/\sqrt{3}$ with alternating signs, and (2) a 4-sparse signal with neighboring spikes of alternating signs and equal magnitude $1/2$. Fig. 2 is a good illustration of the contrasting numerical behaviors of the two methods: the matrix pencil is often the better method in the special case of a signal with 2 spikes, but loses ground to the superset method in various cases of progressively less sparse signals. Understanding the performance of the matrix pencil would require formulating a lower bound on the (typically extremely small) S-th eigenvalues of Y_0 where S is the sparsity of y_0.

V. Conclusion

Empirical evidence is presented for the potential of the superset method as a viable computational method for super-resolution. Further theoretical justifications will be presented elsewhere.

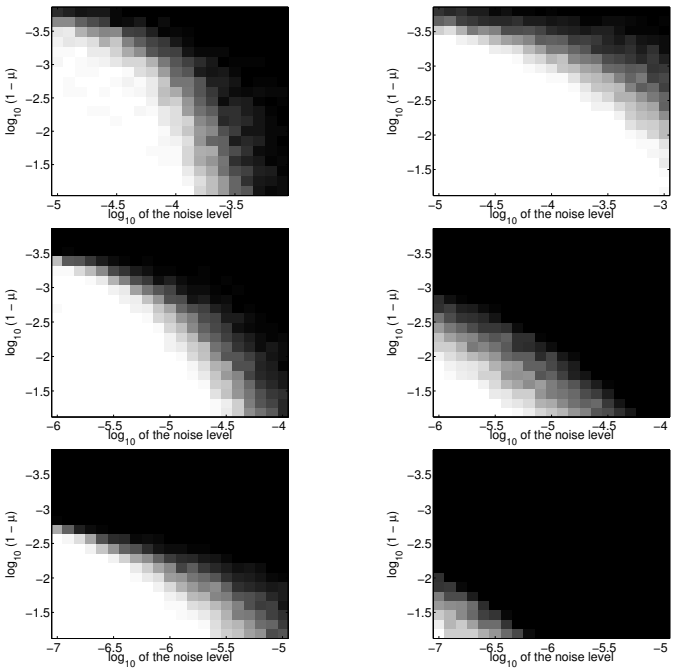


Fig. 2. Probability of recovery, from 1 (white) to 0 (black) for the superset method (left column) and the matrix pencil method (right column). Top row: 2-sparse signal. Middle row: 3-sparse signal. Bottom row: 4-sparse signal. The plots show recovery as a function of the noise level (x-axis, $\log_{10}\sigma$) and the coherence (y-axis, $\log_{10}(1-\mu)$).

References

[1] V.M. Adamjan, D.Z. Arov, and MG Krein. Analytic properties of schmidt pairs for a hankel operator and the generalized schur-takagi problem. *Sb. Math.*, 15(1):31–73, 1971.

[2] E. Candès and C. Fernandez-Granda. Towards a mathematical theory of super-resolution. *Commun. Pure Appl. Math.* To appear.

[3] Y. de Castro and F. Gamboa. Exact reconstruction using Beurling Minimal Extrapolation. *J. Math. Anal. Appl.*, 395(1):336–354, 2012.

[4] D.L. Donoho, I.M. Johnstone, J.C. Hoch, and A.S. Stern. Maximum entropy and the nearly black object. *J. Roy. Stat. Soc. B Met.*, pages 41–81, 1992.

[5] D.L. Donoho and J. Tanner. Sparse nonnegative solutions of underdetermined linear equations by linear programming. In *Proc. Nation. Acad. Scien.*, page 94469451, 2005.

[6] A. Fannjiang and W. Liao. Coherence-pattern guided compressive sensing with unresolved grids. *IAM J. Imaging Sci.*, 5:179–202, 2012.

[7] J.J. Fuchs. Sparsity and uniqueness for some specific underdetermined linear systems. In *Proc. of IEEE ICASSP*, page 729732, Philadelphia, PA, USA, 2005. IEEE.

[8] U. Grenander and G. Szegő. *Toeplitz forms and their applications*. U. California Press, Berkeley, 1958.

[9] Y. Hua and T.K. Sarkar. Matrix pencil method for estimating parameters of exponentially damped/undamped sinusoids in noise. 38(5):814–824, 1990.

[10] Y. Hua and T.K. Sarkar. On svd for estimating generalized eigenvalues of singular matrix pencil in noise. *IEEE T. Signal Proces.*, 39(4):892–900, 1991.

[11] R. O. Schmidt. Multiple emitter location and signal parameter estimation. *IEEE Trans. Atten. Prop.*, 34(3):276–280, Apr. 1986.

[12] J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*, 12(4):389–434, 2012.

[13] M. Vetterli, P. Marziliano, and T. Blu. Sampling signals with finite rate of innovation. *IEEE T. Signal Proces.*, 50(6):1417–1428, 2002.

Support detection in super-resolution

Carlos Fernandez-Granda
Department of Electrical Engineering, Stanford University
Stanford, CA, USA
Email: cfgranda@stanford.edu

***Abstract*—We study the problem of super-resolving a superposition of point sources from noisy low-pass data with a cut-off frequency f_c. Solving a tractable convex program is shown to locate the elements of the support with high precision as long as they are separated by $2/f_c$ and the noise level is small with respect to the amplitude of the signal.**

I. Introduction

The problem of super-resolution is of great importance in applications where the measuring process imposes a physical limit on the resolution of the available measurements. It is often the case that the signal of interest is well modeled as a superposition of point sources. Motivated by this, we consider a signal

$$x = \sum_j a_j \delta_{t_j}, \tag{I.1}$$

consisting of a train of Dirac measures with complex amplitudes a_j located at different locations $\{t_j\}$ in the unit interval. Our aim is to estimate x from the lower end of its spectrum in the form of $n = 2f_c + 1$ Fourier series coefficients (f_c is an integer) perturbed by noise,

$$\begin{aligned} y(k) &= \int_0^1 e^{-i2\pi kt} x(\mathrm{d}t) + z(k) \\ &= \sum_j a_j e^{-i2\pi k t_j} + z(k), \end{aligned} \tag{I.2}$$

for $k \in \mathbb{Z}$, $|k| \leq f_c$. To ease notation, we write (I.2) as $y = \mathcal{F}_n\, x + z$. We model the perturbation $z \in \mathbb{C}^n$ as having bounded ℓ_2 norm,

$$||z||_2 \leq \delta. \tag{I.3}$$

The noise is otherwise arbitrary and can be adversarial.

Even if the signal x is very sparse, without further conditions to ensure that the support of x is not too clustered the super-resolution problem is hopelessly ill-posed. This can be checked numerically, but also formalized thanks to the seminal work of Slepian [4] on discrete prolate spheroidal sequences (see Section 3.2 of [3]). To avoid such extreme ill-posedness, we impose a lower bound on the minimum separation between the elements of the support of the signal.

Definition 1.1 (Minimum separation): Let $\mathbb{T}$ be the circle obtained by identifying the endpoints on $[0, 1]$. For a family of points $T \subset \mathbb{T}$, the minimum separation is the closest distance between any two elements of T,

$$\Delta(T) = \inf_{(t,t') \in T \,:\, t \neq t'} |t - t'|. \tag{I.4}$$

To recover x we propose minimizing the total variation of the estimate, a continuous analog to the ℓ_1 norm for discrete signals (see Appendix A in [3] for a rigorous definition), subject to data constraints:

$$\min_{\tilde{x}} ||\tilde{x}||_{\mathrm{TV}} \quad \text{subject to} \quad ||\mathcal{F}_n \tilde{x} - y||_2 \leq \delta, \tag{I.5}$$

where the minimization is carried out over the set of all finite complex measures $\tilde{x}$ supported on $[0, 1]$. For details on how to solve (I.5) using semidefinite programming see [2], [3].

Previous work established that if

$$\Delta(T) \geq \frac{2}{f_c} := 2\lambda_c \tag{I.6}$$

TV-norm minimization achieves exact recovery in a noiseless setting [3]. Additionally, [2] characterized the reconstruction error for noisy measurements as the target resolution increases. In this work we study support detection using this method. If the original signal contains a spike of a certain amplitude we ask: *How accurately can we recover the position of the spike? How does the accuracy depend on the noise level, the amplitude of the spike and the amplitude of the signal at other locations?* These questions are not addressed by previous work and answering them requires non-trivial modifications to the arguments in [2] and [3]. Our main result establishes that convex programming is in fact a powerful method for support detection in super-resolution.

Theorem 1.2: Consider the noise model (I.3) *and assume the support T satisfies the minimum-separation condition* (I.6)*. The solution to problem* (I.5)[1]

$$\hat{x} = \sum_{\hat{t}_k \in \hat{T}} \hat{a}_k \delta_{\hat{t}_k}$$

[1]This solution can be shown to be an atomic measure with discrete support under very general conditions.

with support $\widehat{T}$ obeys the properties

(i): $\left|a_j - \sum_{\left\{\hat{t}_l \in \widehat{T}:\, \left|\hat{t}_l - t_j\right| \leq c\lambda_{\mathrm{c}}\right\}} \hat{a}_l\right| \leq C_1\delta \quad \forall t_j \in T,$

(ii): $\sum_{\left\{\hat{t}_l \in \widehat{T},\, t_j \in T:\, \left|\hat{t}_l - t_j\right| \leq c\lambda_{\mathrm{c}}\right\}} |\hat{a}_l| \left(\hat{t}_l - t_j\right)^2 \leq C_2\lambda_{\mathrm{c}}^2\delta,$

(iii): $\sum_{\left\{\hat{t}_l \in \widehat{T}:\, \left|\hat{t}_l - t_j\right| > c\lambda_{\mathrm{c}}\ \forall t_j \in T\right\}} |\hat{a}_l| \leq C_3\delta,$

where C_1, C_2 and C_3 are positive numerical constants and $c = 0.1649$.

Properties (i) and (ii) show that the estimate clusters tightly around each element of the signal, whereas (iii) ensures that any spurious spikes detected by the algorithm have small amplitude. These bounds are essentially optimal for the case of adversarial noise, which can be highly concentrated. An intriguing consequence of our result is a bound on the support-detection error for a single spike that does not depend on the value of the signal at other locations.

Corollary 1.3: Under the conditions of Theorem 1.2, for any element t_i in the support of x such that $a_i > C_1\delta$ there exists an element $\hat{t}_i$ in the support of the estimate $\hat{x}$ satisfying

$$\left|t_i - \hat{t}_i\right| \leq \sqrt{\frac{C_2\delta}{|a_i| - C_1\delta}}\lambda_c.$$

Despite the aliasing effect of the low-pass filter applied to the signal, the bound on the support-detection error *only depends on the amplitude of the corresponding spike* (and the noise level). This does not follow from previous analysis. In particular, the bound on the weighted $\mathcal{L}_1$ norm of the error derived in [2] does not allow to derive such local guarantees. A recent paper bounds the support-detection error of a related convex program in the presence of stochastic noise, but the bound depends on the amplitude of the solution rather than on the amplitude of the original spike [1]. As we explain below, obtaining detection guarantees that only depend on the amplitude of the spike of interest is made possible by the existence of a certain low-frequency polynomial, constructed in Lemma 2.2. This is the main technical contribution of the paper.

II. Proof of Theorem 1.2

We begin with an intermediate result proved in Section II-A.

Lemma 2.1: Under the assumptions of Theorem 1.2

$$\sum_{\hat{t}_k \in \hat{T}} |\hat{a}_k| \min\left\{C_a, \frac{C_b\, d\left(\hat{t}_k, T\right)}{\lambda_c^2}\right\} \leq 2\delta,$$

where C_a and C_b are positive numerical constants and

$$d\left(t, T\right) := \min_{t_i \in T} \left(t - t_i\right)^2.$$

Properties (ii) and (iii) are direct corollaries of Lemma (2.1). To establish property (i) we need the following key lemma, proved in Section II-B.

Lemma 2.2: Suppose T obeys condition (I.6) *and $f_c \geq 10$. Then for any $t_j \in T$ there exists a low-pass polynomial*

$$q_{t_j}(t) = \sum_{k=-f_{\mathrm{c}}}^{f_{\mathrm{c}}} b_k e^{i2\pi kt},$$

$b \in \mathbb{C}^n$, such that $|q_{t_j}(t)| < 1$ for all $t \neq t_j$ and

$$q_{t_j}(t_j) = 1, \qquad q_{t_j}(t_l) = 0 \quad t_l \in T \setminus \{t_j\},$$

$$|1 - q_{t_j}(t)| \leq \frac{C_1'\left(t - t_j\right)^2}{\lambda_{\mathrm{c}}^2} \quad \textit{for} \quad |t - t_j| \leq c\lambda_{\mathrm{c}}, \tag{II.1}$$

$$|q_{t_j}(t)| \leq \frac{C_1'\left(t - t_l\right)^2}{\lambda_{\mathrm{c}}^2} \quad \textit{for} \quad t_l \in T \setminus \{t_j\},\ |t - t_l| \leq c\lambda_{\mathrm{c}}, \tag{II.2}$$

$$|q_{t_j}(t)| < C_2' \quad \textit{if} \quad |t - t_l| > c\lambda_{\mathrm{c}}\ \forall\, t_l \in T, \tag{II.3}$$

where $0 < c^2C_2' \leq C_1' < 1$.

The polynomial q_{t_j} provided by this lemma is designed to satisfy $\int_{\mathbb{T}} q_{t_j}(t)x(\mathrm{d}t) = a_j$ and vanish on the rest of the support of the signal. This allows to decouple the estimation error at t_j from the amplitude of the rest of the spikes. Since x and $\hat{x}$ are feasible for (I.5), we can apply Parseval's Theorem and the Cauchy-Schwarz inequality to obtain

$$\begin{aligned}\left|\int_{\mathbb{T}} q_{t_j}(t)x(\mathrm{d}t) - \int_{\mathbb{T}} q_{t_j}(t)\hat{x}(\mathrm{d}t)\right| &= \Big|\sum_{k=-f_c}^{f_c} b_k \mathcal{F}_n(x - \hat{x})_k\Big| \\ &\leq \left|\left|q_{t_j}\right|\right|_{\mathcal{L}_2} ||\mathcal{F}_n(x - \hat{x})||_2 \leq 2\delta, \end{aligned} \tag{II.4}$$

where we have used that the absolute value and consequently the $\mathcal{L}_2$ norm of q_{t_j} is bounded by one. In addition, by Lemmas 2.2 and 2.1 we have

$$\begin{aligned}&\Big|\sum_{\left\{k:\, \left|\hat{t}_k - t_j\right| > c\lambda_c\right\}} \hat{a}_k q_{t_j}(\hat{t}_k) + \sum_{\left\{k:\, \left|\hat{t}_k - t_j\right| \leq c\lambda_c\right\}} \hat{a}_k\left(q_{t_j}(\hat{t}_k) - 1\right)\Big| \\ &\leq \sum_{\left\{k:\, \left|\hat{t}_k - t_j\right| > c\lambda_c\right\}} |\hat{a}_k|\left|q_{t_j}(\hat{t}_k)\right| \\ &\quad + \sum_{\left\{k:\, \left|\hat{t}_k - t_j\right| \leq c\lambda_c\right\}} |\hat{a}_k|\left|1 - q_{t_j}(\hat{t}_k)\right| \\ &\leq \sum_{\hat{t}_k \in \hat{T}} |\hat{a}_k| \min\left\{C_2', \frac{C_1' d\left(\hat{t}_k, T\right)}{\lambda_c^2}\right\} \leq C\delta, \end{aligned} \tag{II.5}$$

for a positive numerical constant C. Finally, Lemma 2.2, the triangle inequality, (II.4) and (II.5)

yield

$$\begin{aligned}&\Big|a_j - \sum_{\left\{k:\,|\hat{t}_k - t_j|\le c\lambda_c\right\}} \hat{a}_k\Big| \\ &= \Bigg| \int_{\mathbb{T}} q_{t_j}(t)x(\mathrm{d}t) - \int_{\mathbb{T}} q_{t_j}(t)\hat{x}(\mathrm{d}t) \\ &\quad + \sum_{\left\{k:\,|\hat{t}_k - t_j|> c\lambda_c\right\}} \hat{a}_k q_{t_j}(\hat{t}_k) \\ &\quad + \sum_{\left\{k:\,|\hat{t}_k - t_j|\le c\lambda_c\right\}} \hat{a}_k \left(q_{t_j}(\hat{t}_k) - 1\right)\Bigg| \le C'\delta.\end{aligned}$$

for a positive numerical constant C'.

A. Proof of Lemma 2.1

The proof relies on a low-pass polynomial provided by Proposition 2.1 and Lemma 2.5 in [3].

Lemma 2.3: Let T obey (I.6)*. For any $v \in \mathbb{C}^{|T|}$ such that $|v_j| = 1$ for all entries v_j there exists a low-pass polynomial $q(t) = \sum_{k=-f_c}^{f_c} d_k e^{i2\pi kt}$, $d \in \mathbb{C}^n$, satisfying*

$$\begin{aligned}&q(t_j) = v_j, \qquad t_j \in T, \\ &|q(t)| < 1 - C_a, \qquad |t - t_j| > c\lambda_{\mathrm{c}} \quad \forall t_j \in T, \\ &|q(t)| \le 1 - \frac{C_b\,(t - t_j)^2}{\lambda_{\mathrm{c}}^2}, \qquad |t - t_j| \le c\lambda_{\mathrm{c}},\, t_j \in T,\end{aligned}$$

with $0 < c^2 C_b \le C_a < 1$.

We set $v_j = \overline{a_j}/\,|a_j|$. The lemma implies

$$\begin{aligned}&\int_{\mathbb{T}} q(t)\hat{x}\,(\mathrm{d}t) \le \sum_k |\hat{a}_k|\,\left|q(\hat{t}_k)\right| \\ &\le \sum_k \left(1 - \min\left\{C_a, \frac{C_b\,d\left(\hat{t}_k, T\right)}{\lambda_c^2}\right\}\right)|\hat{a}_k|\,.\end{aligned} \tag{II.6}$$

The same argument used to prove (II.4) yields

$$\left|\int_{\mathbb{T}} q(t)\hat{x}(\mathrm{d}t) - \int_{\mathbb{T}} q(t)x(\mathrm{d}t)\right| \le 2\delta.$$

Now, taking into account that $\int_{\mathbb{T}} q(t)x\,(\mathrm{d}t) = ||x||_{\mathrm{TV}}$ by construction and $||\hat{x}||_{\mathrm{TV}} \le ||x||_{\mathrm{TV}}$, we have

$$\begin{aligned}&\int_{\mathbb{T}} q(t)\hat{x}\,(\mathrm{d}t) \\ &= \int_{\mathbb{T}} q(t)x\,(\mathrm{d}t) + \int_{\mathbb{T}} q(t)\hat{x}(\mathrm{d}t) - \int_{\mathbb{T}} q(t)x(\mathrm{d}t) \\ &\ge ||x||_{\mathrm{TV}} - 2\delta \ge ||\hat{x}||_{\mathrm{TV}} - 2\delta = \sum_k |\hat{a}_k| - 2\delta.\end{aligned}$$

Combining this with (II.6) completes the proof.

B. Proof of Lemma 2.2

We use a low-frequency kernel and its derivative to construct the desired polynomial exploiting the assumption that the support satisfies the minimum-separation condition (I.6). More precisely, we set

$$q_{t_j}(t) = \sum_{t_k \in T} \alpha_k K(t - t_k) + \beta_k K^{(1)}(t - t_k), \tag{II.7}$$

where $\alpha, \beta \in \mathbb{C}^{|T|}$ are coefficient vectors,

$$K(t) = \left[\frac{\sin\left(\left(\frac{f_c}{2} + 1\right)\pi t\right)}{\left(\frac{f_c}{2} + 1\right)\sin\left(\pi t\right)}\right]^4, \quad t \in \mathbb{T} \setminus \{0\}, \tag{II.8}$$

and $K(0) = 1$; here, $K^{(\ell)}$ is the ℓth derivative of K. Note that K, $K^{(1)}$ and, consequently, q_{t_j} are trigonometric polynomials of the required degree.

We impose $q_{t_j}(t_j) = 1$ and

$$q_{t_j}(t_l) = 0, \quad t_l \in T/\left\{t_j\right\}, \qquad q'_{t_j}(t_k) = 0, \quad t_k \in T.$$

We express these constraints in matrix form. Let $e_{t_j} \in \mathbb{R}^{|T|}$ denote the one-sparse vector with one nonzero entry at the position corresponding to t_j. Then,

$$\begin{bmatrix} D_0 & D_1 \\ D_1 & D_2 \end{bmatrix}\begin{bmatrix}\alpha \\ \beta\end{bmatrix} = \begin{bmatrix} e_{t_j} \\ 0 \end{bmatrix}, \quad \text{where}$$

$$\begin{aligned}&(D_0)_{lk} = K\left(t_l - t_k\right), \quad (D_1)_{lk} = K^{(1)}\left(t_l - t_k\right), \\ &(D_2)_{lk} = K^{(2)}\left(t_l - t_k\right),\end{aligned}$$

and l and k range from 1 to $|T|$. It is shown in Section 2.3.1 of [3] that under the minimum-separation condition this system is invertible. As a result α and β are well defined and q_{t_j} satisfies $q_{t_j}(t_j) = 1$ and $q_{t_j}(t_l) = 0$ for $t_l \in T/\left\{t_j\right\}$. The coefficient vectors can be expressed as

$$\begin{bmatrix}\alpha \\ \beta\end{bmatrix} = \begin{bmatrix} I \\ -D_2^{-1}D_1 \end{bmatrix} S^{-1} e_{t_j}, \quad S := D_0 - D_1 D_2^{-1} D_1,$$

where S is the Schur complement. Let $\|M\|_\infty$ denote the usual infinity norm of a matrix M defined as $\|M\|_\infty = \max_{\|x\|_\infty = 1} \|Mx\|_\infty = \max_l \sum_k |M_{lk}|$. We borrow some results from Section 2.3.1 in [3],

$$\begin{aligned}\left|\left|I - S\right|\right|_\infty &\le 8.747\,10^{-3}, \\ \left|\left|S^{-1}\right|\right|_\infty &\le 1 + 8.824\,10^{-3}, \\ \left|\left|I - S^{-1}\right|\right|_\infty &\le \left|\left|S^{-1}\right|\right|_\infty \left|\left|S - I\right|\right|_\infty \le 8.825\,10^{-3}, \\ \left|\left|\alpha - e_{t_j}\right|\right|_\infty &\le \left|\left|I - S^{-1}\right|\right|_\infty \left|\left|e_{t_j}\right|\right|_\infty \\ &\le 8.825\,10^{-3}, \qquad \text{(II.9)} \\ \left|\left|\beta\right|\right|_\infty &\le 3.294\,10^{-2}\lambda_c. \qquad \text{(II.10)}\end{aligned}$$

Lemma 2.6 in [3] allows to obtain

$$K\left(t\right) \le \frac{1}{\left(f_c t\right)^4} \le 0.333, \quad K'\left(t\right) \le \frac{4\pi}{f_c^3 t^4} \le 4.18\, f_c,$$

for $|t| > c\lambda_c$ as long $f_c \geq 10$. By the same lemma, if we set the minimum separation $\Delta_{\min}$ to $2/f_c$

$$\sum_{t_k \in T \setminus \{t_a, t_b\}} |K(t - t_k)| \leq \sum_{l=0}^{\infty} \frac{1}{\left(f_c \Delta_{\min}(\frac{1}{2} + l)\right)^4} + \sum_{l=0}^{\infty} \frac{1}{\left(f_c \Delta_{\min} l\right)^4} \leq 1.083,$$

$$\sum_{t_k \in T \setminus \{t_a, t_b\}} \left|K^{(1)}(t - t_k)\right| \leq \sum_{l=0}^{\infty} \frac{4\pi}{f_c^3 \left(\Delta_{\min}(\frac{1}{2} + l)\right)^4} + \sum_{j=0}^{\infty} \frac{4\pi}{f_c^3 \left(\Delta_{\min} l\right)^4} \leq 1.75\, f_c,$$

where t_a and t_b are the two spikes nearest to t. Let t_i be the element of $T/\{t_j\}$ that is nearest to t. Combining these inequalities with (II.9) and (II.10) proves that

$$\begin{aligned} |q_{t_j}(t)| &= \Big| \sum_{t_k \in T} \alpha_k K(t - t_k) + \sum_{t_k \in T} \beta_k K^{(1)}(t - t_k) \Big| \\ &\leq |K(t - t_j)| + \left|\left|\alpha - e_{t_j}\right|\right|_\infty \Big(|K(t - t_j)| \\ &\quad + |K(t - t_i)| + \sum_{t_k \in T \setminus \{t_i, t_j\}} |K(t - t_k)| \Big) \\ &\quad + ||\beta||_\infty \Big(\left|K^{(1)}(t - t_j)\right| + \left|K^{(1)}(t - t_i)\right| \\ &\quad + \sum_{t_k \in T \setminus \{t_i, t_j\}} \left|K^{(1)}(t - t_k)\right| \Big) \leq 0.69, \end{aligned}$$

if $|t - t_k| > c\lambda_c$ for all $t_k \in T$ so that (II.3) holds. The proof is completed by two lemmas which prove (II.1) and (II.2) and $|q_{t_j}(t)| < 1$ for any t. They rely on the following bounds borrowed from equation (2.25) in Section 2 of [3],

$$\begin{aligned} K(t) &\geq 0.9539, & K^{(2)}(t) &\leq -2.923\, f_c^2, \\ \left|K^{(1)}(t)\right| &\leq 0.5595\, f_c, & \left|K^{(2)}(t)\right| &\leq 3.393\, f_c^2, \\ \left|K^{(3)}(t)\right| &\leq 5.697\, f_c^3, \end{aligned} \tag{II.11}$$

and on the fact that, due to Lemma 2.7 in [3], for any $t_0 \in T$ and $t \in \mathbb{T}$ obeying $|t - t_0| \leq c\lambda_c$,

$$\sum_{t_k \in T \setminus \{t_0\}} \left|K^{(2)}(t - t_k)\right| \leq 1.06\, f_c^2 \tag{II.12}$$

$$\sum_{t_k \in T \setminus \{t_0\}} \left|K^{(3)}(t - t_k)\right| \leq 18.6\, f_c^3. \tag{II.13}$$

Lemma 2.4: For any t such that $|t - t_j| \leq c\lambda_c$,

$$1 - 4.07\,(t - t_j)^2 f_c^2 \leq q_{t_j}(t) \leq 1 - 2.30\,(t - t_j)^2 f_c^2.$$

Proof We assume without loss of generality that $t_j = 0$. By symmetry, it suffices to show the claim for $t \in (0, c\,\lambda_c]$. By (II.9), (II.10), (II.11), (II.12) and (II.13),

$$\begin{aligned} q_0''(t) &= \sum_{t_k \in T} \alpha_k K^{(2)}(t - t_k) + \sum_{t_k \in T} \beta_k K^{(3)}(t - t_k) \\ &\leq \left(1 + \left|\left|\alpha - e_{t_j}\right|\right|_\infty\right) K^{(2)}(t) \\ &\quad + \left|\left|\alpha - e_{t_j}\right|\right|_\infty \sum_{t_k \in T \setminus \{0\}} \left|K^{(2)}(t - t_k)\right| \\ &\quad + ||\beta||_\infty \Big(\left|K^{(3)}(t)\right| + \sum_{t_k \in T \setminus \{0\}} \left|K^{(3)}(t - t_k)\right| \Big) \\ &\leq -2.30\, f_c^2. \end{aligned}$$

Similar computations yield $|q_0''(t)| \leq 4.07\, f_c^2$. This together with $q_0(0) = 1$ and $q_0'(0) = 0$ implies the desired result. ■

Lemma 2.5: For any $t_l \in T \setminus \{t_j\}$ and t obeying $|t - t_l| \leq c\lambda_c$, we have

$$|q_{t_j}(t)| \leq 16.64\,(t - t_l)^2 f_c^2.$$

Proof We assume without loss of generality that $t_l = 0$ and prove the claim for $t \in (0, c\,\lambda_c]$. By (II.9), (II.10), (II.11), (II.12) and (II.13)

$$\begin{aligned} \left|q_{t_j}''(t)\right| &= \left| \sum_{t_k \in T} \alpha_k K^{(2)}(t - t_k) + \sum_{t_k \in T} \beta_k K^{(3)}(t - t_k) \right| \\ &\leq \left(1 + \left|\left|\alpha - e_{t_j}\right|\right|_\infty\right) \left|K^{(2)}(t - t_j)\right| \\ &\quad + \left|\left|\alpha - e_{t_j}\right|\right|_\infty \left(\left|K^{(2)}(t)\right| + \sum_{t_k \in T \setminus \{0, t_j\}} \left|K^{(2)}(t - t_k)\right| \right) \\ &\quad + ||\beta||_\infty \left(\left|K^{(3)}(t)\right| + \sum_{t_k \in T \setminus \{0\}} \left|K^{(3)}(t - t_k)\right| \right) \\ &\leq 16.64\, f_c^2, \end{aligned}$$

since in the interval of interest $\left|K^{(2)}(t - t_j)\right| \leq \frac{18\pi^2}{f_c^2(\Delta_{\min} - 0.16 f_c)^4} \leq 15.67\, f_c^2$ due to Lemma 2.6 in [3]. This together with $q_{t_j}(0) = 0$ and $q_{t_j}'(0) = 0$ implies the desired result. ■

ACKNOWLEDGEMENTS

This work was supported by a Fundación Caja Madrid Fellowship. The author is grateful to Emmanuel Candès for useful comments regarding this manuscript and for his support.

REFERENCES

[1] J. M. Azais, Y. de Castro and F. Gamboa. Spike detection from innacurate samplings. Preprint.

[2] E. J. Candès and C. Fernandez-Granda. Super-resolution from noisy data. Preprint.

[3] E. J. Candès and C. Fernandez-Granda. Towards a mathematical theory of super-resolution. *Communications on Pure and Applied Mathematics*, to appear.

[4] D. Slepian. Prolate spheroidal wave functions, Fourier analysis, and uncertainty. V - The discrete case. *Bell System Technical Journal*, 57:1371–1430, 1978.

Using Correlated Subset Structure for Compressive Sensing Recovery

Atul Divekar
Alcatel-Lucent
Naperville, IL 60563
Email: atul.divekar@alcatel-lucent.com

Deanna Needell
Department of Mathematics
Claremont McKenna College
Claremont, CA 91711
Email: dneedell@cmc.edu

***Abstract*—Compressive sensing is a methodology for the reconstruction of sparse or compressible signals using far fewer samples than required by the Nyquist criterion. However, many of the results in compressive sensing concern random sampling matrices such as Gaussian and Bernoulli matrices. In common physically feasible signal acquisition and reconstruction scenarios such as super-resolution of images, the sensing matrix has a non-random structure with highly correlated columns. Here we present a compressive sensing recovery algorithm that exploits this correlation structure. We provide algorithmic justification as well as empirical comparisons.**

I. Introduction

Consider the problem of image super-resolution, where one or more low-resolution images of a scene are used to synthesize a single image of higher resolution. If multiple images are used, they are commonly assumed to be subpixel-shifted and downsampled versions of the original high resolution image that is to be reconstructed [1]. Alternatively, super-resolution from a single low resolution image using a dictionary of image patches and compressive sensing recovery has been proposed in [2]. The relationship between the available low resolution and desired high resolution image is commonly modeled by a linear filtering and downsampling operation. Suppose that we wish to reconstruct a size $N \times N$ high resolution image from a lower resolution image, for example of size $\frac{N}{2} \times \frac{N}{2}$, or smaller. Let x and y represent the vectorized high and low resolution images respectively. We model the formation of y from x by the equation $y = SHx + \eta$ where η is the sensor noise, S is a downsampling matrix of size $\frac{N}{2} * \frac{N}{2}$ by N^2, and H is a N^2 by N^2 matrix that represents the filtering (antialiasing) operation. In order to consider super-resolution as a compressive sensing recovery problem we write $x = \Psi c$ where Ψ is a sparsifying basis for the class of images under consideration and c is the coefficient vector corresponding to image x with respect to the basis Ψ. In the simplest case, Ψ is an $N^2 \times N^2$ orthogonal matrix, but can also be generalized to an overcomplete dictionary. Here we have $y = SH\Psi c + \eta = \Phi c + \eta$, where $\Phi = SH\Psi$ is the sampling matrix.

Most of the work in the compressive sensing literature assumes Φ to be random matrix, such as a partial DFT or one drawn from a Gaussian or Bernoulli distribution. However, in this scenario the matrix is not random, but instead has correlated columns whose structure we wish to exploit to improve compressive sensing recovery. Here we assume that H is not a perfect low pass filter, so that it is possible for $\Phi = SH\Psi$ to preserve enough high frequency information for recovery to be possible; SH and Ψ have sufficient incoherency to allow c to be recovered with acceptable error.

Compressed sensing provides techniques for stable sparse recovery [3]–[5], but results for coherent sensing matrices have been limited [6]–[8].

Organization. The structure we wish to exploit is first described. Then we present algorithms that take advantage of this structure for compressive sensing recovery.

II. Correlation Structure

Typical examples of sparsifying bases Ψ for images are wavelets and blockwise discrete cosine transform bases. Images exhibit correlation at each scale: neighboring pixels are heavily correlated except across edges, local averages of neighboring blocks are heavily correlated except across edges, and so on. This makes wavelet-like bases, which have locally restricted atoms, suitable for sparsifying the image. For the super-resolution setting with the low resolution image of size $\frac{N}{2} \times \frac{N}{2}$, the rows of SH consist of shifted versions of the filtering kernel with shifts of 2 horizontally and vertically. Due to the localized nature of wavelet bases, we expect columns of Φ that correspond to spatially distant bases in Ψ to have little correlation. If Ψ is a tree structured orthogonal wavelet basis matrix, columns of Ψ that overlap spatially are orthogonal, however when filtered by H, they result in significant correlation. Then we expect columns in Φ to show significant correlation in tree structured patterns.

We illustrate this with an example. For simplicity we consider only one-dimensional signals, though the discussion is equally valid for images. Suppose that Ψ is a 256×256 matrix whose columns consist of the length 256 Haar basis vectors, and SH is a 128×256 matrix obtained by shifting the filter kernel $h = \{0.1, 0.2, 0.4, 0.2, 0.1\}$ by two from one row to the next. SH represents the filtering and downsampling operation that generates the low resolution signal $y = SHx$ from the length 256 signal x. Then $\Phi = SH\Psi$ is the sampling matrix.

Fig. 1 shows the absolute values of the correlation matrix $C = \Phi^*\Phi$ (here and throughout A^* denotes the adjoint of A). This shows that only a small number of pairs of columns of

Φ are strongly correlated to each other. Each filtered wavelet basis is correlated with other spatially overlapping bases at coarser and finer scale and in the immediate neighborhood, but has no correlation with spatially distant bases.

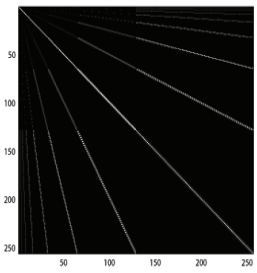


Fig. 1: Absolute values of $\Phi^*\Phi$.

More generally, consider compressive sensing recovery where the columns of the sampling matrix Φ can be grouped into nearly-isolated sets, such that correlation among pairs of columns within a set may be significant, but correlation between two columns that belong to different sets is relatively small. How does one exploit this structure to efficiently reconstruct the signal?

One of the central results in compressive sensing is that if matrix Φ exhibits a property called the Restricted Isometry Property (RIP) [9], [10], convex optimization can recover the sparse signal exactly [11], [12] via

$$\min \ ||c||_1 \text{ such that } y = \Phi c. \tag{1}$$

However, the sampling matrix $\Phi = H\Psi$ described above does not obey the RIP and these results are not readily applicable. On the other hand, it is commonly found in practical applications and has a structure that could be exploited.

Before considering the above problem, a simple modification to CoSaMP [13] is presented that provides some improvement in recovery performance. This algorithm, called Partial Inversion (PartInv) and described by Algorithm 1, also indicates how the above described structure could be exploited.

III. Partial Inversion

Consider the usual CS setting: Given a length M sample vector $y = \Phi c + \eta$ where Φ is an $M \times N$ sampling matrix and c a length N vector with sparsity $K < M$, we wish to obtain the best K-sparse approximation $\hat{c}$ to c. At each step let I be an index set, so that for example, $\hat{c}_I$ represents an estimate of the components of c corresponding to the column indices in I. $\hat{c}$ by itself is an estimate for all the columns $\{1..N\}$. Let L for $K \leq L < M$ be an adjustable parameter for the size of the set I. We get good results with $L = \max\{K, 0.8M\}$. Let Φ_I denote the matrix of columns from Φ corresponding to indices in the set I. Let $\tilde{I} = \{1..N\} \backslash I$ denote the complement of I. For any full rank matrix A, define $A^\dagger = (A^*A)^{-1}A^*$.

For the noiseless case $\eta = 0$, the stopping condition can be obtained by testing the magnitude of $r_2 = y - \Phi\hat{c}$ at the start of each iteration. If set I does not vary from one iteration to the next, the algorithm cannot progress further and can be stopped immediately. In practice the inversion of line 3 can be done efficiently by Richardson's algorithm (see e.g. Sec. 7.2 of [14]).

Algorithm 1 Given $y = \Phi c$, return best K-sparse approximation $\hat{c}$

1: $\hat{c} \leftarrow \Phi^* y; I^0 \leftarrow$ indices of the L-largest magnitudes of $\hat{c}; k \leftarrow 0$
2: **while** Stopping condition not met **do**
3: $\quad \hat{c}_{I^{(k)}} \leftarrow \Phi^\dagger_{I^{(k)}} y$
4: $\quad r \leftarrow y - \Phi_{I^{(k)}} \hat{c}_{I^{(k)}}$
5: $\quad J^{(k)} \leftarrow I^{(k)}$
6: $\quad \hat{c}_{J^{(k)}} \leftarrow \Phi^*_{J^{(k)}} r$
7: $\quad I^{(k+1)} \leftarrow$ indices of the L-largest magnitude components of $\hat{c}$.
8: $\quad k \leftarrow k + 1$
9: **end while**

This algorithm demonstrates improvement relative to CoSaMP when the accurate recovery region is considered on a plot of $\frac{K}{M}$ versus $\frac{M}{N}$. The motivation is the following (for simplicity we drop the iteration indicator k) : From line 3,

$$\hat{c}_I = \Phi^\dagger_I y \tag{2}$$
$$= c_I + (\Phi^*_I \Phi_I)^{-1} \Phi^*_I \Phi_{\tilde{I}} c_{\tilde{I}}. \tag{3}$$

Compare this to the estimator $\hat{c_I} = \Phi^*_I r$ used in CoSaMP. When $r = y$, we have

$$\hat{c_I} = \Phi^*_I y \tag{4}$$
$$= \Phi^*_I \Phi_I c_I + \Phi^*_I \Phi_{\tilde{I}} c_{\tilde{I}} \tag{5}$$
$$= c_I + (\Phi^*_I \Phi_I - I) c_I + \Phi^*_I \Phi_{\tilde{I}} c_{\tilde{I}} \tag{6}$$

If the index set I contains several nonzero coefficients (which we hope is true), then $(\Phi^*_I \Phi_I - I)c_I$, which results from the mutual interference between the columns of Φ_I, is significant and is a source of noise in $\hat{c_I}$. This term is eliminated in (2). Partial inversion does add $(\Phi^*_I \Phi_I)^{-1}$ to the remaining noise term, however, the singular values of this term can be kept from significantly amplifying the noise term by a conservative choice of L, the size of the index set I (for example, empirically we find that $L = s$ tends to be a safe choice, but larger values often lead to noise amplification for certain types of matrices). The improved estimate $\hat{c}_I$ further produces an improved estimate $\hat{c}_{J^{(k)}}$, which leads to a better selection of nonzero coefficients in the next iteration.

The expression (2) also indicates how the correlation structure may be used to improve recovery. The noise term $(\Phi^*_I \Phi_I)^{-1} \Phi^*_I \Phi_{\tilde{I}} c_{\tilde{I}}$ depends upon the correlation between the sets Φ_I and $\Phi_{\tilde{I}}$ given by $\Phi^*_I \Phi_{\tilde{I}}$. This correlation is weak if Φ_I and $\Phi_{\tilde{I}}$ are sufficiently spread.However, the correlation is likely to remain large if L is significant compared to M, as will be the case when $\frac{K}{M}$ is large.

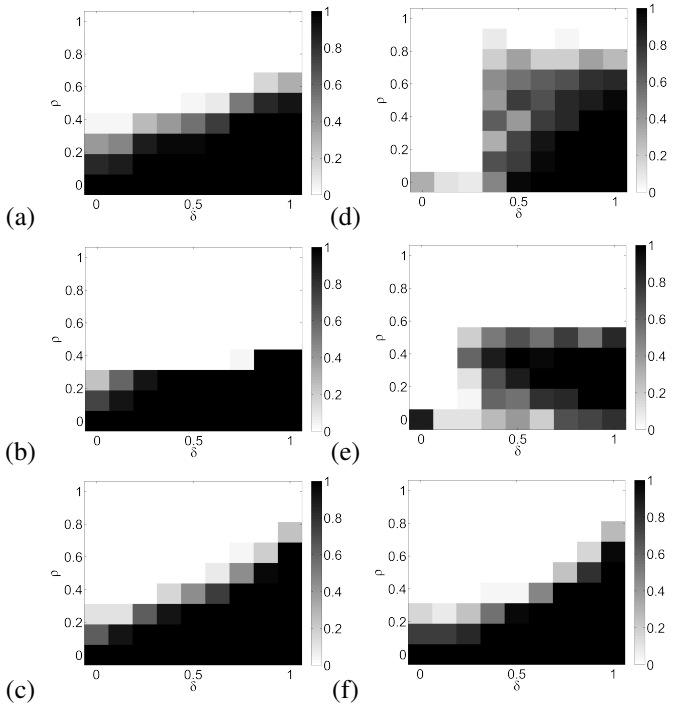


Fig. 2: *Proportion of successes on Gaussian matrices using (a) PartInv, (b) CoSaMP and (c) ℓ_1-minimization, and proportion of successes on correlated column subset matrices using (d) PartInv, (e) CoSaMP and (f) ℓ_1-minimization for various values of $\delta = \frac{M}{N} \in (0,1)$ (horizontal axis) and $\rho = \frac{K}{M} \in (0,1)$ (vertical axis).*

IV. Experimental Comparison

We compare the recovery performance of Partial Inversion with CoSaMP and convex optimization (1) for two classes of matrices: Gaussian random matrices, and matrices constructed to have highly correlated subsets of columns with low correlation across subsets.

In the first case, we construct M by N matrices with $N(0,1)$ elements along with the coefficient vector c containing K nonzero entries taken from a $N(0,1)$ distribution. The nonzero locations are selected uniformly at random from $\{1...N\}$. Each column in each matrix is normalized to have unit l_2 norm. We set $N = 256$ and vary $\delta = \frac{M}{N}$ from 0.1 to 0.9 in steps of 0.1. For each δ we vary $\rho = \frac{K}{M}$ from 0.1 to 0.9 in steps of 0.1. For each (δ, ρ) point we carry out 25 trials, and declare success if $\frac{1}{N}||c - \hat{c}||^2 < 10^{-5}$. For PartInv we considered two cases for the size of subset I : $L = S$ and $L = \max\{S, 0.8M\}$. We see better performance in the $L = S$ case. For l_1 minimization we use the l_1-magic package [15]. We show the results in Fig. 2.

In the second case, we construct M by N matrices with $N = 256$ and variable M and a block diagonal structure. The columns are divided into 16 column subsets. In each subset we set $M/16$ rows to 1. In addition, to every element of the matrix we add noise drawn from a zero-mean normal distribution with variance 0.0025. This produces heavy intra-subset correlation and light correlation across subsets. We let the coefficient vector c contain S nonzeros elements drawn from a $N(0,1)$ distribution. We select 4 of the 16 subsets at random and in each subset select $\frac{S}{4}$ of the indices to have nonzero values, again uniformly at random. If some of the nonzeros were left over, they are accomodated in a fifth subset. For PartInv we set $L = \max\{S, 0.8M\}$. The results are also depicted in Fig. 2.

V. Recovery of Coefficients Concentrated on Wavelet Trees

We next use Partial Inversion to recover nonzero coefficients that are concentrated on wavelet trees, which is commonly seen when a signal or image with discontinuities is decomposed in a wavelet basis. When the coefficients are concentrated on an isolated set (a set of columns that have low correlation with columns outside the set), a setwise estimator is especially useful to identify the sets on which the coefficients are nonzero. Consider the 2D wavelet case. Suppose that I is the index set of columns of the wavelet basis belonging to a particular tree rooted at a coarse scale and containing finer scale coefficients. We have

$$z_I = \Phi_I^* y = \Phi_I^* \Phi_I c_I + \Phi_I^* \Phi_{\tilde{I}} c_{\tilde{I}}. \tag{7}$$

Because Φ_I is relatively isolated from the columns in $\Phi_{\tilde{I}}$, the second term is small, and because most of the elements of c_I are nonzero, the first term is large. This is further intensified by the mutual correlation of the columns of Φ_I which is high because of the spatial overlap of the support of the wavelet bases in the tree. This motivates a simple selection criterion for measuring the strength of the nonzero coefficients in each wavelet tree I: $s_I = \sum_{j \in I} |z_j|$. We use this criterion along with PartInv to select wavelet trees that are known to be nonzero. We denote the number of subsets by SETNUM.

We modify the PartInv algorithm to use this estimator.

Algorithm 2 Given $y = \Phi c$, with K nonzero coefficients concentrated on wavelet trees,return best K-sparse approximation $\hat{c}$

1: $\hat{c} \leftarrow \Phi^* y$;
2: $k \leftarrow -1$
3: **for** $j = 1 \rightarrow$ SETNUM **do**
4: $\quad s_j \leftarrow \sum_{l \in I_j} |\hat{c}_l|$
5: **end for**
6: $I^{k+1} \leftarrow$ indices of columns contained in the sets with the largest magnitude s_k, to include at least K coefficients.
7: $k \leftarrow k + 1$
8: **while** Stopping condition not met **do**
9: $\quad \hat{c}_{I^{(k)}} \leftarrow \Phi^\dagger_{I^{(k)}} y$
10: $\quad r \leftarrow y - \Phi_{I^{(k)}} \hat{c}_{I^{(k)}}$
11: $\quad J^{(k)} \leftarrow \widetilde{I^{(k)}}$
12: $\quad \hat{c}_{J^{(k)}} \leftarrow \Phi^*_{J^{(k)}} r$
13: $\quad$ Repeat lines $3 - 6$
14: $\quad k \leftarrow k + 1$
15: **end while**

VI. Experimental Results

To test this algorithm, we use the Daubechies-5 wavelet basis in two dimensions over 32×32 size patches with 5 levels of decomposition. This gives a size 1024 by 1024 matrix Ψ. We divide this matrix into 49 sets: 1 set of the coarsest scale coefficients in a block of size 4×4 containing the two coarsest scales, and 48 other sets rooted at the coefficients at the next finer scale. Each of these sets contains 21 $(1+4+16)$ coefficients in a quadtree structure. To create matrix Φ we first apply a blurring filter H with a symmetric 5×5 kernel that is close to a delta function. This simulates practical optical sampling acquisition effects such as diffraction and helps prevent rank deficiency problems when carrying out inversion. We use different 2D sampling patterns to carry out the subsampling operation represented by matrix S. Hence the acquisition process is represented by $y = \Phi c$ where $\Phi = SH\Psi$. The sampling patterns are shown in Table I for each sampling rate $\delta = \frac{M}{N}$ used to generate the results. Each pattern is replicated 8 times in horizontal and vertical directions to give the 32×32 sampling pattern used for matrix S. The filter kernel is a 5×5 kernel with 0.29 at the center and 0.02 in other locations. The signals are generated by uniformly selecting at random wavelet trees to make the sparsity of the signal the specified value. The coefficients in these trees are set to values chosen from a standard normal distribution, and the rest are set to zero.

0	0	0	0
0	1	0	0
0	0	0	0
0	0	0	1

(a) $\delta = \frac{2}{16}$

1	0	0	0
0	0	1	0
0	1	0	0
0	0	0	1

(b) $\delta = \frac{4}{16}$

1	0	1	0
0	1	0	1
1	0	0	0
0	0	1	0

(c) $\delta = \frac{6}{16}$

1	0	1	0
0	1	0	1
1	0	1	0
0	1	0	1

(d) $\delta = \frac{8}{16}$

0	1	0	1
1	0	1	0
0	1	1	1
1	1	0	1

(e) $\delta = \frac{10}{16}$

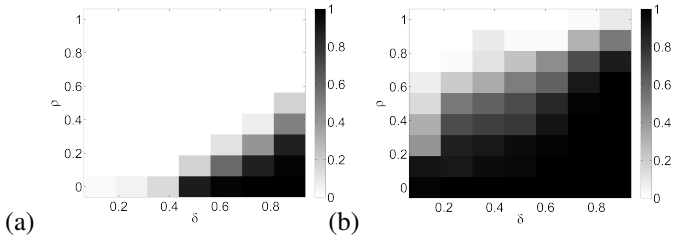
1	1	0	1
0	1	1	1
1	1	1	0
1	0	1	1

(f) $\delta = \frac{12}{16}$

1	1	1	1
1	1	0	1
1	1	1	1
0	1	1	1

(g) $\delta = \frac{14}{16}$

TABLE I: Sampling Patterns



(a) (b)

Fig. 3: *Proportion of successes with nonzero coefficients concentrated on wavelet trees from (a) ℓ_1-minimization and (b) PartInv.*

The results are shown in Fig. 3. For each data point we carry out 100 trials. We declare success if $\frac{1}{N}||c - \hat{c}||^2 < 10^{-5}$ where $N = 32 \times 32$. This shows improvement in selection performance with the sum estimator.

VII. Conclusion

We consider methods of compressive sensing recovery for sampling matrices that have subsets of columns that are strongly intra-correlated, but show low correlation with other subsets. This structure commonly arises in physical sample acquisition/reconstruction scenarios such as image super-resolution. We describe Partial Inversion, an algorithm that improves compressive sensing recovery by removing a source of noise in the initial estimator, and demonstrate its performance by simulations on Gaussian and correlated column subset matrices. We consider compressive sensing recovery when the nonzero coefficients are concentrated on wavelet trees, and demonstrate a simple estimator that improves selection of the trees that carry the nonzero coefficients.

References

[1] S. Farsiu, D. Robinson, M. Elad, and P. Milanfar, "Advances and challenges in super-resolution," *Int. J. Imag. Syst. Tech.*, vol. 14, no. 2, pp. 47–57, 2004.

[2] J. Yang, J. Wright, T. Huang, and Y. Ma, "Image superresolution via sparse representation," *IEEE T. Image Process.*, vol. 19, no. 11, pp. 2861–2873, Nov. 2010.

[3] E. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE T. Inform. Theory*, vol. 52, no. 2, pp. 489–509, 2006.

[4] E. Candes, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Commun. Pure Appl. Math.*, vol. 59, no. 8, pp. 1207–1223, 2006.

[5] D. Donoho and P. Stark, "Uncertainty principles and signal recovery," *SIAM J. Appl. Math.*, vol. 49, no. 3, pp. 906–931, 1989.

[6] E. Candes, Y. Eldar, D. Needell, and P. Randall, "Compressed sensing with coherent and redundant dictionaries," *Appl. Comput. Harmon. A.*, vol. 31, no. 1, pp. 59–73, 2011.

[7] E. Candes and C. Fernandez-Granda, "Towards a mathematical theory of super-resolution," *Preprint*, 2012.

[8] A. Fannjiang and W. Liao, "Coherence pattern-guided compressive sensing with unresolved grids," *SIAM J. Imaging Sci.*, vol. 5, no. 1, pp. 179–202, 2012.

[9] E. Candes and T. Tao, "Decoding by Linear Programming," *IEEE T. Inform. Theory*, vol. 51, no. 12, pp. 4203 – 4215, dec. 2005.

[10] M. Rudelson and R. Vershynin, "On sparse reconstruction from Fourier and Gaussian measurements," *Comm. Pure Appl. Math.*, vol. 61, pp. 1025–1045, 2008.

[11] E. Candes and T. Tao, "Near optimal signal recovery from random projections: Universal encoding strategies?" *IEEE T. Inform. Theory*, vol. 52, pp. 5406–5425, Dec. 2006.

[12] E. Candes, "The Restricted Isometry Property and its implications for Compressed Sensing," *Cr. Acad. Sci. I-Math.*, pp. 589–592, Dec. 2008.

[13] D. Needell and J. Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples," *Appl. Comput. Harmon. A.*, vol. 26, pp. 301–321, 2009.

[14] Å. Björck, *Numerical Methods for Least Squares Problems*. Philadelphia: SIAM, 1996.

[15] ℓ_1-magic. [Online]. Available: http://www.acm.caltech.edu/l1magic/

Sub-Wavelength Coherent Diffractive Imaging based on Sparsity

Yoav Shechtman*, Alexander Szameit*, Eliyahu Osherovich†, Pavel Sidorenko*,Elad Bullkich*,
Hod Dana ‡, Shy Shoham‡, Irad Yavneh†, Michael Zibulevsky†,
Oren Cohen*, Yonina C. Eldar§ and Mordechai Segev*
*Department of Physics Technion, Israel Institute of Technology, Israel 32000
†Computer Science Department, Technion, Israel Institute of Technology, Israel 32000
‡Department of Biomedical Engineering, Technion, Israel Institute of Technology, Israel 32000
§Department of Electrical Engineering, Technion, Israel Institute of Technology, Israel 32000
Email: joe@tx.technion.ac.il

***Abstract*—We propose and experimentally demonstrate a method of performing single-shot sub-wavelength resolution Coherent Diffractive Imaging (CDI), i.e. algorithmic object reconstruction from Fourier amplitude measurements. The method is applicable to objects that are sparse in a known basis. The prior knowledge of the object's sparsity compensates for the loss of phase information, and the loss of all information at the high-spatial frequencies occurring in every microscope and imaging system due to the physics of electromagnetic waves in free-space.**

I. Introduction

Coherent Diffractive Imaging (CDI) is an imaging technique where intricate features are algorithmically reconstructed from measurements of the freely-diffracting intensity pattern ([1], [2]). In CDI, an object is illuminated by a coherent plane wave (LASER light), and the far-field diffraction intensity is measured. That is, the measurements correspond to the absolute value squared of the Fourier components. Recent advances in making lasers in the x-ray regime and in the extreme ultraviolet have made this technique very important for a variety of applications, among them structural biology: mapping out the structure of proteins that cannot be crystalized. However, the physics underlying the propagation of electromagnetic waves acts as a low-pass filter, effectively truncating high Fourier components, and thereby setting a fundamental limit on imaging systems: the finest feature that can be recovered in imaging microscopes is larger than one half of the optical wavelength (the so-called diffraction limit). This stringent limit naturally also limits CDI: the resolution in all current work on CDI is limited by the diffraction limit [3]. Over the past decades, several techniques were developed for sub-wavelength imaging, but none of them works as actual imaging: they all involve scanning or integration over very many acquired images generated by sub-wavelength light sources. These methods include Scanning Near-Field Microscope ([4], [5]), scanning a sub-wavelength "hot spot" ([6], [7], [8]), or ensemble-averaging over multiple experiments with fluorescent particles ([9], [10], [11], [12]). Due to the nature of theses technique – which rely on scanning or averaging - they cannot be used for real-time imaging of dynamics processes (say, a chemical reaction that evolves with time). On the other hand, CDI, being a 'single shot' imaging technique, is suitable for ultra-fast imaging, but it lacks sub-wavelength resolution. Here, we present and demonstrate experimentally a method to enhance CDI resolution beyond the diffraction limit, based on prior knowledge that the object is sparse in a known basis.

II. Problem Formulation

In a typical, plane-wave CDI setting, an object is illuminated by a coherent plane wave, and the far field diffraction pattern intensity is measured. The measured diffraction intensity, in the paraxial approximation, is proportional to the magnitude of the object's Fourier transform, up to the cut-off frequency $1/\lambda$, where λ is the wavelength of the light [3]. Therefore, mathematically, the sub-wavelength CDI problem becomes the problem of recovering a 2D signal from only the magnitude of its truncated Fourier transform. Up to spatial coordinate scaling and normalization, the above relation can be written as:

$$I(j,k) = |LFb|^2(j,k), \tag{1}$$

where I is the measured far-field intensity, F is the 2D Fourier transform operator, L is a low-pass filter with a cutoff frequency of $1/\lambda$, and b is the sought 2D object. The operator $|\cdot|$ here stands for element-wise absolute value.

Inverting Eq.1, i.e. finding b from I, L, F is an ill-posed problem, both because the high frequency information is lost due to the coupling of high spatial frequencies to evanescent waves, and due to the loss of phase information - since only the far-field (Fourier) magnitude is measured. The problem at hand, therefore, is phase-retrieval of a 2D object, combined with bandwidth-extrapolation. In order to invert this ill-posed problem, some additional information is needed, e.g. prior knowledge on the sought signal.

In this work, we focus on objects that can be represented compactly in a known basis, i.e. $b = Ax$ where A is a known basis and x is a sparse vector, namely, containing a small number of nonzero elements. In this case, Eq. 1 can be rewritten as (For simplicity, the indices (j,k) are dropped from now on):

$$I = |LFAx|^2, \tag{2}$$

and the prior knowledge of the sparsity of x adds information that helps the inversion of Eq. 2. The sparsity prior has been used for sub-wavelength imaging [13], but only when the Fourier phase was also known, yielding a linear problem. Since the measurements in our setting are not linear in the unknown (but quadratic), standard linear sparse inversion algorithms cannot be used, and a method to find a sparse solution to a set of quadratic equations is required.

III. Solution Method

The problem of sub-wavelength CDI can be viewed as consisting of two sub-problems: Phase retrieval, and bandwidth extrapolation. The problem of phase retrieval, i.e. recovering a signal from the magnitude of its Fourier transform arises in applications such as holography and crystallography, and there has been a vast amount of work dealing with it ([14], [15]). Usually, some prior knowledge about the object is used (e.g. known support or known real-space megnitude), and the different constaints are imposed iteratively. These techniques have been used in the context of CDI [2], but their application has always been limited to the information contained within numerical aperture of the system.

Here, we devise a phase-retrieval method that can also deal with the loss of high-requencies, by using the prior knowledge that the sought object is sparse in a known basis. The two problems are not handled separately, but rather solved as a combined optimization problem. The logic of the technique is as follows: An iterative thresholding method is used in order to solve Eq. 2 while using the sparsity information. The method attempts to find a solution to the following problem:

$$\begin{aligned} \min \quad & ||x||_0 \\ \text{subject to} \quad & \left|\left|I - |LFAx|^2\right|\right|_2^2 \leq \epsilon \\ & x \geq 0 \end{aligned} \tag{3}$$

The non-negativity constraint on corresponds to the assumption that the real-space object contains no phase information, which is the case we consider in this work. The thresholding method is described in detail in [16], and briefly below. First, an initial support of the vector is approximated from the blurred real-space image. Then, the following two steps are repeated iteratively:

1. Solve the minimization problem:

$$\begin{aligned} \min \quad & \left|\left|I - |LFAx|^2\right|\right|_2^2 \\ \text{subject to} \quad & x \geq 0 \end{aligned} \tag{4}$$

This is a non-convex problem, and in practice we use the L-BFGS method [17] to find a local minimum.

2. Remove the weakest element of x from the support, i.e. set it to zero. This element is constrained to remain zero in the following iterations. Go-to step 1.

The iterations continue as long as the constraint $\left|\left|I - |LFAx|^2\right|\right|_2^2 \leq \epsilon$ can be satisfied. Note that this requires knowledge of the noise level ϵ in the measurements, which might be approximated from knowledge or calibration measurements in the optical system. In addition, a stopping criterion may be defined by analyzing the reconstruction error $\left|\left|I - |LFAx|^2\right|\right|_2^2$.

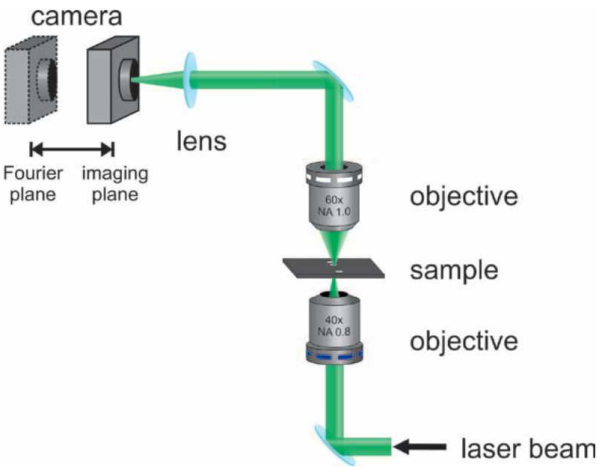


Fig. 1. Experimental setup

IV. Experimental Results

We demonstrate sparsity based sub-wavelength CDI experimentally, using the setup shown in Fig. 1. A coherently illuminated microscope (532nm LASER) is used to image arrangements of sub-wavelength holes, 100nm in diameter, in a 100nm thick Chrome layer covering a transparent substrate of fused silica. The imaging setup consists of a water-immersed objective (NA=1) and a lens imaging onto a 1002×1002 pixel CCD camera. The camera can be moved so that either the real-space (blurred) magnitude of the object is measured, or its truncated Fourier magnitude (Fig. 1). Two different patterns are imaged and recovered experimentally. The first, a star of David, is shown in Fig. 2. Figure 2a shows the Scanning-electron-microscope image of the sample. Figure 2b shows the measured real-space image using our microscope, featuring the blur caused by the diffraction limit. The measured truncated Fourier magnitude is shown in Fig. 2c. The basis for reconstruction is taken as 100nm circles on a grid, and the reconstructed image is shown in Fig. 2d. The circles are recovered with the correct locations, and their recovered amplitude is close to constant - which is consistent with the illumination used for the imaging, which had approximately constant intensity across the sample.

In order to demonstrate our ideas on a non-symmetric sample, exhibiting a truly complex Fourier transform, a second pattern, comprising of a 'random' distribution of twelve 100nm circles, is also recovered. Figure 3a shows the measured blurred real-space image, and Fig. 3b shows the measured truncated Fourier spectrum. The sparse sub-wavelength object is recovered (Fig. 3c) from its truncated Fourier spectrum, using our method, and the SEM image of the true object is shown in Fig. 3d. The reconstruction basis used here is the same as in Fig. 2, namely, 100nm circles on a grid.

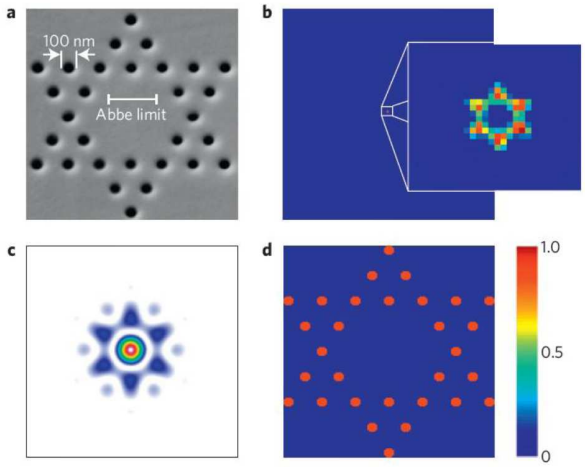


Fig. 2. a) Scanning Electron Microscope (SEM) image of the sample. b) Real-space imaging, blurred due to diffraction limit. c) Measured Fourier magnitude. d) Sparse reconstruction

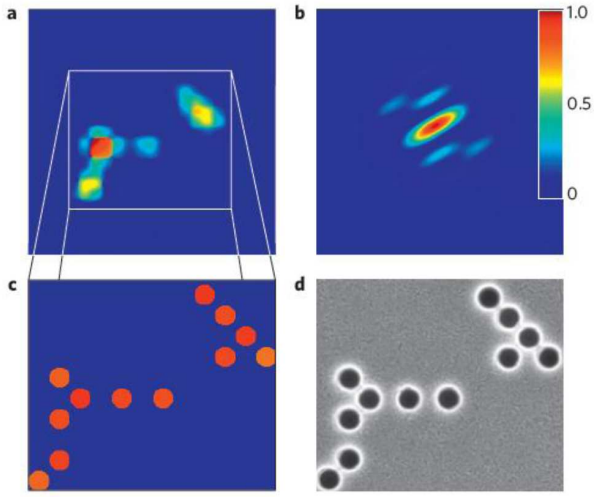


Fig. 3. a) Real-space imaging, blurred due to diffraction limit. b) Measured Fourier magnitude c) Sparse reconstruction d) Scanning Electron Microscope (SEM) image of the sample.

V. Conclusion

In this work, we have presented a technique facilitating the reconstruction of sub-wavelength features, along with phase retrieval, at an unprecedented resolution for single-shot experiments. This work opens the way for ultrafast sub-wavelength coherent diffractive imaging: ultrafast phase retrieval at the sub-wavelength scale. Fundamentally, sparsity-based concepts can be implemented in all imaging systems and achieve sub-wavelength resolution without additional hardware, given only that the image is sparse in a known basis. For example, sparsity-based methods could considerably improve the CDI resolution with x-ray free electron laser [18], without hardware modification.

Acknowledgment

This supported by the National Focal Technology Area on Nanophotonics for Detection and Sensing, and by an Advanced Grant from the European Research Council.

References

[1] D. Sayre, "Some implications of a theorem due to shannon," *Acta Crystallographica*, vol. 5, no. 6, pp. 843–843, 1952.

[2] J. Miao, P. Charalambous, J. Kirz, and D. Sayre, "Extending the methodology of x-ray crystallography to allow imaging of micrometre-sized non-crystalline specimens," *Nature*, vol. 400, no. 6742, pp. 342–344, 1999.

[3] A. Lipson, S. G. Lipson, and H. Lipson, *Optical physics*. Cambridge University Press, 2010.

[4] A. Lewis, M. Isaacson, A. Harootunian, and A. Muray, "Development of a 500 å spatial resolution light microscope: I. light is efficiently transmitted through λ/16 diameter apertures," *Ultramicroscopy*, vol. 13, no. 3, pp. 227–231, 1984.

[5] E. Betzig, J. Trautman, T. Harris, J. Weiner, R. Kostelak, *et al.*, "Breaking the diffraction barrier: optical microscopy on a nanometric scale.," *Science (New York, NY)*, vol. 251, no. 5000, p. 1468, 1991.

[6] G. T. Di Francia, "Super-gain antennas and optical resolving power," *Il Nuovo Cimento (1943-1954)*, vol. 9, pp. 426–438, 1952.

[7] H. J. Lezec, A. Degiron, E. Devaux, R. Linke, L. Martin-Moreno, F. Garcia-Vidal, and T. Ebbesen, "Beaming light from a subwavelength aperture," *Science*, vol. 297, no. 5582, pp. 820–822, 2002.

[8] F. M. Huang and N. I. Zheludev, "Super-resolution without evanescent waves," *Nano letters*, vol. 9, no. 3, pp. 1249–1254, 2009.

[9] A. Yildiz, J. N. Forkey, S. A. McKinney, T. Ha, Y. E. Goldman, and P. R. Selvin, "Myosin v walks hand-over-hand: single fluorophore imaging with 1.5-nm localization," *science*, vol. 300, no. 5628, pp. 2061–2065, 2003.

[10] S. W. Hell and J. Wichmann, "Breaking the diffraction resolution limit by stimulated emission: stimulated-emission-depletion fluorescence microscopy," *Optics letters*, vol. 19, no. 11, pp. 780–782, 1994.

[11] M. J. Rust, M. Bates, and X. Zhuang, "Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (storm)," *Nature methods*, vol. 3, no. 10, pp. 793–796, 2006.

[12] E. Betzig, G. H. Patterson, R. Sougrat, O. W. Lindwasser, S. Olenych, J. S. Bonifacino, M. W. Davidson, J. Lippincott-Schwartz, and H. F. Hess, "Imaging intracellular fluorescent proteins at nanometer resolution," *Science*, vol. 313, no. 5793, pp. 1642–1645, 2006.

[13] S. Gazit, A. Szameit, Y. C. Eldar, and M. Segev, "Super-resolution and reconstruction of sparse sub-wavelength images," *Optics Express*, vol. 17, no. 26, pp. 23920–23946, 2009.

[14] J. R. Fienup, "Phase retrieval algorithms: a comparison," *Applied optics*, vol. 21, no. 15, pp. 2758–2769, 1982.

[15] R. Gerchberg, "A practical algorithm for the determination of phase from image and diffraction plane pictures," *Optik*, vol. 35, p. 237, 1972.

[16] A. Szameit, Y. Shechtman, E. Osherovich, E. Bullkich, P. Sidorenko, H. Dana, S. Steiner, E. Kley, S. Gazit, T. Cohen-Hyams, *et al.*, "Sparsity-based single-shot subwavelength coherent diffractive imaging," *Nature materials*, vol. 11, pp. 455–459, 2012.

[17] D. C. Liu and J. Nocedal, "On the limited memory bfgs method for large scale optimization," *Mathematical programming*, vol. 45, no. 1, pp. 503–528, 1989.

[18] H. N. Chapman, A. Barty, M. J. Bogan, S. Boutet, M. Frank, S. P. Hau-Riege, S. Marchesini, B. W. Woods, S. Bajt, W. H. Benner, *et al.*, "Femtosecond diffractive imaging with a soft-x-ray free-electron laser," *Nature Physics*, vol. 2, no. 12, pp. 839–843, 2006.

Robust Polyhedral Regularization

Samuel Vaiter, Gabriel Peyré
CEREMADE, CNRS-Université Paris-Dauphine,
Place du Maréchal De Lattre De Tassigny,
75775 Paris Cedex 16, France.
Email: {vaiter,peyre}@ceremade.dauphine.fr

Jalal Fadili
GREYC, CNRS-ENSICAEN-Université de Caen,
6, Bd du Maréchal Juin,
14050 Caen Cedex, France.
Email: jalal.fadili@greyc.ensicaen.fr

***Abstract*—In this paper, we establish robustness to noise perturbations of polyhedral regularization of linear inverse problems. We provide a sufficient condition that ensures that the polyhedral face associated to the true vector is equal to that of the recovered one. This criterion also implies that the ℓ^2 recovery error is proportional to the noise level for a range of parameter. Our criterion is expressed in terms of the hyperplanes supporting the faces of the unit polyhedral ball of the regularization. This generalizes to an arbitrary polyhedral regularization results that are known to hold for sparse synthesis and analysis ℓ^1 regularization which are encompassed in this framework. As a byproduct, we obtain recovery guarantees for ℓ^∞ and $\ell^1 - \ell^\infty$ regularization.**

I. Introduction

A. Polyhedral Regularization

We consider the following linear inverse problem

$$y = \Phi x_0 + w, \tag{1}$$

where $y \in \mathbb{R}^Q$ are the observations, $x_0 \in \mathbb{R}^N$ is the unknown true vector to recover, w the bounded noise, and Φ a linear operator which maps the signal domain $\mathbb{R}^N$ into the observation domain $\mathbb{R}^Q$. The goal is to recover x_0 either exactly or to a good approximation.

We call a polyhedron a subset $\mathcal{P}$ of $\mathbb{R}^N$ such that $\mathcal{P} = \left\{x \in \mathbb{R}^N \mid Ax \leqslant b\right\}$ for some $A \in \mathbb{R}^{N_H \times N}$ and $b \in \mathbb{R}^{N_H}$, where the inequality $\leqslant$ should be understood component-wise. This is a classical description of convex polyhedral sets in terms of the hyperplanes supporting their $(N-1)$-dimensional faces.

In the following, we consider polyhedral convex functions of the form

$$J_H(x) = \max_{1 \leqslant i \leqslant N_H} \langle x, h_i \rangle,$$

where $H = (h_i)_{i=1}^{N_H} \in \mathbb{R}^{N \times N_H}$. Thus, $\mathcal{P}_H = \left\{x \in \mathbb{R}^N \mid J_H(x) \leqslant 1\right\}$ is a polyhedron. We assume that $\mathcal{P}_H$ is a bounded polyhedron which contains 0 in its interior. This amounts to saying that J_H is a gauge, or equivalently that it is continuous, non-negative, sublinear (i.e. convex and positively homogeneous), coercive, and $J_H(x) > 0$ for $x \neq 0$. Note that it is in general not a norm because it needs not be symmetric.

In order to solve the linear inverse problem (1), we devise the following regularized problem

$$x^\star \in \operatorname*{argmin}_{x \in \mathbb{R}^N} \frac{1}{2}\|y - \Phi x\|^2 + \lambda J_H(x), \tag{$P_\lambda(y)$}$$

where $\lambda > 0$ is the regularization parameter. Coercivity and convexity of J_H implies the set of minimizers is non-empty, convex and compact.

In the noiseless case, $w = 0$, one usually considers the equality-constrained optimization problem

$$x^\star \in \operatorname*{argmin}_{\Phi x = y} J_H(x). \tag{$P_0(y)$}$$

B. Relation to Sparsity and Anti-sparsity

Examples of polyhedral regularization include the ℓ^1-norm, analysis ℓ^1-norm and ℓ^∞-norm. The ℓ^1 norm reads

$$J_{H_1}(x) = \|x\|_1 = \sum_{i=1}^N |x_i|.$$

It corresponds to choosing $H_1 \in \mathbb{R}^{N \times 2^N}$ where the columns of H_1 enumerate all possible sign patterns of length N, i.e. $\{-1, 1\}^N$. The corresponding regularized problem ($P_\lambda(y)$) is the popular Lasso [1] or Basis Pursuit DeNoising [2]. It is used for recovering sparse vectors. Analysis-type sparsity-inducing penalties are obtained through the (semi-)norm $J_H(x) = \|Lx\|_1$, where $L \in \mathbb{R}^{P \times N}$ is an analysis operator. This corresponds to using $H = L^* H_1$ where * stands for the adjoint. A popular example is the anisotropic total variation where L is a first-order finite difference operator.

The ℓ^∞ norm

$$J_{H_\infty}(x) = \|x\|_\infty = \max_{1 \leqslant i \leqslant N} |x_i|$$

corresponds to choosing $H_\infty = [\mathrm{Id}_N, -\mathrm{Id}_N] \in \mathbb{R}^{N \times 2N}$. This regularization, coined anti-sparse regularization, is used for instance for approximate nearest neighbor search [3].

Another possible instance of polyhedral regularization is the group $\ell^1 - \ell^\infty$ regularization. Let $\mathcal{B}$ be a partition of $\{1, \ldots, N\}$. The $\ell^1 - \ell^\infty$ norm associated to this group structure is

$$J_{H_\mathcal{B}^\infty}(x) = \sum_{b \in \mathcal{B}} \|x_b\|_\infty.$$

This amounts to choosing the block-diagonal matrix $H_\mathcal{B}^\infty \in \mathbb{R}^{N \times \prod_{b \in \mathcal{B}} 2^{|b|}}$ such that each column is chosen by taking for each block a position with sign ± 1, others are 0. If for all $b \in \mathcal{B}, |b| = 1$, then we recover the ℓ^1-norm, whereas if the block structure is composed by one element, we get the ℓ^∞-norm.

C. Prior Work

In the special case of ℓ^1 and analysis ℓ^1 penalties, our criterion is equivalent to those defined in [4] and [5]. To our knowledge, there is no generic guarantee for robustness to noise with ℓ^∞ regularization, but [6] studies robustness of a sub-class of polyhedral norms obtained by convex relaxation of combinatorial penalties. Its notion of support is however completely different from ours. The work [7] studies numerically some polyhedral regularizations.In [8], the authors provide an homotopy-like algorithm for polyhedral regularization through a continuous problem coined adaptive inverse scale space method. The work [9] analyzes some particular polyhedral regularizations in a noiseless compressed sensing setting when the matrix Φ is drawn from an appropriate random ensemble. Again in a compressed sensing scenario, the work of [10] studies a subset of polyhedral

regularizations to get sharp estimates of the number of measurements for exact and ℓ_2-stable recovery.

II. Contributions

Definition 1. *We define the* H-support $\mathrm{supp}_H(x)$ *of a vector* $x \in \mathbb{R}^N$ *to be the set*

$$\mathrm{supp}_H(x) = \{i \in \{1, \dots, N_H\} \mid \langle x,\, h_i\rangle = J_H(x)\}\,.$$

This definition suggests that to recover signals with H-*support* $\mathrm{supp}_H(x)$, it would be reasonable to impose that Φ is invertible on the corresponding subspace $\mathrm{Ker}\, H^*_{\mathrm{supp}_H(x)}$. This is formalised in the following condition.

Definition 2. *A* H*-support* I *satisfies the* restricted injectivity condition *if*

$$\mathrm{Ker}\,\Phi \cap \mathrm{Ker}\, H_I^* = \{0\}, \tag{$\mathcal{C}_I$}$$

where H_I *is the matrix whose columns are those of* H *indexed by* I.

When it holds, we define the orthogonal projection Γ_I on $\Phi\,\mathrm{Ker}\, H_I^*$:

$$M_I = (U^*\Phi^*\Phi U)^{-1} \quad \text{and} \quad \begin{cases}\Gamma_I &= \Phi U M_I U^* \Phi^* \\ \Gamma_I^\perp &= \mathrm{Id} - \Gamma_I.\end{cases}$$

where U is (any) basis of $\mathrm{Ker}\, H_I^*$. The symmetric bilinear form on $\mathbb{R}^N$ induced by $\Gamma_I^\perp$ reads

$$\langle u,\, v\rangle_{\Gamma_I^\perp} = \langle u,\, \Gamma_I^\perp v\rangle,$$

and we denote its associated quadratic form $\|\cdot\|^2_{\Gamma_I^\perp}$.

Definition 3. *Let* I *be a* H*-support such that* ($\mathcal{C}_I$) *holds. The* Identifiability Criterion *of* I *is*

$$\mathbf{IC}_H(I) = \max_{z_I \in \mathrm{Ker}\, H_I} \min_{i\in I} (\tilde\Phi_I^* \Gamma_I^\perp \tilde\Phi_I \mathbb{I}_I + z_I)_i$$

where $\mathbb{I}_I \in \mathbb{R}^{|I|}$ *is the vector with coefficients 1, and* $\tilde\Phi_I = \Phi H_I^{+,*} \in \mathbb{R}^{Q\times|I|}$ *where* $^+$ *stands for the Moore–Penrose pseudo-inverse.*

$\mathbf{IC}_H(I)$ can be computed by solving the linear program

$$\mathbf{IC}_H(I) = \max_{(r,z_I)\in\mathbb{R}\times\mathbb{R}^{|I|}} r \text{ subj. to } \begin{cases}\forall i \in I, r \leqslant (\tilde\Phi_I^*\Gamma_I^\perp\tilde\Phi_I\mathbb{I}_I + z_I)_i \\ H_I z_I = 0.\end{cases}$$

A. Noise Robustness

Our main contribution is the following result.

Theorem 1. *Let* $x_0 \in \mathbb{R}^N \setminus \{0\}$ *and* I *its* H*-support such that* ($\mathcal{C}_I$) *holds. Let* $y = \Phi x_0 + w$. *Suppose that* $\tilde\Phi_I \mathbb{I}_I \neq 0$ *and* $\mathbf{IC}_H(I) > 0$. *Then there exists two constants* $c_I, \tilde c_I$ *satisfying,*

$$\frac{\|w\|_2}{T} < \frac{\tilde c_I}{c_I} \quad \textit{where} \quad T = \min_{j\in I^c} J_H(x_0) - \langle x_0,\, h_j\rangle > 0,$$

such that if λ *is chosen according to*

$$c_I\|w\|_2 < \lambda < T\tilde c_I,$$

the vector $x^\star \in \mathbb{R}^N$ *defined by*

$$x^\star = \mu H_I^{+,*}\mathbb{I}_I + U M_I U^* \Phi^*(y - \mu\tilde\Phi_I\mathbb{I}_I)$$

where U *is any basis of* $\mathrm{Ker}\, H_I^*$ *and*

$$0 < \mu = J_H(x_0) + \frac{\langle \tilde\Phi_I\mathbb{I}_I,\, w\rangle_{\Gamma_I^\perp} - \lambda}{\|\tilde\Phi_I\mathbb{I}_I\|^2_{\Gamma_I^\perp}} \tag{2}$$

is the unique solution of ($P_\lambda(y)$), *and* $x^\star$ *lives on the same face as* x_0, *i.e.* $\mathrm{supp}_H(x^\star) = \mathrm{supp}_H(x_0)$.

Observe that if λ is chosen proportional to the noise level, then $\|x^\star - x_0\|_2 = O(\|w\|_2)$. The following proposition proves that the condition $\mathbf{IC}_H(I) > 0$ is almost a necessary condition to ensure the stability of the H-support. Its proof is omitted for obvious space limitation reasons.

Proposition 1. *Let* $x_0 \in \mathbb{R}^N \setminus \{0\}$ *and* I *its* H*-support such that* ($\mathcal{C}_I$) *holds. Let* $y = \Phi x_0 + w$. *Suppose that* $\tilde\Phi_I\mathbb{I}_I \neq 0$ *and* $\mathbf{IC}_H(I) < 0$. *If* $\frac{\|w\|}{\lambda} < \frac{1}{c_I}$ *then for any solution of* ($P_\lambda(y)$), *we have* $\mathrm{supp}_H(x_0) \neq \mathrm{supp}_H(x^\star)$.

B. Noiseless Identifiability

When there is no noise, the following result, which is a straightforward consequence of Theorem 1, shows that the condition $\mathbf{IC}_H(I) > 0$ implies signal identifiability.

Theorem 2. *Let* $x_0 \in \mathbb{R}^N \setminus \{0\}$ *and* I *its* H*-support. Suppose that* $\tilde\Phi_I\mathbb{I}_I \neq 0$ *and* $\mathbf{IC}_H(I) > 0$. *Then the vector* x_0 *is the unique solution of* ($P_0(y)$).

III. Proofs

A. Preparatory Lemmata

We recall the definition of the subdifferential of a convex function f at the point x is the set $\partial f(x)$ is

$$\partial f(x) = \left\{g \in \mathbb{R}^N \mid f(y) \geqslant f(x) + \langle g,\, y - x\rangle\right\}.$$

The following lemma, which is a direct consequence of the properties of the max function, gives the subdifferential of the regularization function J_H.

Lemma 1. *The subdifferential* ∂J_H *at* $x \in \mathbb{R}^N$ *reads*

$$\partial J_H(x) = H_I \Sigma_I$$

where $I = \mathrm{supp}_H(x)$ *and* Σ_I *is the canonical simplex on* $\mathbb{R}^{|I|}$:

$$\Sigma_I = \left\{v_I \in \mathbb{R}^{|I|} \mid v_I \geqslant 0, \langle v_I,\, \mathbb{I}_I\rangle = 1\right\}.$$

A point $x^\star$ is a minimizer of $\min_x f(x)$ if, and only if, $0 \in \partial f(x^\star)$. Thanks to Lemma 1, this gives the first-order condition for the problem ($P_\lambda(y)$).

Lemma 2. *A vector* $x^\star$ *is a solution of* ($P_\lambda(y)$) *if, and only if, there exists* $v_I \in \Sigma_I$ *such that*

$$\Phi^*(\Phi x - y) + \lambda H_I v_I = 0,$$

where $I = \mathrm{supp}_H(x)$.

We now introduce the following so-called source condition.

($\mathbf{SC}_x$)**:** For $I = \mathrm{supp}_H(x)$, there exists η and $v_I \in \Sigma_I$ such that:

$$\Phi^*\eta = H_I v_I \in \partial J_H(x).$$

Under the source condition, a sufficient uniqueness condition can be derived when v_I lives in the relative interior of Σ_I which is

$$\mathrm{ri}\,\Sigma_I = \left\{v_I \in \mathbb{R}^{|I|} \mid v_I > 0, \langle v_I,\, \mathbb{I}_I\rangle = 1\right\}.$$

Lemma 3. *Let* $x^\star$ *be a minimizer of* ($P_\lambda(y)$) *(resp.* ($P_0(y)$)*) and* $I = \mathrm{supp}_H(x^\star)$. *Assume that* ($\mathbf{SC}_{x^\star}$) *is verified with* $v_I \in \mathrm{ri}\,\Sigma_I$, *and that* ($\mathcal{C}_I$) *holds. Then* $x^\star$ *is the unique solution of* ($P_\lambda(y)$) *(resp.* ($P_0(y)$)*).*

The proof of this lemma is omitted due to lack of space. Observe that in the noiseless case, if the assumptions of Lemma 3 hold at x_0, then the latter is exactly recovered by solving ($P_0(y)$).

Lemma 4. *Let $x^\star \in \mathbb{R}^N$ and $I = \mathrm{supp}_H(x^\star)$. Assume $(\mathcal{C}_I)$ holds. Let U be any basis of $\mathrm{Ker}\, H_I^*$. There exists $z_I \in \mathrm{Ker}\, H_I$ such that*

$$U^*\Phi^*(\Phi x^\star - y) = 0$$
$$v_I = z_I + \frac{1}{\lambda} H_I^+ \Phi^*(y - \Phi x^\star) \in \Sigma_I,$$

if, and only if, $x^\star$ is a solution of $(P_\lambda(y))$. Moreover, if $v_I \in \mathrm{ri}\,\Sigma_I$, then $x^\star$ is the unique solution of $(P_\lambda(y))$.

Proof: We compute

$$\begin{aligned}
&\Phi^*(\Phi x^\star - y) + \lambda H_I v_I \\
=&\Phi^*(\Phi x^\star - y) + \lambda H_I \left(z_I + \frac{1}{\lambda} H_I^+ \Phi^*(y - \Phi x^\star) \right) \\
=&(\mathrm{Id} - H_I H_I^+)\Phi^*(\Phi x^\star - y) = \mathrm{proj}_{H_I^*}(\Phi^*(\Phi x^\star - y)) = 0,
\end{aligned}$$

where $\mathrm{proj}_{H_I^*}$ is the projection on $\mathrm{Ker}\, H_I^*$. Hence, $x^\star$ is a solution of $(P_\lambda(y))$. If $v_I \in \mathrm{ri}\,\Sigma_I$, then according to Lemma 3, $x^\star$ is the unique solution. ∎

The following lemma is a simplified rewriting of the condition introduced in Lemma 4.

Lemma 5. *Let $x^\star \in \mathbb{R}^N$, $I = \mathrm{supp}_H(x^\star)$ and $\mu = J_H(x^\star)$. Assume $(\mathcal{C}_I)$ holds. Let U be any basis of $\mathrm{Ker}\, H_I^*$. There exists $z \in \mathrm{Ker}\, H_I$ such that*

$$v_I = z_I + \frac{1}{\lambda} \tilde{\Phi}_I^* \Gamma_I^\perp (y - \mu \tilde{\Phi}_I \mathbb{I}_I) \in \Sigma_I,$$

if, and only if, $x^\star$ is a solution of $(P_\lambda(y))$. Moreover, if $v_I \in \mathrm{ri}\,\Sigma_I$, then $x^\star$ is the unique solution of $(P_\lambda(y))$.

Proof: Note that any vector $x \in \mathbb{R}^N$ such that the condition $(\mathcal{C}_I)$ holds, where I is the H-support of x, is such that

$$x = \mu H_I^{+,*} \mathbb{I}_I + U\alpha \quad \text{where} \quad \mu = J_H(x),$$

for some coefficients α and U any basis of $\mathrm{Ker}\, H_I^*$. We obtain

$$U\Phi^*(\Phi x^\star - y) = \mu U \Phi^* \Phi H_I^{+,*} \mathbb{I}_I - U\Phi^* y + U\Phi^*\Phi U\alpha = 0$$

Since $(\mathcal{C}_I)$ holds, we have

$$\alpha = (U\Phi^*\Phi U\alpha)^{-1} U\Phi^* \left(y - \mu \tilde{\Phi}_I \mathbb{I}_I \right).$$

Hence,

$$\Phi U\alpha = \Gamma_I \left(y - \mu \tilde{\Phi}_I \mathbb{I}_I \right).$$

Now since, $x^\star = \mu H_I^{+,*} \mathbb{I}_I + U\alpha$, one has

$$\Phi x^\star = \mu \tilde{\Phi}_I \mathbb{I}_I + \Gamma_I \left(y - \mu \tilde{\Phi}_I \mathbb{I}_I \right) = \mu \Gamma_I^\perp \tilde{\Phi}_I \mathbb{I}_I + \Gamma_I y.$$

Subtracting y and multiplying by $\tilde{\Phi}_I^*$ both sides, and replacing in the expression of v_I in Lemma 4, we get the desired result. ∎

B. Proof of Theorem 1

Let I be the H-support of x_0. We consider the restriction of $(P_\lambda(y))$ to the H-support I.

$$x^\star = \underset{\substack{x \in \mathbb{R}^N \\ \mathrm{supp}_H(x) \subseteq I}}{\mathrm{argmax}} \ \frac{1}{2}\|y - \Phi x\|_2^2 + J_H(x). \qquad (\mathcal{P}_\lambda(y)_I)$$

Thanks to $(\mathcal{C}_I)$, the objective function is strongly convex on the set of signals of H-support I. Hence $x^\star$ is uniquely defined. The proof is divided in five parts: We give (**1.**) an implicit form of $x^\star$. We check (**2.**) that the H-support of $x^\star$ is the same as the H-support of x_0. We provide (**3.**) the value of $J_H(x^\star)$. Using Lemma 5, we prove (**4.**) that $x^\star$ is the unique minimizer of $(P_\lambda(y))$.

1. Expression of $x^\star$. One has $x^\star = \mu H_I^{+,*} \mathbb{I}_I + U\alpha$ where $\mu = J_H(x^\star)$. Hence,

$$U^*\Phi^*(\Phi x - y) = \mu U^*\Phi^*\Phi H_I^{+,*}\mathbb{I}_I + (U^*\Phi^*\Phi U)\alpha - U^*\Phi^* y = 0.$$

Thus,

$$U\alpha = U M_I U^* \Phi^* (y - \mu \Phi H_I^{+,*} \mathbb{I}_I).$$

Now, since $y = \Phi x_0 + w$, with $\mathrm{supp}_H(x_0) = I$, then

$$\begin{aligned}
x^\star &= \mu H_I^{+,*}\mathbb{I}_I + U M_I U^* \Phi^* (y - \mu \Phi H_I^{+,*}\mathbb{I}_I) \\
&= \mu H_I^{+,*}\mathbb{I}_I + U M_I U^* \Phi^* ((\mu_0 - \mu)\Phi H_I^{+,*}\mathbb{I}_I + w) + U\alpha_0 \\
&= x_0 - (\mu_0 - \mu) H_I^{+,*}\mathbb{I}_I + U M_I U^* \Phi^* ((\mu_0 - \mu)\Phi H_I^{+,*}\mathbb{I}_I + w),
\end{aligned}$$

where $\mu_0 = J_H(x_0)$. Hence, $x^\star$ is satisfying

$$x^\star = x_0 + (\mu_0 - \mu)[U M_I U^* \Phi^* \Phi - \mathrm{Id}] H_I^{+,*}\mathbb{I}_I + U M_I U^* \Phi^* w. \quad (3)$$

2. Checking that the H-support of $x^\star$ is I. To ensure that the H-support of $x^\star$ is I we have to impose that

$$\begin{aligned}
\forall i \in I, \quad & \langle h_i, x^\star \rangle = J_H(x^\star) = \mu \\
\forall j \in I^c, \quad & \langle h_j, x^\star \rangle < J_H(x^\star) = \mu.
\end{aligned}$$

The components on I of $x^\star$ are satisfying $H_I^* x^\star = \mu \mathbb{I}_I$. Since J_H is subadditive, we bound the components on I^c by the triangular inequality on (3) to get

$$\begin{aligned}
\max_{j \in I^c} \langle h_j, x^\star \rangle \leqslant & \max_{j \in I^c} \langle h_j, x_0 \rangle \\
& + (\mu_0 - \mu) \| H_{I^c}^* [U M_I U^* \Phi^* \Phi - \mathrm{Id}] H_I^{+,*} \mathbb{I}_I \|_\infty \\
& + \| H_{I^c}^* U M_I U^* \Phi^* w \|_\infty.
\end{aligned}$$

Denoting

$$\begin{aligned}
C_1 &= \| H_{I^c}^* [U M_I U^* \Phi^* \Phi - \mathrm{Id}] H_I^{+,*} \mathbb{I}_I \|_\infty, \\
C_2 &= \| H_{I^c}^* U M_I U^* \Phi^* \|_{2,\infty}, \\
T &= \mu_0 - \max_{j \in I^c} \langle h_j, x_0 \rangle,
\end{aligned}$$

we bound the correlations outside the H-support by

$$\max_{j \in I^c} \langle h_j, x^\star \rangle \leqslant \mu_0 - T + (\mu_0 - \mu) C_1 + C_2 \|w\|.$$

There exists some constants c_1, c_2 satisfying $c_1\|w\| < c_2 T + \lambda$ such that

$$0 \leqslant \mu_0 - T + (\mu_0 - \mu) C_1 + C_2 \|w\| < \mu \quad (4)$$

Under this condition, one has

$$\max_{j \in I^c} \langle h_j, x^\star \rangle < \mu,$$

which proves that $\mathrm{supp}_H(x^\star) = I$.

3. Value of $\mu = J_H(x^\star)$. Using Lemma 5 with $H = U^*H$, since $x^\star$ is a solution of $(\mathcal{P}_\lambda(y)_I)$, there exists $z_I \in \mathrm{Ker}\, H_I$ such that

$$v_I = z_I + \frac{1}{\lambda} \tilde{\Phi}_I^* \Gamma_I^\perp (y - \mu \tilde{\Phi}_I \mathbb{I}_I) \in \Sigma_I. \quad (5)$$

We decompose x_0 as

$$x_0 = \mu_0 H_I^{+,*} \mathbb{I}_I + U\alpha_0.$$

Since $y = \Phi x_0 + w$, we have

$$\Gamma_I^\perp y = \Gamma_I^\perp (\mu_0 \tilde{\Phi}_I \mathbb{I}_I + \Phi U \alpha_0 + w).$$

Now since

$$\Gamma_I \Phi U \alpha_0 = \Phi U (U^*\Phi^*\Phi U)^{-1} U^*\Phi^*\Phi U\alpha_0 = \Phi U \alpha_0,$$

one obtains

$$\Gamma_I^\perp y = \mu_0 \Gamma_I^\perp \tilde{\Phi}_I \mathbb{I}_I + \Gamma_I^\perp w.$$

Thus, equation (5) equivalently reads

$$v_I = z_I + \frac{1}{\lambda} \tilde{\Phi}_I^* \Gamma_I^\perp \left((\mu_0 - \mu) \tilde{\Phi}_I \mathbb{I}_I + w \right).$$

In particular, $\langle v_I, \mathbb{I}_I \rangle = \lambda$. Thus,

$$\lambda = \langle \lambda v_I, \mathbb{I}_I \rangle = \langle \lambda \tilde{z}_I, \mathbb{I}_I \rangle + \langle \tilde{\Phi}_I^* \Gamma_I^\perp ((\mu_0 - \mu) \tilde{\Phi}_I \mathbb{I}_I + w, \mathbb{I}_I \rangle.$$

Since $\tilde{z}_I \in \operatorname{Ker} H_I$, one has $\langle z_I, \mathbb{I}_I \rangle = 0$.

$$\begin{aligned} \lambda &= \langle \tilde{\Phi}_I^* \Gamma_I^\perp ((\mu_0 - \mu) \tilde{\Phi}_I \mathbb{I}_I + w, \mathbb{I}_I \rangle \\ &= (\mu_0 - \mu) \|\tilde{\Phi}_I \mathbb{I}_I\|_{\Gamma_I^\perp}^2 + \langle \tilde{\Phi}_I \mathbb{I}_I, w \rangle_{\Gamma_I^\perp}. \end{aligned}$$

Thus the value of μ is given by

$$\mu = \mu_0 + \frac{\langle \tilde{\Phi}_I \mathbb{I}_I, w \rangle_{\Gamma_I^\perp} - \lambda}{\|\tilde{\Phi}_I \mathbb{I}_I\|_{\Gamma_I^\perp}^2} > 0. \tag{6}$$

4. Checking conditions of Lemma 5. Consider now the vector $\tilde{v}_I$ defined by

$$\tilde{v}_I = \tilde{z}_I + \frac{1}{\lambda} \tilde{\Phi}_I^* \Gamma_I^\perp \left((\mu_0 - \mu) \tilde{\Phi}_I \mathbb{I}_I + w \right),$$

where

$$\tilde{z}_I = \frac{1}{\mu - \mu_0} \left(\operatorname*{argmax}_{z_I \in \operatorname{Ker} H_I} \min_{i \in I} (\tilde{\Phi}_I^* \Gamma_I^\perp \tilde{\Phi}_I \mathbb{I}_I + z_I)_i \right)$$

Under condition (4), the H-support of $x^\star$ is I, hence we only have to check that $\tilde{v}_I$ is an element of $\operatorname{ri} \Sigma_I$. Since $\langle \tilde{z}_I, \mathbb{I}_I \rangle = 0$, one has

$$\begin{aligned} &\langle \tilde{v}_I, \mathbb{I}_I \rangle \\ =&\langle z_I + \frac{1}{\lambda} \tilde{\Phi}_I^* \Gamma_I^\perp \left((\mu_0 - \mu) \tilde{\Phi}_I \mathbb{I}_I + w \right), \mathbb{I}_I \rangle + \langle \tilde{z}_I - z_I, \mathbb{I}_I \rangle \\ =&\langle v_I, \mathbb{I}_I \rangle + 0 = \lambda. \end{aligned}$$

Plugging back the expression (6) of $(\mu_0 - \mu)$ in the definition of $\tilde{v}_I$, one has

$$\tilde{v}_I = \tilde{z}_I + \frac{1}{\lambda} \left(\tilde{\Phi}_I^* \Gamma_I^\perp w + \frac{\langle \tilde{\Phi}_I \mathbb{I}_I, w \rangle_{\Gamma_I^\perp} - \lambda}{\|\tilde{\Phi}_I \mathbb{I}_I\|_{\Gamma_I^\perp}^2} \tilde{\Phi}_I^* \Gamma_I^\perp \tilde{\Phi}_I \mathbb{I}_I \right).$$

For some constant c_3 such that $c_3 \|w\| - \mathbf{IC}_H(I) \cdot \lambda > 0$, one has

$$\forall i \in I, \quad v_i > 0.$$

Combining this with the fact that $\langle \tilde{v}_I, \mathbb{I}_I \rangle = \lambda$ proves that $\tilde{v}_I \in \operatorname{ri} \Sigma_I$. According to Lemma 5, $x^\star$ is the unique minimizer of $(P_\lambda(y))$. ∎

C. Proof of Theorem 2

Taking $w = 0$ in Theorem 1, we obtain immediately

Lemma 6. *Let $x_0 \in \mathbb{R}^N \setminus \{0\}$ and I its H-support such that $(\mathcal{C}_I)$ holds. Let $y = \Phi x_0$. Suppose that $\tilde{\Phi}_I \mathbb{I}_I \neq 0$ and $\mathbf{IC}_H(I) > 0$. Let $T = \min_{j \in I^c} J_H(x_0) - \langle x_0, h_j \rangle > 0$ and $\lambda < T \tilde{c}_I$. Then,*

$$x^\star = x_0 + \frac{\lambda}{\|\tilde{\Phi}_I \mathbb{I}_I\|_{\Gamma_I^\perp}^2} [U M_I U^* \Phi^* \Phi - \mathrm{Id}] H_I^{+,*} \mathbb{I}_I,$$

is the unique solution of $(P_\lambda(y))$.

The following lemma shows that under the same condition, x_0 is a solution of $(P_0(y))$.

Lemma 7. *Let $x_0 \in \mathbb{R}^N \setminus \{0\}$ and I its H-support such that $(\mathcal{C}_I)$ holds. Let $y = \Phi x_0$. Suppose that $\tilde{\Phi}_I \mathbb{I}_I \neq 0$ and $\mathbf{IC}_H(I) > 0$. Then x_0 is a solution of $(P_0(y))$.*

Proof: According to Lemma 6, for every $0 < \lambda < T \tilde{c}_I$,

$$x_\lambda^\star = x_0 + \frac{\lambda}{\|\tilde{\Phi}_I \mathbb{I}_I\|_{\Gamma_I^\perp}^2} [U M_I U^* \Phi^* \Phi - \mathrm{Id}] H_I^{+,*} \mathbb{I}_I,$$

is the unique solution of $(P_\lambda(y))$.

Let $\tilde{x} \neq x_0$ such that $\Phi \tilde{x} = y$. For every $0 < \lambda < T \tilde{c}_I$, since $x_\lambda^\star$ is the unique minimizer of $(P_\lambda(y))$, one has

$$\frac{1}{2} \|y - \Phi x_\lambda^\star\|_2^2 + J_H(x_\lambda^\star) < \frac{1}{2} \|y - \Phi \tilde{x}\|_2^2 + J_H(\tilde{x}).$$

Using the fact that $\Phi \tilde{x} = y = \Phi x_0$, one has $J_H(x_\lambda^\star) < J_H(\tilde{x})$. By continuity of the mapping $x \mapsto J_H(x)$, taking the limit for $\lambda \to 0$ in the previous inequality gives

$$J_H(x_0) \leqslant J_H(\tilde{x}).$$

It follows that x_0 is a solution of $(P_0(y))$. ∎

We now prove Theorem 2.

Proof of Theorem 2: Lemma 7 proves that x_0 is a solution of $(P_0(y))$. We now prove that x_0 is in fact the unique solution. Let $\tilde{z}_I$ be the argument of the maximum in the definition of $\mathbf{IC}_H(I)$. We define

$$\tilde{v}_I = \frac{1}{\|\tilde{\Phi}_I \mathbb{I}_I\|_{\Gamma_I^\perp}^2} \left(\tilde{z}_I + \tilde{\Phi}_I^* \Gamma_I^\perp \tilde{\Phi}_I \mathbb{I}_I \right).$$

By definition of $\mathbf{IC}_H(I)$, for every $i \in I, \tilde{v}_I > 0$ and $\langle \tilde{v}_I, \mathbb{I}_I \rangle = 1$. Thus, $H_I \tilde{v}_I \in \operatorname{ri}(\partial J_H(x_0))$. Moreover, since $\tilde{z}_I \in \operatorname{Ker} H_I$, one has

$$H_I v_I = H_I H_I^{+,*} \Phi^* \Gamma_I^\perp \tilde{\Phi}_I \mathbb{I}_I = \Phi^* \eta \quad \text{where} \quad \eta = \Gamma_I^\perp \tilde{\Phi}_I \mathbb{I}_I.$$

Thanks to Lemma 3, x_0 is the unique solution of $(P_0(y))$. ∎

References

[1] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[2] S. Chen, D. Donoho, and M. Saunders, "Atomic decomposition by basis pursuit," *SIAM journal on scientific computing*, vol. 20, no. 1, pp. 33–61, 1998.

[3] H. Jégou, T. Furon, and J. Fuchs, "Anti-sparse coding for approximate nearest neighbor search," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 2029–2032.

[4] J. Fuchs, "On sparse representations in arbitrary redundant bases," *Information Theory, IEEE Transactions on*, vol. 50, no. 6, pp. 1341–1344, 2004.

[5] S. Vaiter, G. Peyré, C. Dossal, and J. Fadili, "Robust sparse analysis regularization," *to appear in IEEE Transactions on Information Theory*, 2012.

[6] F. Bach, "Structured sparsity-inducing norms through submodular functions," *Advances in Neural Information Processing Systems*, 2010.

[7] S. Petry and G. Tutz, "Shrinkage and variable selection by polytopes," *Journal of Statistical Planning and Inference*, vol. 142, no. 1, pp. 48–64, 2012.

[8] M. Moeller and M. Burger, "Multiscale methods for polyhedral regularizations," UCLA, CAM Report 11-74, 2011.

[9] D. Donoho and J. Tanner, "Counting the faces of randomly-projected hypercubes and orthants, with applications," *Discrete & computational geometry*, vol. 43, no. 3, pp. 522–541, 2010.

[10] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky, "The Convex Geometry of Linear Inverse Problems," *Foundations of Computational Mathematics*, vol. 12, no. 6, pp. 805–849, 2012.

On the Performance of Adaptive Sensing for Sparse Signal Inference

Rui M. Castro
Eindhoven University of Technology
The Netherlands
Email: rmcastro@tue.nl

***Abstract*—In this short paper we survey recent results characterizing the fundamental draws and limitations of adaptive sensing for sparse signal inference. We consider two different adaptive sensing paradigms, based either on single-entry or linear measurements. Signal magnitude requirements for reliable inference are shown for two different inference goals, namely signal detection and signal support estimation.**

I. Introduction

In this short paper we survey recent results characterizing the fundamental draws and limitations of adaptive sensing. One of the key aspects of adaptive sensing is that the data collection process is sequential and adaptive. In different fields these sensing/experimenting paradigms are known by different names, such as *sequential experimental design* in statistics and economics (see [1], [2], [3], [4], [5]), *active learning* or *adaptive sensing/sampling* in computer science, engineering and machine learning (see [6], [7], [8], [9], [10], [11], [12], [13], [14]). An essential aspect of adaptive sensing is the intricate coupling between data analysis and acquisition, which creates a powerful feedback structure. This is a double-edged sword: it is key to harness the power of sequential experimental design but also raises challenges in the analysis of such methodologies – indeed it creates complicated and strong dependencies in the data sequence.

We consider a model where the signal of interest is represented by a sparse vector $\boldsymbol{x} \in \mathbb{R}^n$, meaning that most entries of $\boldsymbol{x}$ are zero and only few of the entries are non-zero. Specifically let S be a subset of $\{1,\dots,n\}$ of non-zero entries of $\boldsymbol{x}$, and assume that for all $i \in \{1,\dots,n\}$ such that $i \notin S$ we have $x_i = 0$. We refer to S as the signal support set and this is our main object of interest. We consider two distinct classes of problems: (i) *signal detection*, where we want to test if S belongs to a particular class of subsets of $\{1,\dots,n\}$, and (ii) *support estimation*, where we desire to actually estimate S. The signal $\boldsymbol{x}$ in naturally assumed to be unknown, but we can collect partial information about it through noisy measurements. In particular we consider generalizations of the normal means model allowing for multiple and sequential measurements, therefore enabling adaptive sensing strategies. Our focus is primarily on single-entry observations, but in Section III we discuss also a different (and statistically more powerful) sensing model which allows for linear measurements of the signal - in what is often referred to as Compressive Sensing (CS).

II. Single-entry Measurements

This sensing model was first proposed in [15]. Measurements are of the form

$$Y_k = x_{A_k} + \Gamma_k^{-1/2} W_k \ , \qquad k = 1, 2, \dots \ ,$$

where A_k, Γ_k are taken to be functions of $\{Y_i, A_i, \Gamma_i\}_{i=1}^{k-1}$, and W_k are standard normal random variables, independent of $\{Y_i\}_{i=1}^{k-1}$ and also independent of $\{A_i, \Gamma_i\}_{i=1}^{k}$. In words, each measurement corresponds to a single signal entry corrupted with additive Gaussian noise, and the choice of entry and noise level can be controlled. However, there is a total sensing budget constraint that must be satisfied, namely

$$\sum_{k=1}^{\infty} \Gamma_k \leq m \ , \tag{1}$$

where $m > 0$. In the above model A_k should be viewed as the *sensing action* taken at time k, and Γ_k is the *precision* of the corresponding measurement. We have control over both quantities. Informally stated, measurements are collected sequentially, and for each measurement we can choose which entry of $\boldsymbol{x}$ to observe, and what is the precision (i.e. accuracy) of the measurement. We are allowed to collect as many measurements as desired provided the cumulative precision used satisfies the budget (1). Note that in this model we are allowed to collect an infinite (but countable) number of measurements, provided the precision Γ_k converges to zero as k grows. Although this might seem strange at first, it is not entirely unreasonable in practice - in many sensing modalities the precision is directly proportional to the amount of time necessary to collect a measurement, and therefore (1) can be viewed simply as a time constraint. This is the case in various imaging modalities (e.g. in astronomy) where long exposure times are used to reduce the noise level, which is inversely proportional to the exposure time. It is important to note that there are also settings where the actual number of measurements is limited, and there is little control on the precision level. In that case (1) might represent a constrain on the total number of measurements, provided Γ_k is not a function of k. The results in the latter setting are similar to the ones presented in the current paper, especially when studying asymptotics (when both n and m grow).

It is important to note that we can consider both deterministic sequential designs or random sequential designs. In

the latter we allow the choices A_k and Γ_k to incorporate extraneous randomness, which is not explicitly described in the model. The collection of conditional distributions of A_k, Γ_k given $\{Y_i, A_i, \Gamma_i\}_{i=1}^{k-1}$ for all k is referred to as the *sensing strategy*. Note that, within the sensing paradigm above we can also consider non-adaptive sensing, meaning that the choice of sensing actions and corresponding precision is made before collecting any data. Formally this means that $\{A_k, \Gamma_k\}_{k\in\mathbb{N}}$ is statistically independent from $\{Y_k\}_{k\in\mathbb{N}}$. Note that a non-adaptive design can still be random.

The case $m = n$ is of particular interest, allowing a direct comparison between adaptive and non-adaptive sensing methodologies. When $m = n$ we allow on average one unit of precision per each of the signal entries. So, if there is no reason to give preference to any particular entry of $\boldsymbol{x}$, the natural optimal non-adaptive sensing strategy should simply measure each entry of $\boldsymbol{x}$ exactly once, with precision one. This corresponds to the well studied normal means model.

For simplicity of presentation we consider only signals of the form

$$x_i = \begin{cases} \mu & \text{if } i \in S \\ 0 & \text{if } i \notin S \end{cases} ,$$

where $\mu > 0$ is called the *signal amplitude*. This restriction is also considered in [16], [17] in the non-adaptive sensing context and does not substantially hinder the generality of the results presented in this manuscript.

As stated before we consider two different inference problems: (i) signal detection and (ii) support estimation. For the detection problem (i) the goal is to determine if a signal is present or absent. We formulate the problem as a binary hypothesis testing, and test a simple null hypothesis against a composite alternative. In particular the null hypothesis H_0 is simply $S = \emptyset$, and the alternative hypothesis H_1 is $S \in \mathcal{C}$, where $\mathcal{C}$ is some class of non-empty subsets of $\{1, \ldots, n\}$. For simplicity of presentation we assume that all the sets in $\mathcal{C}$ have the same cardinality s. A test procedure based on the (adaptive) measurements is described by a binary test function $\hat{\phi}\left(\{A_i, \Gamma_i, Y_i\}_{i=1}^{\infty}\right) \in \{0, 1\}$, and a natural way to measure the performance of such a test function is the *worst case* risk

$$R(\hat{\phi}) = \mathbb{P}_{\emptyset}(\hat{\phi} \neq 0) + \max_{S\in\mathcal{C}} \mathbb{P}_S(\hat{\phi} \neq 1) ,$$

where $\mathbb{P}_S$ denotes the joint probability distribution of $\{A_i, \Gamma_i, Y_i\}_{i=1}^{\infty}$ for a given support set S. Characterizing the relation between $R(\hat{\phi})$, n, m, μ, and $\mathcal{C}$ is our main objective.

The goal of the estimation problem (ii) is (statistically) more ambitious, as we seek to actually identify the support set S. An estimation procedure is a function $\hat{S}$ mapping $\{A_i, \Gamma_i, Y_i\}_{i=1}^{\infty}$ to a subset of $\{1, \ldots, n\}$. There are several sensible ways to measure "closeness" between $\hat{S}$ and the true support set S, for instance the worst case probability of making any errors

$$\max_{S\in\mathcal{C}} \mathbb{P}_S[\hat{S} \neq S] .$$

A somewhat more stringent metric is the worst case expected number of errors $\max_{S\in\mathcal{C}} \mathbb{E}_S[|\hat{S}\Delta S|]$, and clearly $\mathbb{P}_S[\hat{S} \neq S] \leq \mathbb{E}_S[|\hat{S}\Delta S|]$. We will focus mainly on the first metric in this manuscript, but remark that the two metrics are essentially equivalent in several cases.

A. Single-entry Measurements: Results

In this section we present the fundamental tradeoffs for the inference problems presented above. Clearly these results bear some dependency on the class of sets $\mathcal{C}$:

Definition II.1 (symmetric class)**.** *Let S be a random set, drawn uniformly at random from $\mathcal{C}$. If for all $i \in \{1, \ldots, n\}$ we have $\mathbb{P}(i \in S) = s/n$ the class $\mathcal{C}$ is said to be symmetric.*

In words, in a symmetric class of sets there is no reason to give a priori preference to any individual entry. Many classes $\mathcal{C}$ of interest satisfy this mild symmetry, for instance all the classes in [16]. Of particular interest is the maximal class of all the subsets of $\{1, \ldots, n\}$ with cardinality s, which corresponds to *lack of structure* in the sparsity pattern S. If the class $\mathcal{C}$ is smaller then we say the sparsity patterns S have *structure*. An example of a structured class is presented later.

Theorem II.1 ([18])**.** *Let $\mathcal{C}$ be a symmetric class, and let $\hat{\Phi}$ be an arbitrary adaptive sensing testing procedure. For any $0 < \epsilon < 1$, if $R(\hat{\Phi}) \leq \epsilon$ then necessarily*

$$\mu \geq \sqrt{\frac{2n}{sm} \log \frac{1}{2\epsilon}} .$$

As argued before, the case $m = n$ is of particular interest, as it allows for comparison between adaptive and non-adaptive sensing performance: in that case the above bound is of the order $\sqrt{2/s}$. It is remarkable that the extrinsic signal dimension n plays no role in this bound, and only the intrinsic signal dimension s is relevant. This is in stark contrast to what is known for the same problem if one restricts to the classical setting of non-adaptive sensing, as in [19], [20], [17]. For instance, for the class of all subsets with cardinality s the non-adaptive sensing lower bound is of the order $\sqrt{\log(n/s^2)}$ if $s < o(\sqrt{n})$. Therefore signals need to be much stronger in order to be reliably detected when using non-adaptive sensing.

The above adaptive sensing lower bound is valid for any symmetric class, and in particular for the maximal class of all subsets S with cardinality s. For this class there is a adaptive sensing methodology able to nearly achieve the lower bound.

Proposition II.1 ([18])**.** *Let $s_n > \log\log\log n$ and consider the class $\mathcal{C}$ of all subsets with cardinality s_n. Furthermore let $\mu > \sqrt{\frac{32 \log\log\log n}{s_n}}$. There is an adaptive sensing testing strategy for which*

$$R(\hat{\Phi}) \to 0 ,$$

as $n \to \infty$.

The mentioned procedure is based on the idea of distilled sensing [15], but it does require some simple modifications to attain the desired bound (see [18]). Note that the order of the bound matches the one of the lower bound up to a factor $\log\log\log n$. It is conjectured that this is an artifact of the specific procedure, however, there are currently no known procedures able to tighten this gap. Perhaps more noteworthy

is the fact that extra structure in the class $\mathcal{C}$ is not helpful in the adaptive sensing detection scenario! This is quite different than in the non-adaptive sensing case, where the structure of the set $\mathcal{C}$ can play a very prominent role as well documented in [16], [21], [22], for instance.

The estimation problem exhibit similar trends, but structure of the set $\mathcal{C}$ can give important cues on the design of adaptive sensing methodologies. We focus first on the unstructured case where $\mathcal{C}$ is the class of all subsets of $\{1,\dots,n\}$ with cardinality s.

Theorem II.2 ([18])**.** *Let $\mathcal{C}$ be the class of all subsets with cardinality s, and let $\hat{S}$ be an arbitrary adaptive sensing support estimator. For any $0<\epsilon<1$, if $\max_{S\in\mathcal{C}}\mathbb{P}_S[\hat{S}\neq S]\leq \max_{S\in\mathcal{C}}\mathbb{E}_S[|\hat{S}\Delta S|]\leq\epsilon$ then necessarily*

$$\mu^2 \geq \sqrt{\frac{2n}{m}\left(\log s+\log\frac{n-s}{n+1}+\log\frac{1}{2\epsilon}\right)}\ .$$

Again, focusing on the case $m=n$ and assuming also the signal is sufficiently sparse (meaning $s_n=o(n)$), we see that μ needs to be on the order of $\sqrt{2\frac{n}{m}\log(s_n)}$ to ensure the probability of making any errors goes to zero as n increases. This result is again in stark contrast with what is possible with non-adaptive sensing, where the signal magnitude μ needs to be on the order of $\sqrt{2\log n}$ to ensure the probability of error goes to zero. Furthermore the above lower bound is tight, as there is a procedure that allows for exact support recovery with probability approaching 1 provided the signal amplitude is of the order $2\sqrt{\log s_n+\log\log n}$ (see [23], [24]). The $\log\log n$ term and the "wrong" constant in the bound are artifacts of their method (which is parameter adaptive and agnostic about s_n), and can be avoided when considering a different approach - running in parallel n entry-wise properly calibrated sequential likelihood ratio tests, which require the knowledge of the sparsity level s_n. Such a procedure achieves precisely the lower bound in the theorem.

It is interesting to notice that, unlike for detection, structure in the class $\mathcal{C}$ can be extremely helpful for estimation. This is the case both for adaptive and non-adaptive sensing. Perhaps the simplest type of structure to consider is when the set S is an "interval", meaning all the entries of S are contiguous (e.g. $S=\{i,i+1,i+s_n-1\}$ for some i). Then adaptive sensing can successfully recover the support with probability approaching 1 provided the signal magnitude is of the order $\sqrt{2\log(s_n)/s_n}$, and this is the optimal rate (unpublished work). Adaptive sensing under other structural constrains (e.g., cliques in a complete graph, paths in a graph) have to the best of our knowledge not been thoroughly studied yet, and therefore remain an important direction for future work.

III. Linear Measurements and Compressed Sensing

The sensing model described in the previous section can be modified to allow for linear measurements, in lieu of single-entry samples. Formally the sensing model becomes

$$\boldsymbol{Y}=\boldsymbol{A}\boldsymbol{x}+\boldsymbol{W}\ ,$$

where $\boldsymbol{Y}\in\mathbb{R}^l$ denotes the observations, $\boldsymbol{A}\in\mathbb{R}^{l\times n}$ is the design/sensing matrix, $\boldsymbol{x}\in\mathbb{R}^n$ is the unknown signal, and $\boldsymbol{W}\in\mathbb{R}^l$ is a normal multivariate vector with zero mean an identity covariance matrix. The rows of $\boldsymbol{A}$ can be designed sequentially, and the i^{th} row (denoted by $\boldsymbol{A}_{i\cdot}$) can depend explicitly on $\{Y_j,\boldsymbol{A}_{j\cdot}\}_{j=1}^{i-1}$. Note that W_i is a normal random variable independent of $\{Y_j,\boldsymbol{A}_{j\cdot},W_j\}_{j=1}^{i-1}$ and also independent of $\boldsymbol{A}_{i\cdot}$. This setting is particularly interesting when we impose norm constrains on $\boldsymbol{A}$, namely

$$\mathbb{E}\left[\|\boldsymbol{A}\|_F^2\right]\leq m\ , \tag{2}$$

where $\|\cdot\|_F$ is the Frobenius matrix norm. Like (1), this sensing budget condition is very natural and the issue of noise is otherwise irrelevant. The norm of each row of $\boldsymbol{A}$ plays here the role of the precision parameters Γ_k in (1).

Inference based on linear measurements is at the heart of compressed sensing. Most existing literature focused on the non-adaptive sensing paradigm, and identified strategies to recover signals from a small number of measurements, see for instance [25], [26], [27]. In our setting this means l is chosen to be as small as possible, while making the restriction $l=m$. In the results described below we consider only the sensing budget restriction (2) and assume the number of measurements l can be potentially infinite.

As linear measurements are more powerful/general than entry-wise ones, we might expect some performance improvement in both the detection and estimation inference tasks. The detection problem was been carefully studied in [28] and the author has shown that for reliable detection it is necessary and sufficient for the signal magnitude to be of the order $\frac{1}{s_n}\sqrt{n/m}$. Although this result is somewhat similar to the one in Theorem II.1 we notice that the dependency on the sparsity level s_n is better, and therefore weaker signals can be detected using linear measurements. Perhaps surprisingly adaptive sensing is of no help in this scenario, and detection procedures achieving the optimal performance can be non-adaptive. Furthermore, the structure of the class $\mathcal{C}$ does not help, provided the class is symmetric. This means that, like in the single-entry measurement case, structure is of no use for detection. However, this statement is true both for adaptive and non-adaptive sensing paradigms, meaning that the extra flexibility of adaptive sensing provides no advantage for detection using linear measurements.

For the estimation problem the story is a bit different: adaptive sensing can exhibit an advantage over non-adaptive sensing, as documented in [29], [30], [31]. Furthermore structural information about S can be extremely helpful. In [18] it is shown that for the unstructured case the same lower bound as in Theorem II.2 applies in the context of linear measurements (although the proof of the result requires a few small modifications). Procedures achieving (or nearly achieving) this bound exist, namely [31], [32]. For the non-adaptive sensing paradigm information theoretical lower bounds have also been shown, namely the signal amplitude must exceed a constant times $\sqrt{\frac{n}{m}\sigma^2\log n}$, as shown for instance in [33].

The factor of n/m is the sensing energy per dimension and $\sqrt{\log n}$ is needed to ensure that the signal is larger than the largest noise contribution. Therefore adaptive sensing is advantageous, especially in the typical case when the signal dimension n is very large.

If the sparsity patterns exhibit some structure there are also results contrasting adaptive and non-adaptive sensing, but the story is far from complete. In [34] the authors devise an algorithm that can identify the support set S with high probability when S is an "interval" (see the last paragraph of Section II) provided the signal magnitude is of the order $\sqrt{(n/m)(\log(s_n)/s_n^2)}$. Furthermore they prove a lower bound of the form $\sqrt{(n/m)/s_n^2}$, which matches the upper bound apart from the $\sqrt{\log s_n}$ factor (which does not appear to be an artifact of the algorithm). Again, note that linear measurements are advantageous over entry-wise ones, for which signal magnitude must scale like $\sqrt{(n/m)(\log(s_n)/s_n)}$ for this problem.

IV. Final Remarks

In this brief note we surveyed existing results over adaptive sensing of sparse signals. We considered both entry-wise and linear measurements and clarified in which situations can adaptive sensing yield interesting gains over non-adaptive designs. A clear picture exists for the unstructured scenario, where one assumes only that the support set S is sparse. If in addition one can make structural assumptions over S than it is clear that support estimation is possible for even weaker signals. With so few results available along those lines this remains an interesting avenue for future research.

Acknowledgment

The author would like to thank the anonymous reviewers for their helpful pointers and remarks.

References

[1] A. Wald, *Sequential Analysis*. John Wiley & Sons, Inc., 1947.

[2] H. Chernoff, "Sequential design of experiments," *The Annals of Mathematical Statistics*, vol. 30, no. 3, pp. 755–770, September 1959.

[3] M. A. El-Gamal, "The role of priors in active Bayesian learning in the sequential statistical decision framework," in *Maximum Entropy and Bayesian Methods*, W. T. Grandy Jr. and L. H. Schick, Eds. Kluwer, 1991, pp. 33–38.

[4] P. Hall and I. Molchanov, "Sequential methods for design-adaptive estimation of discontinuities in regression curves and surfaces," *The Annals of Statistics*, vol. 31, no. 3, pp. 921–941, 2003.

[5] G. Blanchard and D. Geman, "Hierarchical testing designs for pattern recognition," *The Annals of Statistics*, vol. 33, no. 3, pp. 1155–1202, 2005.

[6] D. Cohn, Z. Ghahramani, and M. Jordan, "Active learning with statistical models," *Journal of Artificial Intelligence Research*, no. 4, pp. 129–145, 1996.

[7] Y. Freund, H. S. Seung, E. Shamir, and N. Tishby, "Selective sampling using the query by committee algorithm," *Machine Learning*, vol. 28, no. 2-3, pp. 133–168, August 1997.

[8] E. Novak, "On the power of adaption," *Journal of Complexity*, vol. 12, pp. 199–237, 1996.

[9] A. Korostelev and J.-C. Kim, "Rates of convergence for the sup-norm risk in image models under sequential designs," *Statistics & probability Letters*, vol. 46, pp. 391–399, 2000.

[10] S. Dasgupta, "Coarse sample complexity bounds for active learning," in *Advances in Neural Information Processing (NIPS)*, 2005.

[11] S. Hanneke, "Rates of convergence in active learning," *Annals of Statistics*, vol. 39, no. 1, pp. 333–361, 2010.

[12] V. Koltchiinskii, "Rademacher complexities and bounding the excess risk in active learning," *Journal of Machine Learning Research*, vol. 11, September 2010.

[13] N. Balcan, A. Beygelzimer, and J. Langford, "Agostic active learning," in *23rd International Conference on Machine Learning*, Pittsburgh, PA, 2006.

[14] R. Castro and R. Nowak, "Minimax bounds for active learning," *IEEE Transactions on Information Theory*, vol. 54, no. 5, pp. 2339–2353, July 2008.

[15] J. Haupt, R. Castro, and R. Nowak, "Distilled sensing: Adaptive sampling for sparse detection and estimation," *IEEE Transactions on Information Theory*, vol. 57, no. 9, pp. 6222 – 6235, September 2011.

[16] L. Addario-Berry, N. Broutin, L. Devroye, and G. Lugosi, "On combinatorial testing problems," *The Annals of Statistics*, vol. 38, no. 5, pp. 3063–3092, 2010.

[17] D. Donoho and J. Jin, "Higher criticsm for detecting sparse heterogenous mixtures," *Annals of Statistics*, vol. 32, no. 3, pp. 962–994, 2004.

[18] R. M. Castro, "Adaptive sensing performance lower bounds for sparse signal estimation and detection," *Preprint*, 2012, (available at http://arxiv.org/abs/1206.0648).

[19] Y. Ingster, "Some problem of hypothesis testing leading to infinitely divisible distributions," *Mathematical Methods of Statistics*, vol. 6, pp. 47–69, 1997.

[20] Y. Ingster and I. Suslina, *Nonparametric Goodness-of-Fit Testing under Gaussian Models*, ser. Lecture Notes in Statistics. Springer, 2003.

[21] E. Arias-Castro, E. J. Candès, H. Helgason, and O. Zeitouni, "Searching for a trail of evidence in a maze," *The Annals of Statistics*, vol. 36, no. 4, pp. 1726–1757, 2008.

[22] C. Butucea and Y. Ingster, "Detection of a sparse submatrix of a high-dimensional noisy matrix," *Preprint*, 2011, (available at http://arxiv.org/abs/1109.0898).

[23] M. Malloy and R. Nowak, "Sequential analysis in high-dimensional multiple testing and sparse recovery," in *The IEEE International Symposium on Information Theory*, Saint Petersburg, August 2011, (available at http://arxiv.org/abs/1103.5991v1).

[24] ——, "On the limits of sequential testing in high dimensions," in *Asilomar Conf. on Signals, Systems, and Computers*, Pacific Grove, California, 2011, (available at http://arxiv.org/abs/1105.4540).

[25] E. Candès and M. Davenport, "How well can we estimate a sparse vector?" *Applied and Computational Harmonic Analysis*, vol. 34, no. 2, pp. 317–323, 2013, (available at http://arxiv.org/abs/1111.4646).

[26] D. L. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[27] E. J. Candès and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?" *IEEE Transactions on Information Theory*, vol. 52, no. 12, pp. 5406–5425, Dec. 2006.

[28] E. Arias-Castro, "Detecting a vector based on linear measurements," *Electronic Journal of Statistics*, vol. 6, pp. 547–558, 2012.

[29] E. Arias-Castro and M. Davenport, "Compressive binary search," in *The IEEE International Symposium on Information Theory*, Cambridge, Massachusetts, July 2012, (available at http://arxiv.org/abs/1202.0937).

[30] E. Arias-Castro, E. Candès, and M. Davenport, "On the fundamental limits of adaptive sensing," *Preprint*, 2011, (available at http://arxiv.org/abs/1111.4646).

[31] M. Malloy and R. Nowak, "Near-optimal adaptive compressed sensing," in *Asilomar Conf. on Signals, Systems, and Computers*, Pacific Grove, California, 2012, (available at http://homepages.cae.wisc.edu/~mmalloy/papers/NearOptimalACS.pdf).

[32] J. Haupt, R. Baraniuk, R. Castro, and R. Nowak, "Sequentially designed compressed sensing," August 2012, (available at http://www.win.tue.nl/~rmcastro/publications/SCS.pdf).

[33] M. J. Wainwright, "Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso)," *IEEE Trans Inform Theory*, vol. 55, no. 5, 2009.

[34] S. Balakrishnan, M. Kolar, A. Rinaldo, and A. Singh, "Recovering block-structured activations using compressive measurements," *Preprint*, 2012, (available at http://arxiv.org/abs/1209.3431).

Reconstruction of solutions to the Helmholtz equation from punctual measurements

Gilles Chardon
Acoustics Research Institute
Austrian Academy of Sciences
Vienna, Austria
gilles.chardon@m4x.org

Albert Cohen
Laboratoire Jacques-Louis Lions
Université Pierre et Marie Curie
Paris, France
laurent.daudet@espci.fr

Laurent Daudet
Institut Langevin
Université Paris Diderot
Paris, France
cohen@ann.jussieu.fr

***Abstract*—We analyze the sampling of solutions to the Helmholtz equation (e.g. sound fields in the harmonic regime) using a least-squares method based on approximations of the solutions by sums of Fourier-Bessel functions or plane waves. This method compares favorably to others such as Orthogonal Matching Pursuit with a Fourier dictionary. We show that using a significant proportion of samples on the border of the domain of interest improves the stability of the reconstruction, and that using cross-validation to estimate the model order yields good reconstruction results.**

I. Introduction

Sampling an acoustical field (i.e. the spatial and temporal behavior of sound pressure) or a mechanical field (e.g. distribution of velocities on a vibrating membrane) is an ubiquituous task in experimental acoustics and mechanics. Usually, these fields are sampled on a uniform grid with density chosen according to the sampling theorem. However, in the particular cases mentionned above, the fields are known to satisfy the wave equation

$$\Delta u - \frac{1}{c^2}\frac{\partial^2 u}{\partial t^2} = 0, \tag{1}$$

or, in the harmonic regime, the Helmholtz equation

$$\Delta u + k^2 u = 0, \tag{2}$$

in two or three dimensions, where c is the wave velocity, and k the wavenumber. This fact allows to sample such fields with a reduced number of samples, with a least-squares method described in section 2. Of interest here is the choice of the repartition of the sampling points on the domain of interest, and the choice of the order of approximation used in the least-squares reconstruction. In section 3, we recall the results given in [1] on the stability of the reconstruction in function of the sampling scheme for the case of the disk, and extend it to the 3D-ball. We also gives numerical evidence for the case of the square, further showing that sampling on the border of the domain as well as inside improves the stability of the reconstruction. Finally, we give results of numerical simulations using cross-validation for the determination of the model order in section 4.

II. Reconstruction method

Our goal here, given a solution to the Helmholtz equation (2) in a domain $D \subset \mathbf{R}^d, d = 2$ or 3, is to reconstruct it in a domain $\Omega \subset D$ from a limited number of punctual measurements, without knowing the shape of D or the boundary conditions on ∂D. The reconstruction scheme we use is based on the Vekua theory and least-squares approximations, and has already been shown to compare favorably with existing methods such as OMP using sparsity in the Fourier domain [1], and to give good results in experimental settings [2].

The Vekua theory [3], in its general formulation, allows to build approximations of solutions to general elliptic partial differential equations, by building operators mapping these solutions to harmonic functions and reciprocally. Approximation of harmonic functions by harmonic polynomials can then be translated as approximation of solutions of the PDE by the images of the polynomials. The particular case of the Helmholtz equation in 2 and 3 dimensions has been analyzed by Moiola *et al.* [4]. In this case, the images of the polynomials are the so-called *generalized harmonic polynomials*. In two dimensions, the space of generalized harmonic polynomials of order L is given in polar coordinates (r, θ) by

$$HP_{k,L} = \operatorname*{span}_{l=-L,\cdots,L} e^{il\theta} J_l(kr)$$

where J_l is the l-th Bessel function. In three dimensions, these spaces are defined in spherical coordinates (r, θ, ϕ)

by

$$HP_{k,L} = \operatorname*{span}_{\substack{l=0,\cdots,L \\ m=-l,\cdots,l}} Y_{lm}(\theta,\phi) j_l(kr)$$

where Y_{lm} are the spherical harmonics, and j_l the spherical Bessel functions. Note that in two dimensions, the dimension of $HP_{k,L}$ is $2L+1$, while it is $(L+1)^2$ in three dimensions.

Their main result, given here in a simplified form and for convex domains, is as follows:

Theorem 1. *[4] Let $u \in H^K(\Omega), K \geq 1$ be a solution to the Helmholtz equation in the convex domain $\Omega \in \mathbf{R}^d$, $d = 2, 3$. Then, for $j < K$, there exists a generalized harmonic polynomial $\tilde{u}_L$ of order L such that, in two dimensions,*

$$\|u - \tilde{u}_L\|_{H^j} \leq C \left(\frac{L}{\log L}\right)^{K-j} \|u\|_{H^K},$$

and in three dimensions,

$$\|u - \tilde{u}_L\|_{H^j} \leq C L^{\lambda(K-j)} \|u\|_{H^K},$$

where λ depends only on the shape of Ω.

The result also holds for star-shaped domains, with a slower convergence. Identical results are also available for approximation by plane waves.

To reconstruct a solution u to the Helmholtz equation using n samples, we fix an order of approximation L such that $m = \dim HP_{k,L} \leq n$, and estimate u by the function $\tilde{u} \in HP_{k,L}$ minimizing the sum of the squares of the errors between values u_j sampled at the points x_j and $\tilde{u}(x_j)$, the sampling points being drawn using a predefined density on Ω:

$$\tilde{u} = \min_{\hat{u} \in HP_{k,L}} \sum_{j=1}^{n} |\hat{u}(x_j) - u_j|^2.$$

Such a reconstruction scheme is not always stable. A theorem, from Cohen et al [5], gives indication whether the reconstruction $\tilde{u}$ in a m-dimensional subspace using n samples drawn with probability density ν is stable. With $(L_j)_{j=1\cdots m}$ an orthogonal basis(with respect to the scalar product defined by the density ν) of the subspace, we define

$$K(m) = \max_{x \in \Omega} \sum_{j=1}^{m} |L_j(x)|^2 .$$

The result is as follows:

Theorem 2. *[5] Let $r > 0$ be arbitrary but fixed and let $\kappa := \frac{1-\log 2}{2+2r}$. If m is such that*

$$K(m) \leq \kappa \frac{n}{\log n}, \tag{3}$$

then, one has

$$E(\|u - \tilde{u}\|^2) \leq (1 + \epsilon(n))\sigma_m(u)^2 + 8M^2 n^{-r}, \tag{4}$$

where $\epsilon(n) := \frac{4\kappa}{\log n} \to 0$ as $n \to +\infty$, $\sigma_m(u)$ is the best approximation error, M a upper bound of $|u|$ and $\tilde{u}$ the least square approximation of u thresholded such that $|\tilde{u}| \leq M$.

This suggests that the slowest $K(m)$ increases, the largest m can be, allowing a better reconstruction. The choice of the density ν is here important, as K is dependent on it. This means that choosing a adequate density allows to use a lower number of samples that, e.g. the uniform density. Note however that the choice of the density ν also affects the norm used in theorem 2 to measure the error, which can be different than the norm we are interested in. We are here interested in the stability for the standard L^2 norm. We thus choose a density of the form $\nu = (1-\alpha)\lambda + \alpha\nu'$ where λ is the uniform density, and ν' an arbitrary fixed density. The choice of the density ν' and the parameter α is discussed in the next section for some particular cases.

III. Choice of the sample distribution

Here, we will concentrate on measures $\nu = (1-\alpha)\lambda + \alpha\sigma$, where the support of σ is the boundary of Ω. This heuristic is supported by the following results on the disk and the ball. For these two cases, we will give estimations of K for particular values of m, i.e. the size of the spaces $HP_{k,L}$, $m = 2L+1$ in 2D and $m = (L+1)^2$ in 3D.

For the case of the disk with densities $\nu_\alpha = (1-\alpha)\lambda + \alpha\sigma$ where σ is the uniform measure on the circle, the Fourier-Bessel functions, after normalization, form an orthogonal basis for these measures. Using properties of the Bessel functions, we can estimate the behaviour of $K(m)$ in function of α:

Theorem 3. *[1] For the approximation by generalized harmonic polynomials on the unit disk, one has for sufficiently large m*

$$K(2L+1) \geq c_0 + c_1 L^2$$

when $\alpha = 0$ for any $c_1 < 1/4$ and where c_0 depends on c_1 and λ, and

$$K(2L+1) \leq C + \frac{2L+1}{\alpha}$$

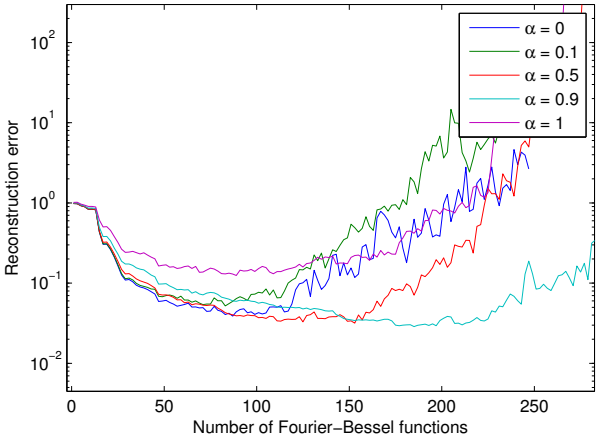


Fig. 1. Reconstruction error in function of the order model for 400 measurements, with different proportions α of samples on the border.

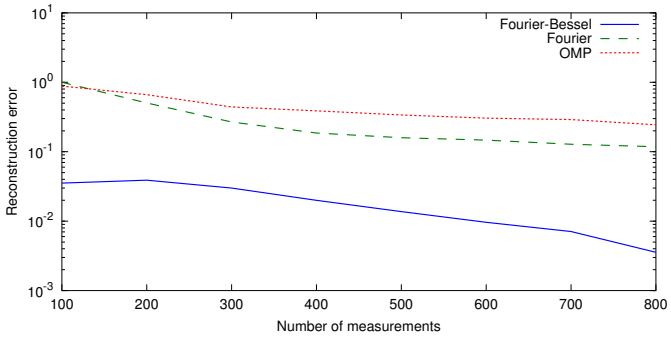


Fig. 2. Reconstruction error in function of the number of measurements for least square with Fourier-Bessel function, Fourier modes, and Orthogonal Matching Pursuit with a dictionnary of Fourier modes

when $\alpha > 0$, and C depends on λ and α.

A similar result is available for plane waves approximations. This theorem shows that using samples on the border needs a number of samples proportionnal to the dimension of $HP_{k,L}$ to ensure stability, while sampling only inside the disk needs more samples.

The effect of the coefficient α is shown on figure 1, where the approximation error in function of α and the order of the model is given, for the recovery of a solution of the Helmholtz equation with $k = 12$, using $n = 400$ measurements. We see that a large proportion of samples on the border allows a large order model, which improve the reconstruction result. However, using samples on the border only ($\alpha = 1$) is detrimental to the reconstruction error, as in this case, theorem 2 controls the error in the norm defined by ν_1 which is the L_2-norm on the circle only.

We compare on figure 2 the results of the least-squares method with Fourier-Bessel functions, OMP with a large dictionary of Fourier modes defined on a square containing the disk, and the least-squares method with a smallest dictionary. The reconstruction of the least-squares method combined with the Fourier-Bessel approximation are clearly better than the two other tested methods.

A slightly modified proof, using properties of the spherical Bessel functions and of the spherical harmonics, yields the following result for the 3D case, with σ the uniform measure on the sphere:

Theorem 4. *For the approximation by generalized harmonic polynomials on the unit ball, one has for sufficiently large L*

$$K((L+1)^2) \geq c_0 + c_1 L^3$$

when $\alpha = 0$ for any $c_1 < 1/9$ and where c_0 depends on c_1 and λ, and

$$K((L+1)^2) \leq C + \frac{(L+1)^2}{\alpha}$$

when $\alpha > 0$, where C depends on λ and α.

In this case, the number of measurements needed to ensure stability grows faster than the dimension of $HP_{k,L}$ for the uniformly dense sampling, while being proportional to this dimension when using additional samples on the border.

We now turn to the case of the square. As neither the Fourier-Bessel functions, nor the plane waves, form an orthogonal basis, we constuct one by orthogonalizing the plane waves, using the Gram matrix of the plane waves families which can be computed exactly in the case of the measures described below.

We numerically compute $K(m)$ for three differents distributions:

- $\nu_0 = \lambda$, the uniform distribution on the square
- $\nu_\alpha = (1-\alpha)\lambda + \alpha\sigma$, where σ is the uniform distribution on the boundary of the square
- $\nu'_\alpha = (1-\alpha)\lambda + \alpha\sigma'$, where σ' is the measure on the boundary with weight $1/4\pi\sqrt{1-s^2}$ where $s = \min(x, y)$.

The estimated values of $K(m)$ for ν_0, $\nu_{1/2}$ and $\nu'_{1/2}$ are given on figure 3. Here, sampling on the border of the square improves the stability of the reconstruction compared to the uniform case, but still needs a high number of samples.

Using the non-uniform sampling on the border, with more samples in the sections of the boundary furthest

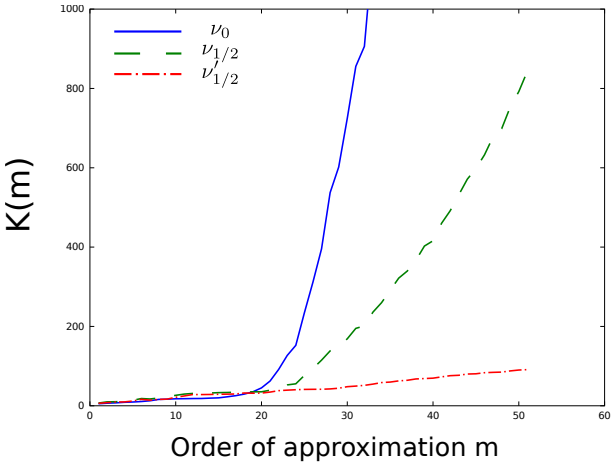


Fig. 3. Numerical evaluation of K for three different samples distribution on the square.

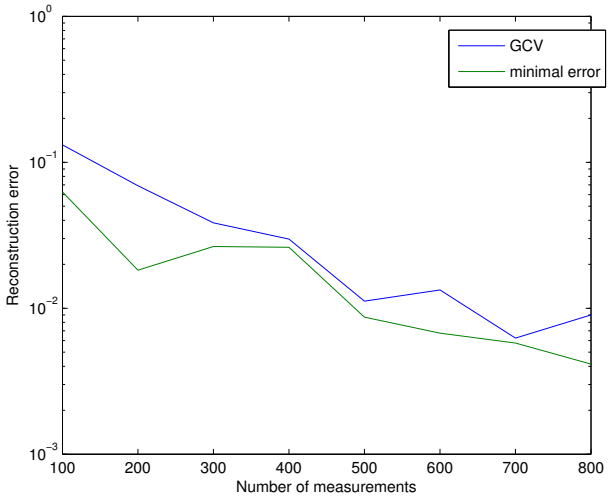


Fig. 4. Comparison of the reconstruction using the generalized cross validation to estimate the order model, and the best reconstruction error.

from the origin, makes the behaviour of $K(m)$ comparable to m, which is the best case possible.

IV. Choice of the model order

Once the number of samples and their distribution are fixed, the model order used in the reconstruction has to be chosen. While a sufficient number of plane waves or Fourier-Bessel functions are needed (physical arguments recommend a number proportionnal to the product of the wavenumber and the diameter of the domain), using a too large order can result in overfitting, as visible figure 1.

A way to estimate the best order m to use is the cross-validation. Given m samples, we reconstruct f from a subset of m' samples, and compute the reconstruction error on the remaining $m - m'$ samples. We then repeat with different subsets, and chose the order for which the average error is minimal. Figure 4 compares the best reconstruction error knowing f, and the reconstruction error using the order estimated using the cross-validation.

V. Conclusion

The sampling of solutions to the Helmholtz equation is interesting both for its experimental applications as well as for theoretical developments. We showed here that a careful choice of the density of the samples can improve the stability of the reconstruction, with theoretical results in simple cases, and numerical simulations in more general settings. We also show that using cross-validation to estimate the model order yields good results.

A general sampling strategy, *i.e.* a choice of the sample density, dependent on the shape of the domain of interest and of the frequency k, and possibly on the order m is yet to be designed.

Acknowledgment

The first author acknowledges support from the Austrian Science Fund (FWF) START-project FLAME (Frames and Linear Operators for Acoustical Modeling and Parameter Estimation; Y 551-N13).

The authors acknowledge partial support from Agence Nationale de la Recherche (ANR), project ECHANGE (ANR-08-EMER-006).

References

[1] G. Chardon, A. Cohen, and L. Daudet, "Approximation of solutions to the helmholtz equation from scattered data." Arxiv preprint 1301.0237.

[2] G. Chardon, A. Leblanc, and L. Daudet, "Plate impulse response spatial interpolation with sub-nyquist sampling," *Journal of Sound and Vibration*, vol. 330, pp. 5678–5689, November 2011.

[3] I. N. Vekua, *New methods for solving elliptic equations*. North-Holland, 1967.

[4] A. Moiola, R. Hiptmair, and I. Perugia, "Plane wave approximation of homogeneous Helmholtz solutions," *Zeitschrift für Angewandte Mathematik und Physik (ZAMP)*, vol. 62, pp. 809–837, 2011. 10.1007/s00033-011-0147-y.

[5] A. Cohen, M. A. Davenport, and D. Leviatan, "On the stability and accuracy of least squares approximations,," tech. rep., arXiv:1111.4422v1, 2011.

A priori convergence of the Generalized Empirical Interpolation Method

Yvon Maday
UPMC Univ Paris 06, UMR 7598,
Laboratoire Jacques-Louis Lions,
F-75005, Paris, France.
and Institut Universitaire de France
and Division of Applied Mathematics,
Brown University, Providence RI, USA.
Email: maday@ann.jussieu.fr

Olga Mula
CEA Saclay,
DEN/DANS/DM2S/SERMA/LLPR
91191 Gif-Sur-Yvette CEDEX - France
and LRC MANON,
Laboratoire de Recherche Conventionnée
CEA/DEN/DANS/DM2S
and UPMC-CNRS/LJLL.
Email: olga.mulahernandez@cea.fr

Gabriel Turinici
CEREMADE
Universite Paris Dauphine
Place du Marechal de Lattre de Tassigny
75016 Paris, France
Email: Gabriel.Turinici@dauphine.fr

***Abstract*—In an effort to extend the classical Lagrangian interpolation tools, new interpolating methods that use general interpolating functions are explored. The Generalized Empirical Interpolation Method (GEIM) belongs to this class. It generalizes the plain Empirical Interpolation Method [1] by replacing the evaluation at interpolating points by application of a class of interpolating linear functions. Since its efficiency depends critically on the choice of the interpolating functions (that are chosen by a Greedy selection procedure), the purpose of this paper is therefore to provide a priori convergence rates for the Greedy algorithm that is used to build the GEIM interpolating spaces.**

I. Introduction

The extension of the Lagrangian interpolation process is an old problem that is still currently subject to active research (see, e.g. [1] and also the activity concerning the kriging [2], [3] in the stochastic community). While this classical method approximates general functions by finite sums of well chosen, linearly independent interpolating functions (e.g. polynomial functions) and the optimal location of the interpolating points is well documented (and completely solved in one dimension), the question remains on how to approximate general functions by finite expansions involving general interpolating functions and how to optimally select the interpolation points in this case.

One step in this direction is the Empirical Interpolation Method (EIM, [4], [5], [1]) that has been developed in the broad framework where the functions f to approximate belong to a compact set F of a Banach space $\mathcal{X}$. The set F is supposed to be such that any $f \in F$ is approximable by linear combinations of small size. In particular, this is the case when the Kolmogorov $n-$width of F in $\mathcal{X}$ is small. Indeed, the Kolmogorov $n-$width of F in $\mathcal{X}$ is defined by $d_n(F,\mathcal{X}) := \inf_{\substack{X_n \subset \mathcal{X} \\ dim(X_n)=n}} \sup_{x\in F} \inf_{y\in X_n} \|x-y\|_{\mathcal{X}}$ (see [6]) and measures the extent to which F can be approximated by finite dimensional spaces $X_n \subset \mathcal{X}$ of dimension n. The Empirical Interpolation Method builds simultaneously the set of interpolating functions and the associated interpolating points by a greedy selection procedure (see [4]).

A recent generalization of this interpolation process consists in replacing the evaluation at interpolating points by application of a class of interpolating continuous linear functions chosen in a given dictionary $\Sigma \subset \mathcal{L}(F)$ and this gives rise to the so-called Generalized Empirical Interpolation Method (GEIM, [7]). In this newly developed method, the particular case where the space $\mathcal{X} = L^2(\Omega)$ is considered, with Ω being a bounded spatial domain of $\mathbb{R}^d$ and F being a compact set of $L^2(\Omega)$.

In the present work, we analyze the quality of the finite dimensional subspaces X_n contained in the span of F built by the greedy selection procedure of GEIM together with the properties of the associated interpolation operator. For this purpose, the accuracy of the approximation in X_n of the elements of F will be compared to the best possible performance which is the Kolmogorov $n-$ width $d_n(F, L^2(\Omega))$.

The methodology developed in this paper is in the spirit of the greedy reduced basis method. Alternative approaches exist like POD and gappy POD or even Adaptive Cross Approximation. We refer to the review paper [8] for a comparative presentation of all these sampling approaches.

The proceeding is organized as follows: after a brief recall of GEIM's Greedy algorithm (section II), we will analyze in sections III and IV some convergence decay rates of the generalized empirical interpolation error as the dimension n of X_n increases and when $d_n(F, L^2(\Omega))$ has a polynomial or an exponential decreasing behavior.

II. The generalized Empirical Interpolation Method

In the following, we assume that the dimension of the vectorial space spanned by F is of dimension $\geq \mathcal{N}$.

In a similar procedure as in the Empirical Interpolation Method (EIM) [4], [5], [1], the Generalized EIM allows to define simultaneously the set of interpolating functions recursively chosen in F together with the associated linear

functions selected from a dictionary of continuous linear forms $\Sigma \subset \mathcal{L}(F)$, with norm 1 in $L^2(\Omega)$. The dictionary has the additional property that if $\varphi \in F$ is such that $\sigma(\varphi) = 0$ for any $\sigma \in \Sigma$, then $\varphi = 0$. The selection of the interpolating functions and linear forms is based on a greedy selection procedure as outlined in [7].

The first interpolating function is, e.g.: $\varphi_0 = \arg\sup_{\varphi \in F} \|\varphi\|_{L^2(\Omega)}$. The first interpolating linear form is $\sigma_0 = \arg\sup_{\sigma \in \Sigma} |\sigma(\varphi_0)|$. We then define the first basis function as $q_0 = \dfrac{\varphi_0}{\sigma_0(\varphi_0)}$. The second interpolating function is $\varphi_1 = \arg\sup_{\varphi \in F} \|\varphi - \sigma_0(\varphi) q_0\|_{L^2(\Omega)}$. The second interpolating linear form is $\sigma_1 = \arg\sup_{\sigma \in \Sigma} |\sigma(\varphi_1 - \sigma_0(\varphi_1) q_0)|$ and the second basis function is defined as $q_1 = \dfrac{\varphi_1 - \sigma_0(\varphi_1) q_0}{\sigma_1(\varphi_1 - \sigma_0(\varphi_1) q_0)}$.

We then proceed by induction : assume that we have built the set of interpolating functions $\{q_0, q_1, \ldots, q_{N-1}\}$ and the set of associated interpolating linear forms $\{\sigma_0, \sigma_1, \ldots, \sigma_{N-1}\}$, for $1 \leq N \leq N_{max}$, with $N_{max} \leq \mathcal{N}$ being an upper bound fixed *a priori*. For $N \leq 1$, we first solve the interpolation problem: find $\{\alpha_j^N(\varphi)\}_j$ such that: $\forall i = 0, \ldots, N-1, \quad \sigma_i(\varphi) = \sum_{j=0}^{N-1} \alpha_j^N(\varphi) \sigma_i(q_j)$. We then compute $\mathcal{J}_N[\varphi] = \sum_{j=0}^{N-1} \alpha_j^N(\varphi) q_j$ and evaluate $\varepsilon_N(\varphi) = \|\varphi - \mathcal{J}_N[\varphi]\|_{L^2(\Omega)}, \ \forall \varphi \in F$. We define $\varphi_N = \arg\sup_{\varphi \in F} \varepsilon_N(\varphi)$ and $\sigma_N = \arg\sup_{\sigma \in \Sigma} |\sigma(\varphi_N - \mathcal{J}_N[\varphi_N])|$. The next basis function is then $q_N = \dfrac{\varphi_N - \mathcal{J}_N[\varphi_N]}{\sigma_N(\varphi_N - \mathcal{J}_N[\varphi_N])}$

We finally set $X_{N+1} \equiv \text{span}\{q_j, \ j \in [0, N]\} = \text{span}\{\varphi_j, \ j \in [0, N]\}$. It has been proven in [7]:

Lemma 1: For any $N \leq \mathcal{N}$, the set $\{q_j, \ j \in [0, N-1]\}$ is linearly independent and X_N is of dimension N. The generalized empirical interpolation procedure is well-posed in $L^2(\Omega)$ and $\forall \varphi \in F$, the interpolation error satisfies:

$$\|\varphi - \mathcal{J}_N[\varphi]\|_{L^2(\Omega)} \leq (1 + \Lambda_N) \inf_{\psi_N \in X_N} \|\varphi - \psi_N\|_{L^2(\Omega)}$$

where Λ_N is the Lebesgue constant in the L^2 norm: $\Lambda_N := \sup_{\varphi \in F} \dfrac{\|\mathcal{J}_N[\varphi]\|_{L^2(\Omega)}}{\|\varphi\|_{L^2(\Omega)}}$.

Remark 1: In a similar way as in the classical Lagrangian interpolation, the Lebesgue constant Λ_N defined in our generalized interpolation procedure depends both on set F and on the choice of the dictionary of continuous linear forms Σ but no detailed analysis of the behavior of Λ_N as a function of F or Σ has been carried out so far.

Remark 2: In practice the selection of the interpolation functions in F and the interpolating elements in the dictionary can be done by discretizing both F and Σ as is the case for standard greedy approximations like in [5], [6]; an alternative approach is [9] where the selection is done through a continuous algorithm based on an iterative sequence of optimization problems (solved by Newton methods) that seek to maximize the error between the RB approximation and the underlying true solution. The interpolants can be efficiently computed recursively as outlined in [10].

III. Preliminary notations and basic properties

In what follows, we denote by $(\varphi_n^*)_{n \geq 0}$ the orthonormal system obtained from $(\varphi_n)_{n \geq 0}$ by Gram-Schmidt orthogonalization.

For any $n \geq 1$, we define the orthogonal projector P_n from $\mathcal{X}$ onto X_n which is given by $P_n(f) = \sum_{j=0}^{n-1} < f, \varphi_j^* > \varphi_j^*$, $\forall f \in F$, where $< ., . >$ is the $L^2(\Omega)$ scalar product. In particular: $\varphi_n = P_{n+1}(\varphi_n) = \sum_{j=0}^{n} a_{n,j} \varphi_j^*$, with $a_{n,j} := < \varphi_n, \varphi_j^* >, 0 \leq j \leq n$.

Finally, let us denote $\tau_0(F)_{L^2(\Omega)} := d_0(F, L^2(\Omega))$ and, for any $n \geq 1$: $\tau_n := \tau_n(F)_{L^2(\Omega)} := \max_{f \in F} \|f - P_n(f)\|_{L^2(\Omega)}$ and by γ_n the constant $\gamma_n = 1/(1 + \Lambda_n)$.

We begin by proving the two following lemmas:

Lemma 2: For any $n \geq 1$, $\|\varphi_n - P_n(\varphi_n)\|_{L^2(\Omega)} \geq \gamma_n \tau_n(F)$.

Proof: From lemma 1 applied to $\varphi = \varphi_n$ we have $\|\varphi_n - P_n(\varphi_n)\|_{L^2(\Omega)} \geq \gamma_n \|\varphi_n - \mathcal{J}_n(\varphi_n)\|_{L^2(\Omega)}$. But $\|\varphi_n - \mathcal{J}_n(\varphi_n)\|_{L^2(\Omega)} \geq \|\varphi - \mathcal{J}_n(\varphi)\|_{L^2(\Omega)}$ for any $\varphi \in F$ according to the definition of φ_n. Thus $\|\varphi_n - P_n(\varphi_n)\|_{L^2(\Omega)} \geq \gamma_n \|\varphi - \mathcal{J}_n(\varphi)\|_{L^2(\Omega)} \geq \gamma_n \|\varphi - P_n(\varphi)\|_{L^2(\Omega)}$. ■

Lemma 3: Let A be the lower triangular matrix defined by $A := (a_{i,j})_{i,j=0}^{\infty}$ ($a_{i,j} := 0, \ j > i$). A has two important properties:

- P1: $\gamma_n \tau_n \leq |a_{n,n}| \leq \tau_n$.
- P2: For every $m \geq n$, $\sum_{j=n}^{m} a_{m,j}^2 \leq \tau_n^2$.

  Proof:
- P1: $\forall f \in F : \ P_n(f) = \sum_{j=0}^{n-1} < f, \varphi_j^* > \varphi_j^*$. In particular: $\varphi_n - P_n(\varphi_n) = a_{n,n} \varphi_n^* \Rightarrow \|\varphi_n - P_n(\varphi_n)\|_{L^2(\Omega)}^2 = a_{n,n}^2$. The upper bound is thus obvious and Lemma 2 gives the lower bound.
- P2: For every $m \geq n$: $\sum_{j=n}^{m} |a_{m,j}|^2 = \|\varphi_m - P_n(\varphi_m)\|_{L^2(\Omega)}^2 \leq \max_{f \in F} \|f - P_n(f)\|^2 = \tau_n^2$.

■

IV. A priori convergence rates of the GEIM Greedy method

In order to get convergence decay rates in the generalized interpolation error of our method, we first note that lemma 2 shows that the GEIM's Greedy algorithm is what is called in [11] a "weak Greedy algorithm" of parameter $\gamma_n = 1/(1+\Lambda_n)$ that depends on the dimension of X_n.

Thanks to this observation, we shall derive convergence decay rates in the sequence $(\tau_n)_{n \geq 0}$. This task consists in extending the proofs of [11] where the constant case $\gamma_n = \gamma$ was addressed and where the following two results were proven in Corollary 3.3:

i) If $d_n(F) \leq C_0 n^{-\alpha}$ for $n \geq 1$, then $\tau_n \leq C_0 2^{5\alpha+1} \gamma^{-2} n^{-\alpha}$ for $n \geq 1$.
ii) If $d_n(F) \leq C_0 e^{-c_0 n^{\alpha}}$ for $n \geq 1$, then $\tau_n \leq \sqrt{2C_0} \gamma^{-1} e^{-c_1 n^{\alpha}}$ for $n \geq 1$, where $c_1 := 2^{-1-2\alpha} c_0$.

In order to extend $i)$ and $ii)$ to the more general case where γ depends on the dimension n, the following preliminary theorem is required:

***Theorem** 4:* For any $N \geq 0$, consider the weak Greedy algorithm with constant γ_N in $L^2(\Omega)$ associated with the compact set F. We have the following inequalities between τ_N and $d_N := d_N(F, L^2(\Omega))$: for any $K \geq 1,\ 1 \leq m < K$

$$\prod_{i=1}^{K} \tau_{N+i}^2 \leq \frac{1}{\prod_{i=1}^{K} \gamma_{N+i}^2} \left(\frac{K}{m}\right)^m \left(\frac{K}{K-m}\right)^{K-m} \tau_{N+1}^{2m} d_m^{2(K-m)}.$$

Proof: This result is an extension of Theorem 3.2 of [11] to the case where the parameter of the weak Greedy algorithm (γ_N) depends on the dimension of the reduced space X_N. Its proof is a slight modification to the one provided in [11] using γ_N and the properties P1 and P2 stated in Lemma 3. ■

Using theorem 4, convergence rates in the sequence $(\tau_n)_{n\geq 0}$ when $(d_n)_{n\geq 0}$ has a polynomial or an exponential decay can be inferred and lead to lemmas 5 and 6:

***Lemma** 5 (Polynomial decay of $(d_n)_{n\geq 0}$):* For any $n \geq 1$, let $n = 4\ell + k$ (where $\ell \in \{0, 1, \dots\}$ and $k \in \{0, 1, 2, 3\}$). Assume that there exists a constant $C_0 > 0$ such that $\forall n \geq 1$, $d_n(F, L^2(\Omega)) \leq C_0 n^{-\alpha}$, then $\tau_n \leq C_0 \beta_n n^{-\alpha}$, where $\beta_1 = 2$ and for $n \geq 2$: $\beta_n = \beta_{4\ell+k} := \sqrt{2\beta_{\ell_1}} \frac{1}{\prod_{i=1}^{\ell_2} \gamma_{\ell_1 - \lceil \frac{k}{4} \rceil + i}^{\frac{1}{\ell_2}}} (2\sqrt{2})^{\alpha}$ and $\ell_1 = 2\ell + \lfloor \frac{2k}{3} \rfloor$, $\ell_2 = 2\left(\ell + \lceil \frac{k}{4} \rceil\right)$.

Proof: The proof is done by recurrence over n. We initialize the reasoning by proving that $\tau_1 \leq 2C_0$ and then prove the general statement for $n \geq 2$.

Case $n = 1$: We recall that $\varphi_0 = \arg\sup_{\varphi \in F} \|\varphi\|_{L^2(\Omega)}$ and that P_1 is the projector operator onto $\text{span}\{\varphi_0\}$. We set: $f_1 = \arg\tau_1 = \arg\max_{f\in F} \|f - P_1(f)\|_{L^2(\Omega)}$ and let $\mu \in F$ span the one dimensional subspace of F for which $d_1 \geq \|f - P_\mu(f)\|_{L^2(\Omega)}$ for any $f \in F$ (P_μ being the projector operator onto $\text{span}\{\mu\}$). We have: $\tau_1 = \|f_1 - P_1(f_1)\|_{L^2(\Omega)} = \|f_1 - P_\mu(f_1) + P_\mu(f_1) - P_1(f_1)\|_{L^2(\Omega)} = \|f_1 - P_\mu(f_1) - P_1(f_1 - P_\mu(f_1)) + P_\mu(f_1) - P_1 P_\mu(f_1)\|_{L^2(\Omega)} \leq d_1 + \|P_\mu(f_1) - P_1 P_\mu(f_1)\|_{L^2(\Omega)}$.

We have: $\|P_\mu(f_1) - P_1 P_\mu(f_1)\|_{L^2(\Omega)} = \|\frac{< f_1, \mu > \mu}{\|\mu\|^2_{L^2(\Omega)}} - \frac{\langle < f_1, \mu > \mu, \varphi_0 \rangle \varphi_0}{\|\mu\|^2_{L^2(\Omega)} \|\varphi_0\|^2_{L^2(\Omega)}}\|_{L^2(\Omega)} = \frac{|< f_1, \mu >|}{\|\mu\|_{L^2(\Omega)}} \|\frac{\mu}{\|\mu\|_{L^2(\Omega)}} - \frac{< \varphi_0, \mu > \varphi_0}{\|\mu\|_{L^2(\Omega)} \|\varphi_0\|^2_{L^2(\Omega)}}\|_{L^2(\Omega)}$.

Since for any $x, y \in F$ with norm 1 we have $\|x - < x, y > y\|_{L^2(\Omega)} = \|y - < x, y > x\|_{L^2(\Omega)}$, we deduce that : $\|P_\mu(f_1) - P_1 P_\mu(f_1)\|_{L^2(\Omega)} = \frac{|< f_1, \mu >|}{\|\mu\|_{L^2(\Omega)}} \|\frac{\varphi_0}{\|\varphi_0\|_{L^2(\Omega)}} - \frac{< \varphi_0, \mu > \mu}{\|\mu\|^2_{L^2(\Omega)} \|\varphi_0\|_{L^2(\Omega)}}\|_{L^2(\Omega)}$.

Hence: $\tau_1 \leq d_1 + \frac{|< f_1, \mu >|}{\|\mu\|_{L^2(\Omega)} \|\varphi_0\|_{L^2(\Omega)}} \|\varphi_0 - \frac{< \varphi_0, \mu > \mu}{\|\mu\|^2_{L^2(\Omega)}}\|_{L^2(\Omega)} \leq d_1 \left(1 + \frac{|< f_1, \mu >|}{\|\mu\|_{L^2(\Omega)} \|\varphi_0\|_{L^2(\Omega)}}\right) \leq 2d_1$.

***Remark** 3:* In the case where $\|\varphi_0\|_{L^2(\Omega)} \geq \gamma_0 \|f\|_{L^2(\Omega)}$ for any $f \in F$ $(0 < \gamma_0 \leq 1)$, we would have obtained $\tau_1 \leq d_1 \left(1 + \frac{1}{\gamma_0}\right)$.

Case $n \geq 2$: Let $n = N + K$ for any $N \geq 0$, $K \geq 2$. If $i \leq K$, we have $\tau_n = \tau_{N+K} \leq \tau_{N+i}$ from the monotonicity of $(\tau_n)_{n\geq 0}$. By combining this inequality with theorem 4, if $1 \leq m < K$, we derive that $\tau_n \leq \frac{1}{\prod_{i=1}^{K} \gamma_{N+i}^{\frac{1}{K}}} \sqrt{\left(\frac{K}{m}\right)^{\frac{m}{K}} \left(\frac{K}{K-m}\right)^{1-\frac{m}{K}}} \tau_{N+1}^{\frac{m}{K}} d_m^{1-\frac{m}{K}} \leq \frac{1}{\prod_{i=1}^{K} \gamma_{N+i}^{\frac{1}{K}}} \sqrt{2} \tau_{N+1}^{\frac{m}{K}} d_m^{1-\frac{m}{K}}$, since $x^{-x}(1-x)^{x-1} \leq 2$ for any x, $0 < x < 1$. We now use that $d_m \leq C_0 m^{-\alpha}$ and the recurrence hypothesis in $N + 1 < n : \tau_{N+1} \leq C_0 \beta_{N+1} (N+1)^{-\alpha}$ which yields: $\tau_{N+K} \leq C_0 \sqrt{2} \frac{1}{\prod_{i=1}^{K} \gamma_{N+i}^{\frac{1}{K}}} \beta_{N+1}^{\frac{m}{K}} \xi(n)^{\alpha} (N+K)^{-\alpha}$ where $\xi(n) = \frac{n}{m} \left(\frac{m}{N+1}\right)^{\frac{m}{K}}$.

Any $n \geq 2$ can be written as $n = 4\ell + k$ with $\ell \in \{0, 1, \dots\}$ and $k \in \{0, 1, 2, 3\}$. If $k = 1, 2$ or 3, it can easily be proven that $\xi(n) \leq 2\sqrt{2}$ by setting $N = 2\ell - 1,\ K = 2\ell + 2,\ m = \ell + 1$ if $k = 1$ and $\ell \geq 1$, $N = 2\ell,\ K = 2\ell + 2,\ m = \ell + 1$ if $k = 2$ and $\ell \geq 0$ and $N = 2\ell + 1,\ K = 2\ell + 2,\ m = \ell + 1$ if $k = 3$ and $\ell \geq 0$. These choices of N, K and m combined with the upper bound of ξ yield the result $\tau_n \leq C_0 \beta_n n^{-\alpha}$ in the case $k = 1, 2$ or 3.

In the case $n = 4\ell$ $(\ell \geq 1)$, using the fact that $\tau_{N+1} \leq \tau_N$, we can derive that $\tau_n \leq \frac{1}{\prod_{i=1}^{K} \gamma_{N+i}^{\frac{1}{K}}} \sqrt{2} \tau_N^{\frac{m}{K}} d_m^{1-\frac{m}{K}}$. If we choose $N = K = 2\ell$ and $m = \ell$, the previous inequality directly yields $\tau_{4\ell} \leq C_0 \sqrt{2\beta_{2\ell}} \frac{1}{\prod_{i=1}^{2\ell} \gamma_{2\ell+i}^{\frac{1}{2\ell}}} (2\sqrt{2})^{\alpha} (4\ell)^{-\alpha}$.

■

***Lemma** 6 (Exponential decay in $(d_n)_{n\geq 0}$):* Assume that there exists a constant $C_0 > 0$ such that $\forall n \geq 1$, $d_n(F, L^2(\Omega)) \leq C_0 e^{-c_1 n^{\alpha}}$, then $\tau_n \leq C_0 \beta_n e^{-c_2 n^{\alpha}}$, where $\beta_n := \frac{1}{\prod_{i=1}^{\lceil \frac{n}{2} \rceil} \gamma_{\lfloor \frac{n}{2} \rfloor + i}^{\frac{1}{\lceil \frac{n}{2} \rceil}}} \sqrt{2\beta_{\lfloor \frac{n}{2} \rfloor}}$ for $n \geq 2$, $\beta_1 = 2$ and $c_2 := 2^{-1-3\alpha} c_1$.

Proof: The proof is done by recurrence over n.

The case $n = 1$ is addressed by following the same lines as in lemma 5.

In the case $n = 2$, we have: $\tau_2 \leq \tau_1 \leq 2C_0$.

For $n \geq 3$, we start from $\tau_{N+K} \leq \frac{1}{\prod_{i=1}^{K} \gamma_{N+i}^{\frac{1}{K}}} \sqrt{2} \tau_{N+1}^{\frac{m}{K}} d_m^{1-\frac{m}{K}}$ and treat the cases $n = N + K = 2\ell$ and $n = N + K = 2\ell + 1$ separately $(\ell \geq 1)$.

If $n = N + K = 2\ell$, we choose $N = K = \ell$ and $m = \lfloor \frac{K}{2} \rfloor$.

The inequality yields $\tau_{2\ell} \leq \dfrac{1}{\prod_{i=1}^{\ell} \gamma_{\ell+i}^{\frac{1}{\ell}}} \sqrt{2\tau_\ell} e^{-c_2(2\ell)^\alpha}$.
In a similar procedure, the desired result can be inferred for $n = N + K = 2\ell + 1$ if we choose $N = \ell$, $K = \ell + 1$ and $m = \lfloor \frac{K}{2} \rfloor$. ■

***Remark** 4:* 1) In the case where γ_n is constant $\gamma_n = \gamma$, lemmas 5 and 6 yield results that are similar to the ones obtained in [11] (see results i) and ii) above).

2) In the case where $(\gamma_n)_{n\geq 1}$ is a monotonically decreasing sequence, the following bounds can be derived for τ_n:
   - If $d_n(F, L^2(\Omega)) \leq C_0 n^{-\alpha}$ for any $n \geq 1$, then $\tau_n \leq C_0 \beta n^{-\alpha}$ for $n \geq 1$, with $\beta := 2^{3\alpha+1} \left(\min_{1\leq j\leq n} \gamma_j\right)^{-2} = 2^{3\alpha+1} \gamma_n^{-2}$.
   - If $d_n(F, L^2(\Omega)) \leq C_0 e^{-c_1 n^\alpha}$ for any $n \in \{1, 2, \dots\}$, then $\tau_n \leq C_0 \beta e^{-c_2 n^{-\alpha}}$ for $n \geq 1$, with $\beta := 2 \left(\min_{1\leq j\leq n} \gamma_j\right)^{-2} = 2\gamma_n^{-2}$.

Lemmas 5 and 6 are the keys to derive the decay rates of the interpolation error of the GEIM Greedy algorithm. This is the purpose of the following theorem:

***Theorem** 7:* 1) Assume that $d_n(F, L^2(\Omega)) \leq C_0 n^{-\alpha}$ for any $n \geq 1$, then the interpolation error of the GEIM Greedy selection process satisfies for any $\varphi \in F$ the inequality $\|\varphi - \mathcal{J}_n[\varphi]\|_{L^2(\Omega)} \leq C_0(1 + \Lambda_n)\beta_n n^{-\alpha}$, where the parameter β_n is defined as in lemma 5.

2) Assume that $d_n(F, L^2(\Omega)) \leq C_0 e^{-c_1 n^\alpha}$ for any $n \geq 1$, then the interpolation error of the GEIM Greedy selection process satisfies for any $\varphi \in F$ the inequality $\|\varphi - \mathcal{J}_n[\varphi]\|_{L^2(\Omega)} \leq C_0(1 + \Lambda_n)\beta_n e^{-c_2 n^\alpha}$, where β_n and c_2 are defined as in lemma 6.

Proof: It can be inferred from lemma 1 that, $\forall \varphi \in F$, $\|\varphi - \mathcal{J}_n[\varphi]\|_{L^2(\Omega)} \leq (1 + \Lambda_n)\|\varphi - P_n(\varphi)\|_{L^2(\Omega)} \leq (1 + \Lambda_n)\tau_n$ according to the definition of τ_n. We conclude the proof by bounding τ_n thanks to lemmas 5 and 6. ■

***Remark** 5:* If $(\Lambda_n)_{n\geq 1}$ is a monotonically increasing sequence, then the sequence $(\gamma_n)_{n\geq 1}$ in the GEIM procedure is monotonically decreasing. Using remark 4, the following decay rates in the generalized interpolation error can be derived:

- For any $\varphi \in F$, if $d_n(F, L^2(\Omega)) \leq C_0 n^{-\alpha}$ for any $n \geq 1$, then the interpolation error of the GEIM Greedy selection process can be bounded as $\|\varphi - \mathcal{J}_n[\varphi]\|_{L^2(\Omega)} \leq C_0 2^{3\alpha+1}(1 + \Lambda_n)^3 n^{-\alpha}$.
- For any $\varphi \in F$, if $d_n(F, L^2(\Omega)) \leq C_0 e^{-c_1 n^\alpha}$ for any $n \geq 1$, then the interpolation error of the GEIM Greedy selection process can be bounded as $\|\varphi - \mathcal{J}_n[\varphi]\|_{L^2(\Omega)} \leq C_0 2(1 + \Lambda_n)^3 e^{-c_2 n^\alpha}$.

***Remark** 6:* The evolution of the Lebesgue constant Λ_N as a function of N is a subject of great interest. From the theoretical point of view, crude estimates exist and provide an exponential upper bound that is far from being what we get in the applications. As is shown in ([4], [5], [1]), the growth is lower than linear in N in the EIM situations. Our first numerical experiments with the GEIM reveal cases where it is uniformly bounded when evaluated in the $\mathcal{L}(L^2)$ norm (see [7], [10] for an illustration of this topic as well as for an application of the method to data assimilation coupled with simulation). We do not pretend that this is universal, but it only shows that the theoretical exponentially increasing upper bound is far from being optimal in a class of sets F that have a small Kolmogorov n-width.

V. Conclusion

In this work, it has been proven that the approximation properties of the generalized interpolating spaces X_n lead to an error that has the same qualitative decay as the best possible choice and that is distant by a (multiplicative) factor $(1 + \Lambda_n)\beta_n$ from it. This has been proven in the case of a polynomial or exponential convergence in the $n-$width and is a first step towards the explanation of efficiency of this method in practice (as outlined in [7]).

Acknowledgements

This work was supported in part by the joint research program MANON between CEA-Saclay and University Pierre et Marie Curie-Paris 6 and also by the research grant ApProCEM - FP7-PEOPLE - PIEF-GA-2010-276487.

References

[1] Y. Maday, N. Nguyen, A. Patera, and G. Pau, "A general multipurpose interpolation procedure: the magic points," *Commun. Pure Appl. Anal.*, vol. 8(1), pp. 383–404, 2009.

[2] J. P. Kleijnen and W. C. van Beers, "Robustness of kriging when interpolating in random simulation with heterogeneous variances: Some experiments," *European Journal of Operational Research*, vol. 165, no. 3, pp. 826 – 834, 2005.

[3] H. Liu and S. Maghsoodloo, "Simulation optimization based on taylor kriging and evolutionary algorithm," *Applied Soft Computing*, vol. 11, no. 4, pp. 3451 – 3462, 2011.

[4] M. Barrault, Y. Maday, N. Y. Nguyen, and A. Patera, "An empirical interpolation method: Application to efficient reduced-basis discretization of partial differential equations." *C. R. Acad. Sci. Paris, Série I.*, vol. 339, pp. 667–672, 2004.

[5] M. Grepl, Y. Maday, N. Nguyen, and A. Patera, "Efficient reduced-basis treatment of nonaffine and nonlinear partial differential equations." *M2AN (Math. Model. Numer. Anal.)*, vol. 41(3), pp. 575–605, 2007.

[6] A. Kolmogoroff, "Über die beste Annäherung von Funktionen einer gegebenen Funktionenklasse," *Annals of Mathematics*, vol. 37, pp. 107–110, 1936.

[7] Y. Maday and O. Mula, "A generalized empirical interpolation method: application of reduced basis techniques to data assimilation," *Analysis and Numerics of Partial Differential Equations*, vol. XIII, pp. 221–236, 2013.

[8] M. Bebendorf, Y. Maday, and B. Stamm, "Comparison of some reduced representation approximations," in *Reduced Order Methods for modeling and computational reduction, Proceedings of the CECAM Workshop: Reduced Basis, POD and Reduced Order Methods for model and computational reduction: towards real-time computing and visualization?*, 2013.

[9] T. Bui-Thanh, K. Willcox, and O. Ghattas, "Model reduction for large-scale systems with high-dimensional parametric input space," *SIAM Journal on Scientific Computing*, vol. 30, no. 6, pp. 3270–3288, 2008.

[10] O. Mula, "in progress," Ph.D. dissertation, Univ. Paris VI, 2014.

[11] R. DeVore, G. Petrova, and P. Wojtaszczyk, "Greedy algorithms for reduced bases in banach spaces," *Constructive Approximation*, pp. 1–12, 2012.

Test-size Reduction Using Sparse Factor Analysis

Divyanshu Vats*, Christoph Studer, and Richard G. Baraniuk

Rice University, TX, USA; e-mail: {dvats, studer, richb}@sparfa.com

***Abstract*—Consider a large database of questions that test the knowledge of learners (e.g., students) about a range of different concepts. While the main goal of personalized learning is to obtain accurate estimates of each learner's concept understanding, it is additionally desirable to reduce the number of questions to minimize each learner's workload. In this paper, we propose a novel method to extract a small subset of questions (from a large question database) that still enables the accurate estimation of a learner's concept understanding. Our method builds upon the SPARse Factor Analysis (SPARFA) framework and chooses a subset of questions that minimizes the entropy of the error in estimating the level of concept understanding. We approximate the underlying combinatorial optimization problem using a mixture of convex and greedy methods and demonstrate the efficacy of our approach on real educational data.**

I. Introduction

There has been a recent surge in providing free and high-quality online education through ventures, such as Coursera, Udacity, and edX.[1] Among the key challenges of such systems is in the estimation of each learner's concept understanding. Such information is essential in order to automatically recommend remediation about concepts each learner has weak knowledge of (see, e.g., [6] for the details). In practice, accurate estimates for each learner's concept understanding can be extracted automatically by analyzing responses to large sets of questions about the concepts underlying the given class. To minimize each learner's workload, however, it is of paramount importance to reduce the test-size (compared to the size of the entire question database), while still enabling accurate estimates of each learner's concept understanding. We refer to this problem as *test-size reduction* (TeSR).

In this paper, we propose a novel algorithm for test-size reduction (TeSR), i.e., the problem of selecting a small number of questions from a large dataset, while enabling the accurate estimation of conceptual understanding of each learner. Our approach builds upon the *SPARse Factor Analysis* (SPARFA) framework proposed in [6] to automatically estimate the latent concepts associated with each question. Then, using theory of maximum likelihood (ML) estimators, we formulate the TeSR problem as a combinatorial optimization problem that minimizes the entropy of the asymptotic error in estimating the concept understanding of each learner. We show how the optimization problem can be solved approximately using a combination of convex and greedy methods. We then highlight the advantages of the proposed method by carrying out an experiment with real educational data.

*Also affiliated with the Institute for Mathematics and its Applications, University of Minnesota - Twin Cities, USA.

[1] https://www.coursera.org ; https://www.udacity.com ; https://www.edx.org

Prior work on selecting a subset of questions mainly use statistical models that rely on a single parameter that captures the concept understanding of a learner [3]. In contrast, the SPARFA model used in this work assumes that there are multiple concepts involved in a database of questions. This scenario is more realistic in practice, since it is often the case that questions test knowledge from multiple concepts simultaneously. Several authors have considered the problem of selecting questions in an adaptive manner, see, e.g., [2], [7]. All these adaptive algorithms require a set of starting questions to gauge the adaptive process. Our proposed method can be used for this purpose and is designed to minimize the error of the initial concept understanding estimates, which eventually improves the performance of adaptive methods. We finally note that the problem of selecting questions is related to the problem of sensor selection [5]. The main difference is that the data in sensor network problems is typically real valued, whereas the SPARFA model focuses on binary-valued measurements (i.e., right and wrong answers to questions).

II. Problem Formulation

We begin by reviewing the SPARFA model [6] for extracting relationships between questions and concepts from graded question responses. We then detail the TeSR problem to select "good" subsets of questions for concept estimation.

A. The SPARFA Framework in a Nutshell

Suppose we have a total of Q questions that test knowledge from K concepts. For example, in a signal processing course, questions can test knowledge on concepts like convolution, the sampling theorem, or the Fourier transform. For each question $i = 1, \dots, Q$, let $\mathbf{w}_i \in \mathbb{R}^{K \times 1}$ be a column vector that represents the association of question i to all concepts. Note that a question can test knowledge from multiple concepts. For example, a question on the convolution theorem (i.e., the Fourier transform of a convolution is the product of Fourier transforms of the two signals to be convoluted) in signal processing may test the learner's knowledge on both convolution and the Fourier transform.

The j^{th} entry in $\mathbf{w}_i$, which we denote by w_{ij}, measures the association of question i to concept j. In other words, if question i does not test any knowledge from concept j, then $w_{ij} = 0$. Let $\mathbf{W} = [\mathbf{w}_1, \dots, \mathbf{w}_Q]^T$ be a sparse $Q \times K$ matrix with non-negative entries so that each question only tests a subset of all concepts. Let $\mu_i \in \mathbb{R}$ be a scalar that represents the intrinsic difficulty of a question. A larger (smaller) μ_i corresponds to an easier (harder) question. Let $\boldsymbol{\mu} = [\mu_1, \dots, \mu_Q]^T$ be a $Q \times 1$ column vector that represents the difficulty of

Notation	Description
$\mathbf{W}$	A *sparse* non-negative matrix that characterizes the relationship between questions and knowledge concepts
$\boldsymbol{\mu}$	A vector that specifies the intrinsic difficulty of each question
$\mathbf{c}^*$	A vector that represents a learner's concept knowledge

TABLE I
MAIN PARAMETERS OF THE SPARFA MODEL.

each question. Finally, let $\mathbf{c}^* \in \mathbb{R}^K$ be a column vector that represents the concept understanding of a particular learner. It is this parameter vector that personalized learning systems are naturally interested in estimating accurately.

To model the interaction between $\mathbf{W}$, $\boldsymbol{\mu}$, and $\mathbf{c}^*$, we use the SPARFA framework proposed in [6]. Let Y_i be a binary random variable that indicates whether question i has been answered correctly or not, indicated by 1 and 0, respectively. More specifically, we assume that $Y_i \in \{0,1\}$ admits the following distribution:

$$\Pr(Y_i = 1 \,|\, \mathbf{w}_i, \mu_i, \mathbf{c}^*) = \Phi(\mathbf{w}_i^T \mathbf{c}^* + \mu_i)\,, \qquad (1)$$

where $\Phi(\cdot)$ is an appropriate link function. In this paper, we consider the logistic link function, i.e., $\Phi(x) = 1/(1+e^{-x})$. Assuming that all the random variables $Y_1, \ldots, Y_Q$ are independent of each other, the joint probability distribution of the random vector $\mathcal{Y} = [Y_1, \ldots, Y_Q]^T$ can be written as

$$\Pr(\mathcal{Y} = \mathbf{y} \,|\, \mathbf{W}, \boldsymbol{\mu}, \mathbf{c}^*) = \prod_{i=1}^{Q} \frac{\exp(y_i(\mathbf{w}_i^T \mathbf{c}^* + \mu_i))}{1 + \exp(\mathbf{w}_i^T \mathbf{c}^* + \mu_i)}\,, \qquad (2)$$

where $\mathbf{y} = [y_1, \ldots, y_Q]^T \in \{0,1\}^Q$ is the response of a learner to all the questions. Given graded question responses from multiple learners, the problem of computing the factors $\mathbf{W}$, $\boldsymbol{\mu}$, and the concept understanding vector for each learner can be solved using either the SPARFA-M or SPARFA-B algorithm proposed in [6].

B. Problem Statement: Test-size Reduction (TeSR)

The problem we consider here is to select an appropriate subset of $q < Q$ questions so that $\mathbf{c}^*$, the unknown concept understanding vector of a learner, can be estimated accurately. We assume that prior data, a binary-valued matrix $\widetilde{\mathbf{Y}}$, is known such that an entry $\widetilde{\mathbf{Y}}_{i,j}$ refers to whether a learner j answered question i correct or incorrect. This data matrix can be easily obtained in real educational settings by looking at past offerings of a course, for example. As mentioned in Section II-A, we can compute $\mathbf{W}$ for all the Q questions in the database using $\widetilde{\mathbf{Y}}$.

Suppose, hypothetically, that we choose a subset $\mathcal{I}$ of $q < Q$ questions and we are given a response vector $\mathbf{y}_\mathcal{I}$. Using the model in (2), the maximum likelihood (ML) estimate $\widehat{\mathbf{c}}$ can be computed as follows:

$$\begin{aligned}\widehat{\mathbf{c}} &= \arg\max_{\mathbf{c}\in\mathbb{R}^K} \log \Pr(\mathcal{Y}_\mathcal{I} = \mathbf{y}_\mathcal{I} \,|\, \mathbf{W}, \boldsymbol{\mu}, \mathbf{c}) \\ &= \arg\max_{\mathbf{c}\in\mathbb{R}^K} \sum_{i\in\mathcal{I}} \left[y_i(\mathbf{w}_i^T \mathbf{c} + \mu_i) - \log(1 + e^{\mathbf{w}_i^T \mathbf{c} + \mu_i}) \right]. \quad (3)\end{aligned}$$

Given $\mathbf{y}_\mathcal{I}$, the result of (3) can be solved via standard convex optimization algorithms [1]. Our main objective is to find a subset $\mathcal{I}$ so that $|\mathcal{I}| = q$ and the ML estimate $\widehat{\mathbf{c}}$ is as close to the ground truth $\mathbf{c}^*$ as possible. To do this, we make use of the following asymptotic normality property of ML estimators (see, e.g., [4] for the details). First, define the Fisher information matrix as follows:

$$\mathbf{F}(\mathbf{W}_\mathcal{I}, \boldsymbol{\mu}_\mathcal{I}, \mathbf{c}^*)) = \sum_{i\in\mathcal{I}} \frac{\exp(\mathbf{w}_i^T \mathbf{c}^* + \mu_i)}{(1 + \exp(\mathbf{w}^T \mathbf{c}^* + \mu_i))^2} \mathbf{w}_i \mathbf{w}_i^T\,, \quad (4)$$

where the notation $\mathbf{W}_\mathcal{I}$ refers to the rows of $\mathbf{W}$ indexed by $\mathcal{I}$ and $\boldsymbol{\mu}_\mathcal{I}$ refers to the entries in $\boldsymbol{\mu}$ indexed by $\mathcal{I}$.

Theorem II.1. *Let $\mathcal{I}_r$ for $r = 1, \ldots, q$ be a fixed sequence of q subsets of size r. Assume that there exists a $q_0 < q$ such that $\mathbf{F}(\mathbf{W}_{\mathcal{I}_q}, \boldsymbol{\mu}_{\mathcal{I}_q}, \mathbf{c}^*))$ is invertible for all $r > q_0$. Then, the random vector $\sqrt{q}(\widehat{\mathbf{c}} - \mathbf{c}^*)$ converges in distribution to a multivariate normal vector with mean zero and covariance $\mathbf{F}(\mathbf{W}_{\mathcal{I}_q}, \boldsymbol{\mu}_{\mathcal{I}_q}, \mathbf{c}^*))^{-1}$, i.e., $\sqrt{q}(\widehat{\mathbf{c}} - \mathbf{c}^*) \xrightarrow{d} \mathcal{N}(0, \mathbf{F}(\mathbf{W}_{\mathcal{I}_q}, \boldsymbol{\mu}_{\mathcal{I}_q}, \mathbf{c}^*))^{-1})$.*

Theorem II.1 states that as the number of questions q gets large, the covariance of the error $\sqrt{q}(\widehat{\mathbf{c}} - \mathbf{c}^*)$ can be approximated by the inverse of the Fisher information matrix. This motivates a natural strategy to choose a subset of questions $\mathcal{I}$ that minimizes the differential entropy[2] of a multivariate normal random vector with mean zero and covariance $\mathbf{F}(\mathbf{W}_\mathcal{I}, \boldsymbol{\mu}_\mathcal{I}, \mathbf{c}^*))^{-1}$, which intuitively minimizes the uncertainty in the error $\sqrt{q}(\widehat{\mathbf{c}} - \mathbf{c}^*)$. Consequently, the optimization problem considered in the remainder of the paper, referred to as the *test-size reduction* (TeSR) problem, corresponds to

$$\text{(TeSR)} \quad \widehat{\mathcal{I}} = \arg\max_{\mathcal{I}\subset\{1,\ldots,Q\}, |\mathcal{I}|=q} \log\det(\mathbf{F}(\mathbf{W}_\mathcal{I}, \boldsymbol{\mu}_\mathcal{I}, \mathbf{c}^*))\,.$$

The main challenges in solving (TeSR) are (i) the TeSR problem is a combinatorial optimization problem and (ii) the concept knowledge vector $\mathbf{c}^*$ is *unknown*, so the objective function cannot be evaluated exactly. In the next section, we outline a data-driven approach for approximating the (TeSR) objective function. We then develop a computationally efficient algorithm that delivers good approximations to the combinatorial TeSR problem.

III. TEST-SIZE REDUCTION ALGORITHM

We start by noting that the scalar term in the summation in (4) is equivalent to the variance of the random variable Y_i

[2]Note that the differential entropy of $X = (X_1, \ldots, X_q) \sim \mathcal{N}(0, \Sigma)$ is given by $\log\left((2\pi e)^q \det(\Sigma)\right)$.

Algorithm 1: Nonadaptive test-size reduction (NA-TeSR)

Step 1) First choose K questions by solving

$$\widehat{\mathcal{I}}_{[K]} = \underset{\mathcal{I}\subset\{1,\ldots,Q\},|\mathcal{I}|=K}{\arg\max} \log\det\left(\mathbf{W}_{\mathcal{I}}^T\widehat{\mathbf{V}}\mathbf{W}_{\mathcal{I}}\right) \quad (5)$$

using the convex optimization method in (8). The entries of the diagonal matrix $\widehat{\mathbf{V}}$ are defined as $\widehat{\mathbf{V}}_{kk} = \widehat{v}_k$, where $\bar{v}_k$ specified in (6).

Step 2) Select questions $K+1,\ldots,q$ in a greedy manner:

$$\widehat{\mathcal{I}}_{j+1} = \underset{i\in\{1,\ldots,Q\}\setminus\widehat{\mathcal{I}}_{[j]}}{\arg\max}\ \widehat{v}_i\mathbf{w}_i^T\left(\mathbf{W}_{\mathcal{I}_{[j]}}^T\widehat{\mathbf{V}}_{\mathcal{I}_{[j]}}\mathbf{W}_{\mathcal{I}_{[j]}}\right)^{-1}\mathbf{w}_i\,.$$

conditioned on $\mathbf{c}^*$, i.e.,

$$\mathbb{V}\mathrm{ar}[Y_i|\mathbf{c}^*] = \frac{\exp(\mathbf{w}_i^T\mathbf{c}^*+\mu_i)}{(1+\exp(\mathbf{w}^T\mathbf{c}^*+\mu_i))^2}. \quad (6)$$

The variance $\mathbb{V}\mathrm{ar}[Y_i|\mathbf{c}^*]$ captures the variability of a learner in answering the i^{th} question. By defining $\mathbf{V}$ as a $Q\times Q$ diagonal matrix with entries $\mathbf{V}_{ii} = \mathbb{V}\mathrm{ar}[Y_i|\mathbf{c}^*]$, the TeSR problem can be rewritten in matrix form as

$$\text{(TeSR)}\quad \widehat{\mathcal{I}} = \underset{\mathcal{I}\subset\{1,\ldots,Q\},|\mathcal{I}|=q}{\arg\max} \log\det(\mathbf{W}_{\mathcal{I}}^T\mathbf{V}_{\mathcal{I}}\mathbf{W}_{\mathcal{I}})\,.$$

We first address the problem of approximating the objective function using a graded question response matrix $\widetilde{\mathbf{Y}}$ acquired in, e.g., a previous offering of a course. Since the vector $\mathbf{c}^*$ is not known, we need to make some assumptions on $\widetilde{\mathbf{Y}}$ so that the objective function can be estimated. As it turns out, a natural, and convenient, assumption is for the prior data to be chosen in such a way that the concept understanding of the learners in the response matrix $\widehat{\mathbf{Y}}$ is roughly equal to $\mathbf{c}^*$. Using this assumption, we can easily estimate $\mathbb{V}\mathrm{ar}[Y_i|\mathbf{c}^*]$ to be the sample variance of the data $\widetilde{\mathbf{Y}}$:

$$\widehat{v}_i = \mathbb{V}\mathrm{ar}[Y_i|\mathbf{c}^*] = \frac{1}{N}\sum_{j=1}^{N}\left(\widetilde{Y}_{ij} - \frac{1}{N}\sum_{j=1}^{N}\widetilde{Y}_{ij}\right)\,, \quad (7)$$

where $\widetilde{Y}_{ij}$ is the $(i,j)^{\text{th}}$ entry of $\widetilde{\mathbf{Y}}$. Using the sample variance, (TeSR) can be rewritten as

$$\text{(TeSR)}\quad \widehat{\mathcal{I}} = \underset{\mathcal{I}\subset\{1,\ldots,Q\},|\mathcal{I}|=q}{\arg\max} \log\det(\mathbf{W}_{\mathcal{I}}^T\widehat{\mathbf{V}}_{\mathcal{I}}\mathbf{W}_{\mathcal{I}})\,,$$

where $\widehat{\mathbf{V}}$ is a diagonal matrix with entries $\widehat{\mathbf{V}}_{kk} = \widehat{v}_k$. In the above formulation, there is no longer any dependence on $\mathbf{c}^*$.

Algorithm 1 summarizes a nonadaptive method for solving the TeSR problem. The first step is to find the "best" K questions, where K is the number of concepts in the Q questions. Next, we select the remaining questions $K+1,\ldots,q$ in an iterative manner. Note that selecting less than K questions would inhibit estimating the K-dimensional concept knowledge vector.

For any subset $\mathcal{I}$, let $\mathcal{I}_{[K]}$ denote the first K elements. To select the initial K questions $\widehat{\mathcal{I}}_{[K]}$, we use methods in [5] to formulate the combinatorial optimization problem in (5) as a convex optimization problem. More specifically, we can obtain an approximate solution to (5) by solving the following convex optimization problem:

$$\begin{aligned} \text{maximize} \quad & \log\det\left(\mathbf{W}_{\mathcal{I}}^T\widehat{\mathbf{V}}\mathbf{Z}\mathbf{W}_{\mathcal{I}}\right) \\ \text{subject to} \quad & \text{diagonal matrix } \mathbf{Z} \text{ with } Z_{kk} = z_k \\ & \textstyle\sum z_k = K \text{ and } 0\le z_k\le 1 \end{aligned} \quad (8)$$

Once (8) has been computed, $\widehat{\mathcal{I}}_{[K]}$ can be approximated as the indices corresponding to the top K largest values of the diagonal elements $Z_{kk} = z_k$ of the matrix $\mathbf{Z}$.

The second step in Algorithm 1 chooses the remaining $q-K$ questions in a greedy manner. Using the identity

$$\det(\mathbf{X}+\mathbf{b}\mathbf{b}^T) = \det(\mathbf{X})(1+\mathbf{b}^T\mathbf{X}^{-1}\mathbf{b}),$$

where $\mathbf{X}$ is a square matrix and $\mathbf{b}$ is a column vector, the quantity $\log\det(\mathbf{W}_{\mathcal{I}_{[j+1]}}^T\widehat{\mathbf{V}}_{\mathcal{I}_{[j+1]}}\mathbf{W}_{\mathcal{I}_{[j+1]}})$ can be rewritten as

$$\log\det(\mathbf{W}_{\mathcal{I}_{[j]}}^T\widehat{\mathbf{V}}_{\mathcal{I}_{[j]}}\mathbf{W}_{\mathcal{I}_{[j]}}) + \log(1+F) \quad (9)$$

with the definition

$$F = \widehat{\mathbf{V}}_{\mathcal{I}_{j+1},\mathcal{I}_{j+1}}\mathbf{w}_{\mathcal{I}_{j+1}}^T(\mathbf{W}_{\mathcal{I}_{[j]}}^T\widehat{\mathbf{V}}_{\mathcal{I}_{[j]}}\mathbf{W}_{\mathcal{I}_{[j]}})^{-1}\mathbf{w}_{\mathcal{I}_{j+1}}. \quad (10)$$

Thus, once j questions $\widehat{\mathcal{I}}_{[j]}$ have been selected, the next question, $\widehat{\mathcal{I}}_{j+1}$, can be selected so that the quantity F defined above is maximized.

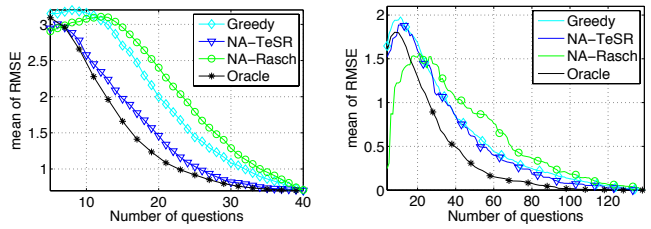
Remark 1: The computational complexity of Step 1 of Algorithm 1 is rather low when using the convex optimization relaxation approach outlined in (TeSR). We refer to [5] for iterative methods that solve (8). We note that although Step 2 requires computing an inverse of a $K\times K$ matrix multiple times, this inverse can be computed recursively once $(\mathbf{W}_{\mathcal{I}_{[K]}}^T\widehat{\mathbf{V}}\mathbf{W}_{\mathcal{I}_{[K]}})^{-1}$ has been computed. Finally, we can directly solve (TeSR) using the convex relaxation in (8). However, the computational complexity of this approach can be large, especially when q is large.

Remark 2: Note that when $\mathbf{W}$ is a $Q\times 1$ vector of all ones, the SPARFA model reduces to the Rasch model [9]. In this case, (TeSR) reduces to a problem of maximizing the sum of the variance terms over the selected questions. Thus, all the questions can be selected independently of the others when using the Rasch model. On the other hand, when using SPARFA, since we account for the statistical dependencies amongst questions, the questions can no longer be chosen independently as it is evident from Algorithm 1.

IV. Experimental Results

In this section, we assess the performance of our algorithms for test-size reduction (TeSR) using synthetic and real educational datasets.

Baseline algorithms: We compare NA-TeSR to three baseline algorithms. The first, referred to as NA-Rasch, uses the Rasch model [9] and selects questions in a non-adaptive manner



(a) Synthetic Data (b) Real Data

Fig. 1. TeSR and baseline methods for synthetic data and real data.

(see Remark 2). The second, referred to as Greedy, iteratively selects a question from each concept until the required number of q questions has been selected. If all questions from a given concept have been exhausted, then Greedy skips to the next concept to select a question. Note that this approach completely ignores the variability of a learner in answering various questions. Finally, we also compare to an oracle algorithm, referred to as Oracle, that uses the true underlying (but in practice unknown) vector $\mathbf{c}^*$ to solve the TeSR problem. Note that the oracle algorithm is not practical and is only used to characterize the performance limits of TeSR.

Performance measure: We assess the performance of the algorithms using the root mean-square error (RMSE), defined as $\mathsf{RMSE} = \|\hat{\mathbf{c}} - \mathbf{c}^*\|_2$, where $\hat{\mathbf{c}}$ is the estimate delivered by each method and $\mathbf{c}^*$ is the ground truth. Although $\mathbf{c}^*$ is known for synthetic experiments, for real data, we assume that the ground truth is the concept vector estimated when asking all Q available questions.

Methods: In the experiments shown next, we assume that a matrix $\mathbf{Y}$ is given that contains graded responses of Q questions from M students. As mentioned in Section 2, for real data, we use SPARFA-M [6] to estimate $\mathbf{W}$ and the ground truth concept values of each learner. For each learner, we apply the baseline and our proposed TeSR algorithms using $\mathbf{W}$ and a training data $\widetilde{\mathbf{Y}}$ obtained after removing the responses of the learner from the matrix $\mathbf{Y}$. To show the performance of our TeSR algorithms, we report the mean and standard deviation of the RMSE evaluated over all M learners.

MLE convergence: It turns out that the maximum likelihood estimate (MLE) may not converge for certain patterns of the response vectors. In the case of inexistent ML estimates, we make use of the sign of the ML estimates to compute the RMSE. We then assign each entry in $\hat{\mathbf{c}}$ to the worst (for $-\infty$) or best (for $+\infty$) value obtained from a prior set of learners who have taken the course. In our simulations, these worst and best concept values are computed using the training data $\widetilde{\mathbf{Y}}$.

Results: We generated a sparse 50×5 matrix $\mathbf{W}$ that maps 50 questions to 5 concepts. There were roughly 30% non-zero entries in $\mathbf{W}$ with the non-zero entries chosen from an exponential random variable with parameter $\lambda = 2/3$. Each entry in the intrinsic difficulty vector $\boldsymbol{\mu}$ was generated from a standard normal distribution. We assumed 25 learners whose concept understanding vectors were again generated from a standard normal distribution. For each $\mathbf{Y}$, we computed the reduced test-size with $q = 5, 6, \ldots, 44$. Fig. 1 shows the mean value of the RMSE over 100 randomly generated response vectors $\mathbf{Y}$. Note that the mean RMSE is taken over all 25 learners. We observe that NA-TeSR is superior to all practical baseline algorithms. This observation suggests that the Rasch model is not an appropriate model for selecting questions for the purpose of test-size reduction in courses having more than one underlying concept.

Fig. 1(b) shows results on real educational dataset corresponding to graded response data obtained from the ASSISTment system [8]. The original data contained responses from 4354 learners on 240 questions. There is a large number of missing responses in this dataset, i.e., not every learner answered all problems. In order to get a dataset with a sufficient number of observed entries, we focused on a subset of 219 questions answered by 403 learners. The resulting trimmed $\mathbf{Y}$ matrix has roughly 75% missing values. Fig. **??**(b) shows the associated results and we observe similar trends as for synthetic data set. The main difference is that the performance of the Greedy algorithm is almost as good as the NA-TeSR algorithm in certain domains. This may be a result of the several missing values present in the dataset that does not allow for accurate computations of the variability in answering each question. We note that we have extended the NA-TeSR algorithm in [10] to an adaptive algorithm where each question selected by the greedy step in NA-TeSR uses prior responses to form an estimate of $\widehat{\mathbf{c}}$. This method leads to results that are closer to the Oracle algorithm. We refer to [10] for more details.

References

[1] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[2] S. Buyske. *Applied optimal designs*, chapter Optimal design in educational testing, pages 1–16. John Wiley & Sons Inc, 2005.

[3] H. Chang and Z. Ying. Nonlinear sequential designs for logistic item response theory models with applications to computerized adaptive tests. *The Annals of Statistics*, 37(3):1466–1488, Jun. 2009.

[4] L. Fahrmeir and H. Kaufmann. Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics*, 13(1):342–368, 1985.

[5] S. Joshi and S. Boyd. Sensor selection via convex optimization. *IEEE Transactions on Signal Processing*, 57(2):451–462, 2009.

[6] A. S. Lan, A. E. Waters, C. Studer, and R. G. Baraniuk. Sparse factor analysis for learning and content analytics. *Journal of Machine Learning Research*, Nov. 2012, submitted.

[7] W. J. Linden and P. J. Pashley. *Elements of adaptive testing*, chapter Item selection and ability estimation in adaptive testing, pages 3–30. Springer, 2010.

[8] Z. Pardos and N. Heffernan. Modeling individualization in a Bayesian networks implementation of knowledge tracing. *User Modeling, Adaptation, and Personalization*, pages 255–266, 2010.

[9] G. Rasch. *Probabilistic Models for Some Intelligence and Attainnment Tests*. Studies in mathematical psychology. Danmarks paedagogiske Institut, 1960.

[10] D. Vats, C. Studer, A. Lan, L. Carin, and R. Baraniuk. Test-size reduction for concept estimation. In *International Conference on Educational Data Mining (EDM)*, 2013.

Special Frames

Luis Daniel Abreu
Acoustic Research Institute
Austrian Academy of Sciences
Email: daniel@mat.uc.pt

***Abstract*—Three classes of special frames are presented: special Fourier-type frames, special Gabor frames and special wavelet frames. Known information about density of Fourier-Bessel frames, Gabor (super)frames with Hermite functions and wavelet (super)frames with Laguerre functions will be outlined.**

I. Introduction

Given a Hilbert space $\mathcal{H}$, a vector $g \in \mathcal{H}$ and a family of operators $\{\pi_\lambda g\}_{\lambda\in\Lambda}$, *the special frame problem* consists of the following question:

- *What conditions one should impose on a discrete set Λ, such that $\{\pi_\lambda g\}_{\lambda\in\Lambda}$ is a frame for $\mathcal{H}$?*

More precisely, we want to find constants $A, B > 0$ such that, for every $f \in \mathcal{H}$,

$$A\|f\|^2 \leq \sum_{\lambda\in\Lambda} |\langle f, \pi_\lambda g\rangle_{\mathcal{H}}|^2 \leq B\|f\|^2. \qquad (1)$$

The term *special* refers to a *viewpoint*: rather than looking at general properties of frames, we want to know detailed information about a *specific example* of a frame with particular interesting structure. We will consider three classes of frames.

1) Fourier Frames: $\mathcal{H} = L^2(-\pi,\pi)$, $g(x) = e^{ix}$ and $\pi_\lambda g(x) = g(\lambda x)$. For g other that e^{ix} we will talk about *Fourier-type frames*.
2) Gabor frames: $\mathcal{H} = L^2(R)$ and $\pi_{\lambda=(\lambda_1,\lambda_2)}g(x) = e^{2\pi i\lambda_2 t}g(t-\lambda_1)$. Several choices of g are possible.
3) Wavelet frames (positive frequencies): $\mathcal{H} = H^2(C^+)$ and $\pi_\lambda g(t) = \lambda_1^{-\frac{1}{2}} g(\lambda_1^{-1}(t-\lambda_2)), \quad t \in R$. Several choices of g are possible.

II. Special Fourier-type frames

While the *Fourier orthogonal basis* is of the form $\{e^{ikx}\}_{k\in Z}$, *Fourier frames* are of the form $\{e^{i\lambda x}\}_{\lambda\in\Lambda}$, allowing the set Λ to be nonuniform and redundant. The orthogonal basis case $\Lambda = Z$ works as a threshold for Fourier frames: we know that frames requires Λ to be "denser than Z" [16]. We can think of Fourier frames as being made out of the special function $f(x) = e^x$. Frames of the form $\{f(\lambda x)\}_{\lambda\in\Lambda}$ will be called *Fourier-type frame*s. To keep intact the rich set up of the Fourier frames we want to be able to transfer our Fourier-type frames to a Paley–Wiener-type space using some Fourier-type transform. Moreover, we are interested in cases displaying a second order differential operator commuting with the respective concentration operators. In the case of Fourier frames, the existence of such an operator is regarded as a "fortunate accident", according to Daubechies exposition in [9, page 22]. In the work of Tracy and Widom about the local statistics of the asymptotics of certain random matrices [19], [18], one can find two more instances where this "fortunate accident" occurs. This motivated our investigation of Fourier-Bessel frames [5] and Airy frames [6]. Let us say a bit more about the results in [5].

Let $J_\alpha(x)$ be the Bessel function of order $\alpha > -1/2$ and $j_{n,\alpha}$ its n^{th} zero. While the *Fourier-Bessel orthogonal basis* is of the form $\{x^{\frac{1}{2}}J_\alpha(j_{n,\alpha}x)\}_{n=0}^{\infty}$, *Fourier-Bessel frames* are of the form $\{(\lambda x)^{\frac{1}{2}}J_\alpha(\lambda x)\}_{\lambda\in\Lambda}$, allowing the set Λ to be nonuniform and redundant. To obtain the definition of a Fourier-Bessel frame, choose in (1) $\mathcal{H} = L^2[0,1]$, $g(x) = (x)^{\frac{1}{2}}J_\alpha(x)$ and $(\pi_\lambda g)(x) = g(\lambda x)$. In [5], we have considered a more general situation than frames and obtained analogues of the Landau conditions [16] for interpolation and sampling. As a particular case we obtain precise necessary density conditions for Fourier Bessel frames. Let $n_a(r)$ denote the number of points of $\Lambda \subset (0,\infty)$ to be found in $[a, a+r]$. Then the lower density of Λ is given by $D^-(\Lambda) = \lim_{r\to\infty} \inf\inf_{a\geq 0} \frac{n_a(r)}{r}$. The main result in [5] is the following Landau-type necessary condition for sampling in spaces of functions $\mathcal{B}_\alpha(S)$ whose Hankel transform (the analogue of the Fourier transform in this context) is supported on a set S of bounded measure:

Theorem [5]: Let S be a measurable subset of $(0,\infty)$ and $\alpha > -1/2$. If a separated set Λ is of sampling for $\mathcal{B}_\alpha(S)$, then

$$D^-(\Lambda) \geq \frac{1}{\pi} m(S). \qquad (2)$$

III. Special Gabor (super)frames

The investigation of special Gabor frames has been a topic of high interest in the last twenty years. See the recent paper [13] and the outline in the Introduction. We can construct Gabor superframes with Hermite functions, which are useful in the multiplexing of non-stationary signals. Consider the Hilbert space $L^2(R, C^n)$ of vector-valued functions $\overrightarrow{\mathbf{f}} = (f_0, ..., f_{n-1})$ together with the inner product $\left\langle \overrightarrow{\mathbf{f}}, \overrightarrow{\mathbf{g}} \right\rangle_{\mathcal{H}} = \sum_{0\leq k\leq n-1} \langle f_k, g_k\rangle_{L^2(R)}$. To obtain the definition of a *Gabor superframe* for the vector valued system $\mathcal{G}(\overrightarrow{\mathbf{g}}, \Lambda) = \{\pi_\lambda \overrightarrow{\mathbf{g}}\}_{\lambda\in\Lambda}$, choose in (1) $\mathcal{H} = L^2(R, C^n)$, $g = \overrightarrow{\mathbf{g}}$ and, given a point $\lambda = (\lambda_1, \lambda_2)$ in R^2, define π_λ as the time-frequency shift $\pi_\lambda g(t) = e^{2\pi i\lambda_2 t}g(t-\lambda_1)$, $t \in R$.

There is a characterization of all lattices generating Gabor superframes with Hermite functions h_n [12], which is equivalent to a sampling problem in a Fock space of polyanalytic functions [1].

Theorem [12] Let $\overrightarrow{\mathbf{h}_n} = (h_0, ..., h_{n-1})$ be the vector of the first n Hermite functions. Then $\mathcal{G}(\overrightarrow{\mathbf{h}_n}, \alpha Z + i\beta Z)$ is a frame for $L^2(R, C^n)$, if and only if $\alpha\beta < \frac{1}{n+1}$.

For a special frame generated by a single Hermite function, the characterization is still an open problem. Nevertheless, some interesting results are known. If $\alpha\beta < \frac{1}{n+1}$ then $\mathcal{G}(h_n, \alpha Z + i\beta Z)$ is a frame [11] but if $\alpha\beta = 1 - \frac{1}{j}$ then $\mathcal{G}(h_1, \alpha Z + i\beta Z)$ is not [14]. Supported by their results and by some numerical evidence, the authors of [14] conjectured that if $\alpha\beta < 1$ and $\alpha\beta \neq 1 - \frac{1}{j}$, then $\mathcal{G}(h_1, \alpha Z + i\beta Z)$ is a frame.

IV. Special Wavelet (super)frames

We can also construct wavelet superframes which are useful in the multiplexing of non-stationary signals of positive frequencies, leading to a sampling problem in certain (Bergman) spaces of polyanalytic functions. We should emphasize again that our viewpoint of wavelet frames is different of those ones documented in [9] and in the more recent monograph [15]. For a vector $\mathbf{g} = (g_1, ..., g_n)$ such that the Fourier transforms of any two functions g_i and g_j are orthogonal in $L^2(R^+, t^{-1})$, define π_z pointwise as $\pi_z \mathbf{g} = (\pi_z g_1, ..., \pi_z g_n)$. To obtain the definition of a *wavelet superframe* for the vector valued system $\mathcal{W}(\overrightarrow{\mathbf{g}}, \Lambda) = \{\pi_\lambda \overrightarrow{\mathbf{g}}\}_{\lambda\in\Lambda}$, let in (1) $\mathcal{H} = H^2(C^+, C^n)$ be the inner product space whose vector components belong to $H^2(C^+)$, the standard Hardy space of the upper half-plane, $g = \overrightarrow{\mathbf{g}}$ and, given a point $\lambda = (\lambda_1, \lambda_2)$ in R^2, define π_λ as the time-scale shift $\pi_\lambda g(t) = \lambda_1^{-\frac{1}{2}} g(\lambda_1^{-1}(t - \lambda_2)), \quad t \in R$.

We consider wavelet superframes with analyzing wavelets $\overrightarrow{\Phi_n^\alpha} = (\frac{\Phi_0^\alpha}{c_{\Phi_0^\alpha}}, ..., \frac{\Phi_n^\alpha}{c_{\Phi_n^\alpha}})$, where $c^2_{\Phi_n^\alpha} = \frac{\Gamma(n+\alpha+1)}{n!}$ is the admissibility constant of the vector component Φ_n^α defined via its Fourier transform as

$$\mathcal{F}\Phi_n^\alpha(t) = t^{\frac{1}{2}} l_n^\alpha(2t), \quad with \quad l_n^\alpha(t) = t^{\frac{\alpha}{2}} e^{-\frac{t}{2}} L_n^\alpha(t), \qquad (3)$$

where $L_n^\alpha(t)$ is the standard notation for the Laguerre polynomial.

The problem of, given a wavelet g, to characterize the sets of points Λ such that $\mathcal{W}(g, \Lambda)$ is a wavelet frame (and the corresponding problem for the superframes defined above), is more difficult than the corresponding one for Gabor frames. The only characterization known so far concerns the case $n = 0$ in (3). In this case, the problem can be reduced to the density of sampling in the Bergman spaces, which has been completely understood in [17]. An important research problem is to understand how Seip's results extend to the whole family $\{\Phi_n^\alpha\}$. The only thing known to the present date is a necessary condition obtained in [2] in terms of a set of points known as the "hyperbolic lattice" $\Gamma(a, b) = \{a^m bk, a^m\}_{k,m\in Z}$. The quantity $b \log a$ replaces the time-frequency $\alpha\beta$ for purposes of measuring frame density.

Theorem [2]: If $\mathcal{W}(\Phi_n^{2\alpha-1}, \Gamma(a, b))$ is a wavelet frame for $H^2(C^+)$, then $b \log a < 2\pi \frac{n+1}{\alpha}$.

Using the polyanalytic structure of the underlying Bergman spaces [3] one can also prove a result which shows that it is necessary to oversample by a rate of n to obtain superframes. This matches what one would expect from [10].

Theorem [4]: If $\mathcal{W}(\overrightarrow{\Phi_n^{2\alpha-1}}, \Gamma(a, b))$ is a wavelet superframe for $H^2(C^+, C^n)$, then $b \log a < \frac{2\pi}{n+\alpha}$.

Actually in [4] we obtain a much stronger result using Seip´s density [17], as part of our sampling results in polyanalytic Bergman spaces.

Aknowledgements: The author would like to thank to the three referees for their remarks, corrections and suggestions which help to improve the presentation of the manuscript and to Diana Stoeva for her reading of the final version. Supported by European program COMPETE/FEDER via FCT project PTDC/MAT/114394/2009, by Austrian Science Foundation (FWF) project "Frames and Harmonic Analysis" and START-project FLAME ('Frames and Linear Operators for Acoustical Modeling and Parameter Estimation').

References

[1] L. D. Abreu, *Sampling and interpolation in Bargmann-Fock spaces of polyanalytic functions*, Appl. Comp. Harm. Anal., 29 (2010), 287-302.
[2] L. D. Abreu, *Wavelet frames with Laguerre functions,* C. R. Acad. Sci. Paris, Ser. I, 349 (2011), 255-258.
[3] L. D. Abreu, *Super-wavelets versus poly-Bergman spaces,* Int. Eq. Oper. Theory, 73 (2012), 177–193.
[4] L. D. Abreu, *Wavelets (super)frames with Laguerre functions and sampling in polyanalytic spaces*, manuscript.
[5] L. D. Abreu, A. Bandeira, *Landau's necessary density conditions for the Hankel transform*, J. Funct. Anal. 162 (2012), 1845-1866.
[6] L. D. Abreu, A. Bandeira, *Landau's necessary density conditions for the Airy transform*, work in progress.
[7] L. D. Abreu, K. Gröchenig, *Banach Gabor frames with Hermite functions: polyanalytic spaces from the Heisenberg group*, Appl. Anal., 91 (2012), 1981-1997.
[8] L. D. Abreu, H. G. Feichtinger, *Function spaces of polyanalytic functions*, HCAA special volume, 32pp, to appear.
[9] I. Daubechies, *"Ten Lectures On Wavelets"*, CBMS-NSF Regional conference series in applied mathematics, (1992).
[10] D. E. Dutkay, P. Jorgensen, *Oversampling generates super-wavelets.* Proc. Amer. Math. Soc. 135 (2007), 2219–2227.
[11] K. Gröchenig, Y. Lyubarskii, *Gabor frames with Hermite functions*, C. R. Acad. Sci. Paris, Ser. I 344 (2007), 157-162.
[12] K. Gröchenig, Y. Lyubarskii, *Gabor (Super)Frames with Hermite Functions*, Math. Ann., 345 (2009), 267-286.
[13] K. Gröchenig, J. Stoeckler, *Gabor frames and totally positive functions,* Duke Math. J., to appear.
[14] Y. Lyubarskii, P. G. Nes, *Gabor frames with rational density,* Appl. Comp. Harm. Anal., 34 (2013), 488494.
[15] G. Kuttyniok, *Affine Density in Wavelet Analysis,* Lecture Notes in Mathematics 1914 , Springer-Verlag, Berlin, 2007.
[16] H. J. Landau: *Necessary Density Conditions for Sampling and Interpolation of Certain Entire Functions*, Acta Math., 117 (1967), 37-52.
[17] K. Seip, *Beurling type density theorems in the unit disc*, Invent. Math., 113 (1993), 21-39.
[18] C. A. Tracy, H. Widom, *Level-spacing distributions and the Airy kernel*, Commun. Math. Phys. 159 (1994), 151-174.
[19] C. A. Tracy, H. Widom, *Level-spacing distributions and the Bessel kernel*, Commun. Math. Phys. 161 (1994), 289-309.

Variation and approximation for Mellin-type operators

Laura Angeloni
Dipartimento di Matematica e Informatica
Università degli Studi di Perugia
Via Vanvitelli 1, 06123, Perugia (Italy)
Email: angeloni@dmi.unipg.it

Gianluca Vinti
Dipartimento di Matematica e Informatica
Università degli Studi di Perugia
Via Vanvitelli 1, 06123, Perugia (Italy)
Email: mategian@unipg.it

***Abstract*—Mellin analysis is of extreme importance in approximation theory, also for its wide applications: among them, for example, it is connected with problems of Signal Analysis, such as the Exponential Sampling. Here we study a family of Mellin-type integral operators defined as**

$$(T_w f)(\mathtt{s}) = \int_{\mathbb{R}^N_+} K_w(\mathtt{t}) f(\mathtt{st}) \frac{d\mathtt{t}}{\langle \mathtt{t} \rangle}, \ \mathtt{s} \in \mathbb{R}^N_+, \ w > 0, \qquad \text{(I)}$$

where $\{K_w\}_{w>0}$ are (essentially) bounded approximate identities, $\langle \mathtt{t} \rangle := \prod_{i=1}^N t_i$, $\mathtt{t} = (t_1, \ldots, t_N) \in \mathbb{R}^N_+$, and $f : \mathbb{R}^N_+ \to \mathbb{R}$ is a function of bounded $\varphi-$variation. We use a new concept of multidimensional $\varphi-$variation inspired by the Tonelli approach, which preserves some of the main properties of the classical variation. For the family of operators (I), besides several estimates and a result of approximation for the $\varphi-$modulus of smoothness, the main convergence result that we obtain proves that

$$\lim_{w \to +\infty} V^{\varphi}[\mu(T_w f - f)] = 0,$$

for some $\mu > 0$, provided that f is $\varphi-$absolutely continuous. Moreover, the problem of the rate of approximation is studied, taking also into consideration the particular case of Fejér-type kernels.

I. Introduction

An important topic in approximation theory is the study of convergence of classes of integral operators in the frame of $BV-$spaces, namely spaces of functions of bounded variation. This problem was faced in the literature from several points of view, using different families of operators and different notions of variation, such as the classical variation ([4]), the distributional variation ([7]), the Cesari variation ([16]) or the Musielak-Orlicz $\varphi-$variation ([26], [15], [24], [28], [13], [17], [5]). An important direction of this research is the multidimensional case, in particular in view of the application of such results in several fields, such as image reconstruction. Results in this sense can be found, for example, in [10], [4] in the case of Tonelli variation and in [6], where the authors introduce a new multidimensional concept of $\varphi-$variation and give approximation results for functions of bounded $\varphi-$variation by means of the classical convolution integral operators. The nonlinear case was explored in [3].

An interesting development of the theory is the case of Mellin-type integral operators. Mellin operators are well known and widely used in approximation theory (see, e.g., [23], [19]), also because of their important applications in various fields, for example in Signal Processing. Indeed, Mellin analysis is strictly connected to Signal Analysis, in particular to the Exponential Sampling. A seminal paper in this sense is [20], where the authors establish a Sampling Theorem in which the samples are not equally spaced, as in the classical Shannon Sampling Theorem, but exponentially spaced, by means of Mellin transform methods. This theory has important applications, for example in optical physics and engineering (see, e.g., [22], [18]), in problems in which information accumulates near time $t = 0$. With this respect, to develop a theory about Mellin-type operators becomes useful and interesting. Results in this sense can be also found, for example, in [11], [12].

Here we consider a family of Mellin-type integral operators of the form

$$(T_w f)(\mathtt{s}) = \int_{\mathbb{R}^N_+} K_w(\mathtt{t}) f(\mathtt{st}) \frac{d\mathtt{t}}{\langle \mathtt{t} \rangle}, \ \mathtt{s} \in \mathbb{R}^N_+, \ w > 0 \qquad \text{(I)}$$

and we develop an approximation theory in the frame of $BV-$spaces. In particular, $f : \mathbb{R}^N_+ \to \mathbb{R}$ will be a function of bounded $\varphi-$variation on $\mathbb{R}^N_+$ and $\{K_w\}_{w>0}$ will be a family of (essentially) bounded approximate identities (see Section IV). Here φ is a convex $\varphi-$function (see Section II) such that $u^{-1}\varphi(u) \to 0$ as $u \to 0^+$. The above operators (I) allow us to obtain, as particular cases, several classes of integral operators well-known and used in approximation theory, such as, for example, the moment-type or average operators, the Gauss-Weierstrass-type operators and others.

The new multidimensional concept of variation that we will use is inspired to the Tonelli approach ([29]) (see also [27] and [30]). Such concept of variation was introduced in [8], and it was adapted to the setting of $\mathbb{R}^N_+$ from the multidimensional $\varphi-$variation defined in [6] in the case of $\mathbb{R}^N$ endowed with the Lebesgue measure. Indeed, in order to treat the Mellin case, it is natural to frame the theory in $\mathbb{R}^N_+$ endowed with the Haar measure $\mu(A) := \int_A \langle \mathtt{t} \rangle^{-1} \, d\mathtt{t}$, where A is a Borel subset of $\mathbb{R}^N_+$ and $\langle \mathtt{t} \rangle := \prod_{i=1}^N t_i$, $\mathtt{t} = (t_1, \ldots, t_N) \in \mathbb{R}^N_+$. We recall that, in the case of the Lebesgue measure, similar approximation results were obtained in [2], [9], while the one-dimensional case was explored in [15] and [14] (nonlinear case).

In order to get convergence of the family $\{T_w f\}_{w>0}$ to f, a crucial tool is to prove that

$$\lim_{\delta\to 0^+} \omega(f,\delta) = 0, \tag{1}$$

where $\omega(f,\delta)$ denotes the modulus of smoothness of f. It is well known that (1) holds if and only if f is absolutely continuous working with the classical (Jordan or Tonelli) variation (see, e.g., [21], [10], [4]). On the contrary, dealing with the $\varphi-$variation, due to the lack of an integral representation of $\varphi-$variation in terms of $\varphi-$absolute continuity, the result is no more trivial. In particular, working with the Musielak-Orlicz $\varphi-$variation, the result can be obtaind by means of a direct construction (see, e.g., [26], [5]). In the multidimensional setting, the situation becomes more delicate. The result was obtaind in [8] where, through an approximation technique by means of step-type functions, we proved that

$$\lim_{\delta\to 0^+} \omega^{\varphi}(\lambda f,\delta) = 0, \tag{2}$$

for some $\lambda > 0$, provided that the function f is $\varphi-$absolutely continuous. Here $\omega^{\varphi}(\lambda f,\delta) := \sup_{|\mathtt{1}-\mathtt{t}|\leq\delta} V^{\varphi}[\lambda(\tau_{\mathtt{t}} f - f)]$ ($\tau_{\mathtt{t}} f(\mathtt{s}) = f(\mathtt{st})$, $\mathtt{s},\mathtt{t}\in\mathbb{R}^N_+$ is the dilation operator and $\mathtt{1}$ is the unit vector of $\mathbb{R}^N_+$) represents the natural reformulation of the classical modulus of smoothness in terms of $\varphi-$variation (see, e.g., [25], [13]). The above result proves that the situation is analogous to the one-dimensional case (see [15], [14]) and to the case of the Lebesgue measure (see, e.g., [1], [2]).

In this paper we develop a new theory about convergence and rate of approximation for the operators (I). In particular we first obtain several estimates for $\{T_w f\}_{w>0}$. Then, by means of such results and using (2), we are able to prove the main convergence theorem, which states that there exists a constant $\mu > 0$ such that

$$\lim_{w\to+\infty} V^{\varphi}[\mu(T_w f - f)] = 0,$$

whenever $f \in AC^{\varphi}(\mathbb{R}^N_+)$ (the space of $\varphi-$absolutely continuous functions). Introducing suitable Lipschitz classes, we also study the problem of the rate of approximation. Moreover, in the particular case of Fejér-type kernels, we obtain that all the assumptions for the rate of approximation are implied by the classical condition that the absolute moments of order α of the kernels are finite.

We finally point out that the case of the classical variation can be also treated, by using a direct approach: indeed, taking the identity function instead of the $\varphi-$function φ, it is possible to obtain a new multidimensional version of the classical Jordan variation in the sense of Tonelli for functions defined on $\mathbb{R}^N_+$ equipped with the logarithmic measure.

II. Notations and definitions

We denote by Φ the class of all the functions φ such that

1) φ is a convex $\varphi-$function, where a $\varphi-$function is a nondecreasing continuous function $\varphi : \mathbb{R}^+_0 \to \mathbb{R}^+_0$ such that $\varphi(0) = 0$, $\varphi(u) > 0$ for $u > 0$ and $\lim_{u\to+\infty}\varphi(u) = +\infty$;
2) $u^{-1}\varphi(u) \to 0$ as $u \to 0^+$.

From now on we will assume that $\varphi \in \Phi$.

Given $f : \mathbb{R}^N_+ \to \mathbb{R}$ and $\mathtt{x} = (x_1,\dots,x_N) \in \mathbb{R}^N_+$, $N \in \mathbb{N}$, if we are interested in particular in the $j-$th coordinate, $j = 1,\dots,N$, we will write

$$x'_j = (x_1,\dots,x_{j-1},x_{j+1},\dots,x_N) \in \mathbb{R}^{N-1}_+,$$

so that $\mathtt{x} = (x'_j, x_j)$ and $f(\mathtt{x}) = f(x'_j, x_j)$. For a fixed interval $I = \prod_{i=1}^{N}[a_i,b_i]$, we will denote by $[a'_j,b'_j]$ the $(N-1)-$dimensional interval obtained deleting by I the $j-$th coordinate, so that

$$I = [a'_j,b'_j]\times[a_j,b_j].$$

Moreover, given two vectors $\mathtt{s},\mathtt{t} \in \mathbb{R}^N_+$, we put $\mathtt{st} = (s_1t_1,\dots,s_Nt_N)$.

In order to define the multidimensional $\varphi-$variation, the first step is to compute the Musielak-Orlicz $\varphi-$variation of the $j-$th section of f, i.e., $V^{\varphi}_{[a_j,b_j]}[f(x'_j,\cdot)]$, and then to consider the $(N-1)-$dimensional integrals

$$\Phi^{\varphi}_j(f,I) := \int_{a'_j}^{b'_j} V^{\varphi}_{[a_j,b_j]}[f(x'_j,\cdot)]\frac{dx'_j}{\langle x'_j\rangle},$$

where by $\langle x'_j\rangle$ we denote the product $\prod_{i=1,i\neq j}^{N} x_i$. We recall that the $\varphi-$variation of a function $g : [a,b] \to \mathbb{R}$ is defined as

$$V^{\varphi}_{[a,b]}[g] := \sup_D \sum_{i=1}^{n} \varphi(|g(s_i) - g(s_{i-1})|),$$

where $D = \{s_0 = a, s_1,\dots,s_n = b\}$ is a partition of $[a,b]$ ([26], [25]), and g is said to be of bounded $\varphi-$variation ($g \in BV^{\varphi}([a,b])$) if $V^{\varphi}_{[a,b]}[\lambda g] < +\infty$, for some $\lambda > 0$.

Let now $\Phi^{\varphi}(f,I)$ be the Euclidean norm of the vector $(\Phi^{\varphi}_1(f,I),\dots,\Phi^{\varphi}_N(f,I))$, namely

$$\Phi^{\varphi}(f,I) := \left\{\sum_{k=1}^{N}[\Phi^{\varphi}_k(f,I)]^2\right\}^{\frac{1}{2}}.$$

We set $\Phi^{\varphi}(f,I) = +\infty$ if $\Phi^{\varphi}_k(f,I) = +\infty$ for some $k = 1,\dots,N$.

We define the multidimensional $\varphi-$variation of f on an interval $I \subset \mathbb{R}^N_+$ as

$$V^{\varphi}[f,I] := \sup \sum_{i=1}^{m} \Phi^{\varphi}(f,J_i),$$

where the supremum is taken over all the finite families of $N-$dimensional intervals $\{J_1,\dots,J_m\}$ which form partitions of I.

The $\varphi-$variation of f over the whole space $\mathbb{R}^N_+$ is defined as

$$V^{\varphi}[f] := \sup_{I\subset\mathbb{R}^N_+} V^{\varphi}[f,I],$$

where the supremum is taken over all the intervals $I \subset \mathbb{R}^N_+$. By $BV^{\varphi}(\mathbb{R}^N_+)$ we denote the *space of functions of bounded $\varphi-$variation* over $\mathbb{R}^N_+$, i.e.,

$$BV^{\varphi}(\mathbb{R}^N_+) = \{f \in L^1_{\mu}(\mathbb{R}^N_+) :\ \exists\lambda > 0 \text{ s.t. } V^{\varphi}[\lambda f] < +\infty\}.$$

We will say that a function $f:\mathbb{R}_+^N\to\mathbb{R}$ is *locally $\varphi-$absolutely continuous* ($f\in AC^{\varphi}_{loc}(\mathbb{R}_+^N)$) if f is (uniformly) $\varphi-$absolutely continuous in the Tonelli sense: this means that for every $I=\prod_{i=1}^{N}[a_i,b_i]\subset\mathbb{R}_+^N$ and for every $j=1,2,\dots,N$, the $j-$th sections of f, $f(x'_j,\cdot):[a_j,b_j]\to\mathbb{R}$, are (uniformly) $\varphi-$absolutely continuous for almost every $x'_j\in[a'_j,b'_j]$, i.e., for every $\varepsilon>0$ there exists $\delta>0$ for which

$$\sum_{i=1}^{n}\varphi(\lambda|f(x'_j,\beta^i)-f(x'_j,\alpha^i)|)<\varepsilon,$$

for a.e. $x'_j\in[a'_j,b'_j]$ and for all finite collections of non-overlapping intervals $[\alpha^i,\beta^i]\subset[a_j,b_j]$, $i=1,\dots,n$, such that $\sum_{i=1}^{n}\varphi(\beta^i-\alpha^i)<\delta$.

By $AC^{\varphi}(\mathbb{R}_+^N)$ we will denote the subspace of $BV^{\varphi}(\mathbb{R}_+^N)$ of the *$\varphi-$absolutely continuous functions*, namely all the functions of bounded $\varphi-$variation which are locally $\varphi-$absolutely continuous.

III. Results about the multidimensional $\varphi-$variation

Our multidimensional $\varphi-$variation satisfies similar properties to the Musielak-Orlicz $\varphi-$variation and to the Jordan variation. In particular we prove that $BV^{\varphi}(\mathbb{R}_+^N)$ is a vector space, namely, $\alpha f_1+\beta f_2\in BV^{\varphi}(\mathbb{R}_+^N)$ whenever $f_1,f_2\in BV^{\varphi}(\mathbb{R}_+^N)$, $\alpha,\beta\in\mathbb{R}$. Indeed this is a consequence of the following property ([8])

$$V^{\varphi}[\lambda(f_1+f_2)]\le\frac{1}{2}\Big(V^{\varphi}[2\lambda f_1]+V^{\varphi}[2\lambda f_2]\Big),\quad\lambda>0,$$

and of the trivial consideration that $V^{\varphi}[\lambda f]\le V^{\varphi}[\mu f]$, if $0<\lambda\le\mu$.

Another classical property of variation which is preserved by our definition is the lower semicontinuity with respect to pointwise convergence. Indeed, in this paper we prove that, if $(f_k)_{k\in\mathbb{N}}$ is pointwise convergent to f, then

$$V^{\varphi}[f]\le\liminf_{k\to+\infty}V^{\varphi}[f_k].$$

Finally it is also possible to prove results about additivity on intervals which are quite similar to the classical ones. Nevertheless we recall that a crucial difference with the Jordan variation is that, in the frame of $\varphi-$variation, even in the one-dimensional case, we don't have at our disposal an integral representation of $\varphi-$variation for absolutely continuous functions.

IV. Mellin operators and convergence results

We will study the following family of Mellin-type integral operators of the form

$$(T_wf)(\mathtt{s})=\int_{\mathbb{R}_+^N}K_w(\mathtt{t})f(\mathtt{st})\frac{d\mathtt{t}}{\langle\mathtt{t}\rangle},\quad w>0,\ \mathtt{s}\in\mathbb{R}_+^N,\quad\text{(I)}$$

for $f\in BV^{\varphi}(\mathbb{R}_+^N)$, where $\{K_w\}_{w>0}$ is a family of bounded approximate identities, i.e.,

$\mathbf{K_w}.\mathbf{1}$) $K_w:\mathbb{R}_+^N\to\mathbb{R}$ is a measurable essentially bounded function such that $K_w\in L^1_{\mu}(\mathbb{R}_+^N)$, $\|K_w\|_{L^1_{\mu}}\le A$ for an absolute constant $A>0$ and $\int_{\mathbb{R}_+^N}K_w(\mathtt{t})\langle\mathtt{t}\rangle^{-1}d\mathtt{t}=1$, for every $w>0$,

$\mathbf{K_w}.\mathbf{2}$) for every fixed $0<\delta<1$, $\int_{|\mathbf{1}-\mathtt{t}|>\delta}|K_w(\mathtt{t})|\langle\mathtt{t}\rangle^{-1}d\mathtt{t}\to0$, as $w\to+\infty$.

We point out that, since K_w is essentially bounded, if $f\in BV^{\varphi}(\mathbb{R}_+^N)$, $(T_wf)(\mathtt{s})$ is well-defined for every $\mathtt{s}\in\mathbb{R}_+^N$ and $w>0$.

We first obtain two estimates for our integral operators (I). The first one proves that $\{T_w\}_{w>0}$ map $BV^{\varphi}(\mathbb{R}_+^N)$ into itself.

Proposition 1: Let $f\in BV^{\varphi}(\mathbb{R}_+^N)$ and let $\{K_w\}_{w>0}$ be such that $K_w.1)$ holds. Then there exists $\lambda>0$ such that

$$V^{\varphi}[\lambda(T_wf)]\le V^{\varphi}[\zeta f],\qquad(3)$$

where $\zeta>0$ is the constant for which $V^{\varphi}[\zeta f]<+\infty$. Therefore, $T_w:BV^{\varphi}(\mathbb{R}_+^N)\to BV^{\varphi}(\mathbb{R}_+^N)$.

The second estimate will be the main tool in order to prove the convergence result. By $\omega^{\varphi}(\lambda f,\delta):=\sup_{|\mathbf{1}-\mathtt{t}|\le\delta}V^{\varphi}[\lambda(\tau_{\mathtt{t}}f-f)]$, where $\tau_{\mathtt{t}}f(\mathtt{s}):=f(\mathtt{st})$, $\mathtt{s},\mathtt{t}\in\mathbb{R}_+^N$ is the dilation operator, we denote the $\varphi-$modulus of smoothness of f.

Proposition 2: Let $f\in BV^{\varphi}(\mathbb{R}_+^N)$ and let $\{K_w\}_{w>0}$ be such that $K_w.1)$ is satisfied. Then for every $\lambda>0$, $\delta\in]0,1[$ and $w>0$,

$$V^{\varphi}[\lambda(T_wf-f)]\le\omega^{\varphi}(\lambda Af,\delta)+A^{-1}V^{\varphi}[2\lambda Af]\int_{|\mathbf{1}-\mathtt{t}|>\delta}|K_w(\mathtt{t})|\langle\mathtt{t}\rangle^{-1}\,d\mathtt{t}.$$

This estimate links the $\varphi-$variation of the error of approximation (T_wf-f) to the $\varphi-$modulus of smoothness, hence the convergence result will follow by the assumptions on kernel functions and by the following result of convergence for $\omega^{\varphi}(f,\delta)$ ([8]):

Theorem 1: Let $f\in AC^{\varphi}(\mathbb{R}_+^N)$. Then there exists $\lambda>0$ such that $\lim_{\delta\to0^+}\omega^{\varphi}(\lambda f,\delta)=0$.

By means of Propositions 1 and 2, and using Theorem 1, we can therefore prove the main convergence result:

Theorem 2: Let $f\in AC^{\varphi}(\mathbb{R}_+^N)$ and let $\{K_w\}_{w>0}$ be such that $K_w.1)$ and $K_w.2)$ are satisfied. Then there exists a constant $\mu>0$ such that

$$\lim_{w\to+\infty}V^{\varphi}[\mu(T_wf-f)]=0.$$

We also obtain results about the order of approximation, with suitable singularity assumptions on kernels, for functions which belong to the Lipschitz class $V^{\varphi}Lip_N(\alpha)$, $\alpha>0$, defined as

$$V^{\varphi}Lip_N(\alpha):=\{f\in BV^{\varphi}(\mathbb{R}_+^N):\ \exists\mu>0\text{ s.t. }V^{\varphi}[\mu\Delta_{\mathtt{t}}f]=O(|\log\mathtt{t}|^{\alpha}),\ \text{as }|\mathbf{1}-\mathtt{t}|\to0\},$$

where $\Delta_{\mathtt{t}}f(\mathtt{x}):=(\tau_{\mathtt{t}}f-f)(\mathtt{x})=f(\mathtt{xt})-f(\mathtt{x})$, for $\mathtt{x},\mathtt{t}\in\mathbb{R}_+^N$, and $\log\mathtt{t}:=(\log t_1,\dots,\log t_N)$.

We point out that, in the particular case of Fejér-type kernels, namely kernels of the form

$$K_w(\mathtt{t})=w^NK(\mathtt{t}^w),\quad\mathtt{t}\in\mathbb{R}_+^N,\ w>0,$$

where $K \in L^1_\mu(\mathbb{R}^N_+)$ is essentially bounded and such that $\int_{\mathbb{R}^N_+} K(\mathtt{t})\langle \mathtt{t}\rangle^{-1} d\mathtt{t} = 1$, it is possible to prove that all the assumptions of the results about the rate of approximation are implied by the classical condition that the absolute moments of order α of K are finite.

V. The particular case of $BV(\mathbb{R}^N_+)$

It is immediate to see that assumption 2) on the $\varphi-$functions implies that the identity function does not belong to the class Φ. Nevertheless all the theory can be developed also for the space $BV(\mathbb{R}^N_+)$, i.e., taking $\varphi(u) = u$, $u \in \mathbb{R}^+_0$, in the definition of the variation, and hence replacing everywhere the Musielak-Orlicz $\varphi-$variation with the Jordan variation. In this setting we obtain a new multidimensional concept of variation in the sense of Tonelli in the frame of Mellin theory and approximation results for Mellin-type integral operators in $BV(\mathbb{R}^N_+)$. Indeed assumption 2) on the $\varphi-$function, that now fails, is just used to prove the convergence result for the $\varphi-$modulus of smoothness (Theorem 1) and it replaces the lack of the integral representation of $\varphi-$variation. On the contrary, working with the classical variation, we have at our disposal the integral representation for absolutely continuous functions, and the convergence of the modulus of smoothness can be derived from it: hence, by means of different techniques, we prove the following

Theorem 3: If $f \in AC(\mathbb{R}^N_+)$, then $\lim_{\delta\to 0^+} \omega(f,\delta) = 0$.

Here $AC(\mathbb{R}^N_+)$ and $\omega(f,\delta)$ denote the space of the absolutely continuous functions and the modulus of smoothness, respectively, in the case $\varphi(u) = u$, $u \in \mathbb{R}^+_0$.

The other results (estimates, convergence and rate of approximation) can be proved in a similar fashion, and therefore we obtain new results also in the case of the classical multidimensional variation in the present setting.

Acknowledgment

The authors would like to thank the Fondazione Cassa di Risparmio di Perugia, Project n. 2010.011.0403 for supporting this research.

References

[1] L. Angeloni, *A characterization of a modulus of smoothness in multidimensional setting,* Boll. Unione Mat. Ital., Serie IX, **4** (1) (2011), 79–108.

[2] L. Angeloni, *Convergence in variation for a homothetic modulus of smoothness in multidimensional setting,* Comm. Appl. Nonlinear Anal., **19**(1) (2012), 1–22.

[3] L. Angeloni, *Approximation results with respect to multidimensional $\varphi-$variation for nonlinear integral operators*, Z. Anal. Anwendungen, **32**(1) (2013), 103–128.

[4] L. Angeloni and G. Vinti, *Convergence in Variation and Rate of Approximation for Nonlinear Integral Operators of Convolution Type,* Results Math., **49**(1-2) (2006), 1–23.

[5] L. Angeloni and G. Vinti, *Approximation by means of nonlinear integral operators in the space of functions with bounded $\varphi-$variation,* Differential Integral Equations, **20**(3) (2007), 339–360.

[6] L. Angeloni and G. Vinti, *Convergence and rate of approximation for linear integral operators in $BV^\varphi-$spaces in multidimensional setting,* J. Math. Anal. Appl., **349** (2009), 317–334.

[7] L. Angeloni and G. Vinti, *Approximation with respect to Goffman-Serrin variation by means of non-convolution type integral operators,* Numer. Funct. Anal. Optim., **31** (2010), 519–548.

[8] L. Angeloni and G. Vinti, *A sufficient condition for the convergence of a certain modulus of smoothness in multidimensional setting,* Comm. Appl. Nonlinear Anal., **20**(1) (2013), 1–20.

[9] L. Angeloni and G. Vinti, *Approximation in variation by homothetic operators in multidimensional setting*, Differential Integral Equations, **26**(5-6) (2013), 655–674.

[10] C. Bardaro, P.L. Butzer, R.L. Stens, and G. Vinti, *Convergence in variation and rates of approximation for Bernstein-type polynomials and singular convolution integrals,* Analysis, **23** (2003), 299–340.

[11] C. Bardaro and I. Mantellini, *Voronovskaya-type estimates for Mellin convolution operators*, Results Math., **50** (2007), 1–16.

[12] C. Bardaro and I. Mantellini, *Quantitative Voronovskaya formula for Mellin convolution operators*, Mediterr. J. Math., **7**(4) (2010), 483–501.

[13] C. Bardaro, J. Musielak, and G. Vinti, *Nonlinear Integral Operators and Applications,* De Gruyter Series in Nonlinear Analysis and Applications, New York, Berlin, 9, 2003.

[14] C. Bardaro, S. Sciamannini, and G. Vinti, *Convergence in BV_φ by nonlinear Mellin-type convolution operators,* Funct. Approx. Comment. Math., **29** (2001), 17–28.

[15] C. Bardaro and G. Vinti, *On convergence of moment operators with respect to $\varphi-$variation,* Appl. Anal., **41** (1991), 247–256.

[16] C. Bardaro and G. Vinti, *General convergence theorem with respect to Cesari variation and applications*, Nonlinear Anal., **22** (1994), 505–518.

[17] C. Bardaro and G. Vinti, *On the order of $BV^\varphi-$approximation of convolution integrals over the line group,* Comment. Math., Tomus Specialis in Honorem Iuliani Musielak (2004), 47–63.

[18] M. Bertero and E.R. Pike, *Exponential-sampling method for Laplace and other dilationally invariant transforms: I. Singular-system analysis. II. Examples in photon correlation spectroscopy and Fraunhofer diffraction,* Inverse Problems, **7** (1991), 1–20, 21–41.

[19] P.L. Butzer and S. Jansche, *A direct approach to the Mellin Transform*, J. Fourier Anal. Appl., **3** (1997), 325–376.

[20] P.L. Butzer and S. Jansche, *The Exponential Sampling Theorem of Signal Analysis*, Atti Sem. Mat. Fis. Univ. Modena, Suppl. Vol. **46**, a special issue of the International Conference in Honour of Prof. Calogero Vinti (1998), 99–122.

[21] P.L. Butzer and R.J. Nessel, *Fourier Analysis and Approximation, I,* Academic Press, New York-London, 1971.

[22] D. Casasent (Ed.), *Optical Data Processing*, Springer, Berlin (1978), 241–282.

[23] R.G. Mamedov, *The Mellin transform and approximation theory*, "Elm", Baku, 1991, (in Russian).

[24] I. Mantellini and G. Vinti, *$\Phi-$variation and nonlinear integral operators,* Atti Sem. Mat. Fis. Univ. Modena, Suppl. Vol. **46**, a special issue of the International Conference in Honour of Prof. Calogero Vinti (1998), 847–862.

[25] J. Musielak, *Orlicz Spaces and Modular Spaces,* Springer-Verlag, Lecture Notes in Math., 1034, 1983.

[26] J. Musielak and W. Orlicz, *On generalized variations (I),* Studia Math., **18** (1959), 11–41.

[27] T. Radò, *Length and Area,* Amer. Math. Soc. Colloquium Publications, **30**, 1948.

[28] S. Sciamannini and G. Vinti, *Convergence and rate of approximation in BV_φ for a class of integral operators,* Approx. Theory Appl., **17** (2001), 17–35.

[29] L. Tonelli, *Su alcuni concetti dell'analisi moderna,* Ann. Scuola Norm. Super. Pisa, **11**(2) (1942), 107–118.

[30] C. Vinti, *Perimetro—variazione,* Ann. Scuola Norm. Sup. Pisa, **18**(3) (1964), 201–231.

Iterative methods for random sampling and compressed sensing recovery

Masoumeh Azghani
ACRI and EE Dept.
sharif university of Technology
Tehran, Iran
Email:azghani@ee.sharif.ir

Farokh Marvasti
ACRI and EE Dept.
sharif university of Technology
Tehran, Iran
Email:Marvasti@sharif.edu

Abstract— **In this paper, two methods are proposed which address the random sampling and compressed sensing recovery problems. The proposed random sampling recovery method is the Iterative Method with Adaptive Thresholding and Interpolation (IMATI). Simulation results indicate that the proposed method outperforms existing random sampling recovery methods such as Iterative Method with Adaptive Thresholding (IMAT). Moreover, the suggested method surpasses compressed sensing recovery methods such as Orthogonal Matching Pursuit (OMP) in terms of recovery performance. We propose a compressed sensing recovery method, named Iterative Method with Adaptive Thresholding for Compressed Sensing recovery (IMATCS). Unlike its counterpart, Iterative Hard Thresholding (IHT), the thresholding function of the proposed method is adaptive i.e. the threshold value changes with the iteration number, which enables IMATCS to reconstruct the sparse signal without having any knowledge of the sparsity number. The simulation results indicate that IMATCS outperforms IHT and OMP in both computational complexity and quality of the recovered signal.**

I. Introduction

Sparse recovery methods have found broad applications in various areas such as imaging systems, multipath channel estimation, spectral estimation, and coding. Depending on various kinds of sparsity (low pass, high pass, or random) and various sampling techniques (uniform or random), different methods have been suggested in the literature for reconstruction of sparse signals [1]. When the location of sparsity is known, the number of samples required for exact reconstruction equals the sparsity number. Some of the recovery methods in this case are suggested in [1]. When the location of the sparsity is unknown, the number of samples must be at least twice the sparsity number to identify both the locations and the values of the coefficients[2]. More sophisticated recovery methods are required in this case which can be grouped based on the sampling strategy. One sampling strategy is to take linear combinations of the signal entries which is the focus of the Compressed Sensing (CS) techniques. The second sampling scheme is to take random samples of the signal entries. In CS [3, 4], linear combinations of the signal coefficients are taken instead of directly sampling the signal. Many compressed sensing recovery algorithms have been proposed, ranging from convex relaxation techniques to greedy approaches such as Orthogonal Matching Pursuit (OMP) [5] to iterative thresholding schemes such as Iterative Hard Thresholding (IHT) [6, 7]. IHT is proposed for compressed sensing recovery of sparse signals when the sparsity number of the signal is known. In [8], normalized IHT algorithm is proposed which is a stabilized version of IHT. In [2, 9], the Iterative Method with Adaptive Thresholding (IMAT) is proposed to recover the signal from its random samples. The random samples in this case are random selection of the signal entries. The IMAT recovers the underlying sparse signal by alternating projections between the information domain and the sparsity domain (the domain in which the signal is sparse). In order to take advantage of the sparsity of the embedded signal, IMAT thresholds adaptively the signal (by decreasing or increasing the threshold levels) in such a way that the coefficients are picked up gradually after some iterations. In this paper, two methods are proposed which address the random sampling and CS recovery problems. The proposed random sampling recovery method is the Iterative Method with Adaptive Thresholding and Interpolation (IMATI) which is a modified version of the IMAT. At each iteration, a crude reconstruction of the signal based on linear interpolation is obtained. The adaptive thresholding scheme is exploited to promote sparsity. The proposed compressed sensing recovery method is Iterative Method with Adaptive Thresholding for Compressed Sensing recovery (IMATCS). We note that IMATCS is closely related to IHT method [6, 7], except that the thresholding function is adaptive, i.e., the threshold value changes with the iteration number, which enables IMATCS to reconstruct the sparse signal without having any knowledge of the sparsity number. The simulation results indicate that the IMATI method outperforms the IMAT method. Also, we conclude that random sampling recovery (using IMAT or IMATI) is a good choice for signal compression compared to CS recovery techniques such as Orthogonal Matching Pursuit (OMP), and there is no need to add more complexity to take linear combination of the signal coefficients. However, in some applications the linear combinations of the signal coefficients are imposed by the problem. In such cases, the compressed sensing recovery techniques are the only solution. The simulation results indicate that IMATCS provides better and faster reconstruction compared to normalized IHT, although IHT has an extra information of the sparsity number. Also, the recovery performance of IMATCS is better than that of OMP with less computational complexity.

The rest of the paper is organized as follows: The IMATCS method is proposed in Section II. The proposed IMATI method is presented in section III. The simulation results are given in Section IV. Finally, Section V concludes this work.

II. Iterative Method for Compressed Sensing recovery (IMATCS)

In this section, the proposed Iterative Method for Compressed Sensing recovery (IMATCS) is illustrated. Let S be $M \times 1$ signal and Φ be $L \times M$ ($L \langle M$) measurement matrix. The problem is to recover S from its measurement vector Y=$\Phi \times S$ with the constraint that S is sparse in the Ψ domain, S=$\Psi \times X$. In other words, the coefficient vector X has a small number of non-zero entries. The transformation matrix, can be DCT, DWT or DFT. The IMATCS method can be considered as a variant of IHT based on adaptive thresholding. The mathematical formulation of the method is as follows:

$$x_{k+1} = T(x_k + \times A^H(Y - A \times x_k)) \quad (1)$$

$$A = \Phi \times \Psi \tag{2}$$

$$S_{recovered} = \Psi \times x_{itermax} \tag{3}$$

λ is the relaxation parameter which controls the convergence of the algorithm. T is the thresholding function decreased iteration by iteration in an exponential manner as follows:

$$T = T_0 \times exp(-\alpha \times K) \tag{4}$$

where K is the iteration number and λ indicates the threshold step and is determined empirically. The algorithm starts from zero initial value, $x_0 = 0$. After a number of iterations, indicated by itermax, the coefficient vector is recovered as $x_{itermax}$. The adaptivity of the threshold enables us to recover the embedding signal from its linear measurements without any knowledge of the sparsity number of the signal.

III. Iterative Method with Adaptive Thresholding and Interpolation (IMATI)

Another problem which is addressed here is the recovery of the sparse signal from a random selection of its entries, i.e. random sampling. The proposed method in this case is IMATI which rely on some modifications to the well-known iterative method [10]. The conventional iterative method has originally been proposed in the field of non-uniform sampling recovery for low pass or high pass signals (a special kind of sparse signals). In order to promote sparsity, a thresholding operator is used at the end of each iteration. In random sampling, the measurements are a subset of signal entries. Hence, the random sampling measurement matrix, Φ_R, consists of a random selection of the rows of the identity matrix. The formulation of the IMATI method is given as:

$$x_{k+1} = T(x_k + \times Interpl(Y - A_R \times x_k)) \tag{5}$$

$$A_R = \Phi_R \times \Psi \tag{6}$$

$$S_{recovered} = \Psi \times x_{itermax} \tag{7}$$

The above formulation of IMATI shows the analogy of the two proposed methods, IMATCS and IMATI. A crude reconstruction scheme is used successively and the recovered signal at each iteration is sparsified using an adaptive threshold. In IMATCS method, the measurements are linear combinations of the signal entries and the iterated recovery is based on the transpose of the matrix, i.e. AH. In IMATI, a random selection of the signal entries is available as measurements and the crude reconstruction scheme is based on linear interpolation. Furthermore, in order to promote sparsity, exponential adaptive thresholding is used in the proposed methods. The IMATI method can be implemented in a more efficient way according to the block diagram depicted in Figure 1.

The G operator applies the sampling and interpolation. The random sampling scheme can be implemented by an inner product of the image with a binary sampling mask. Moreover, the linear interpolation can be applied to the sampled image using a sliding interpolating window. Therefore, the above implementation enables IMATI to process the whole image at once.

IV. Simulation Results

In this part, the simulation results are reported. The parameters of IMATI method are set as: $T_0 = 66363$, $\alpha = 0.6$, $\lambda = 1.8$,itermax=35. The parameters of IMATCS are set as follows: $T_0 = 900$,$\alpha = 0.2$, $\lambda = 0.3$, itermax=100.

Two kinds of interpolators have been exploited in IMATI method:

- Linear interpolation using sliding window 3×3

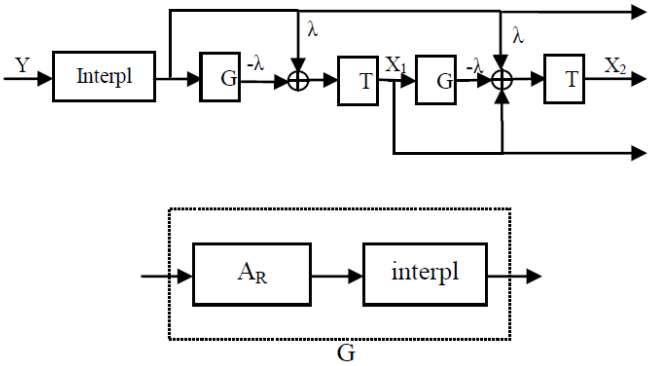


Fig. 1. Block diagram of IMATI method.

The missing pixel is replaced by a weighted average of the 3×3 neighbors. The IMATI method in this case is named IMATLI.

- Sample and hold interpolation
  the missing pixels are replaced by their neighboring samples in the top or left. The IMATI method in this case is called IMATSH.

In the case of random sampling recovery methods such as IMAT, IMATSH and IMATLI, the whole of the image is processed at once without dividing it into small blocks, while 8×8 blocks of the image are processed separately for compressed sensing recovery methods such as OMP, normalized IHT and IMATCS. The performances of IMATLI, IMATSH, IMAT and OMP methods are compared in Figure 2.

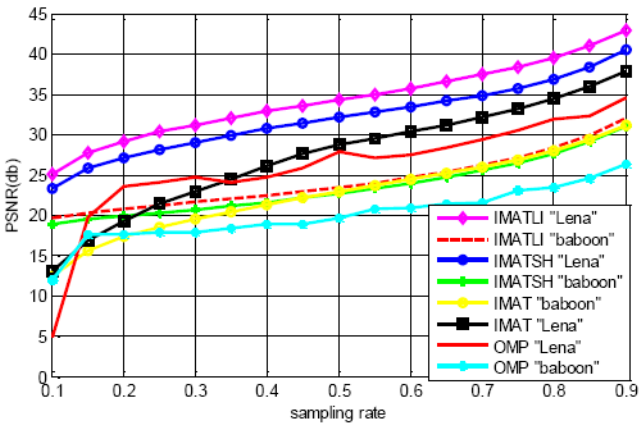


Fig. 2. comparison of IMATI method with IMAT and OMP

According to Figure 2, the IMATLI method has better recovery performance than the IMATSH and both of them outperform IMAT. The OMP method has the worst recovery performance among the all. The simulation time of the methods are compared in Figure 3 as a trustable complexity measure.

Comparing the simulation times of the methods, we observe that IMAT is much more complex than IMATLI and IMATSH especially for lower sampling rates. Furthermore, the IMATSH is faster than IMATLI. The simulation time of OMP goes up as the sampling rate increases and its complexity is more than those of IMAT and IMATSH especially for higher sampling rates. The IMATCS method

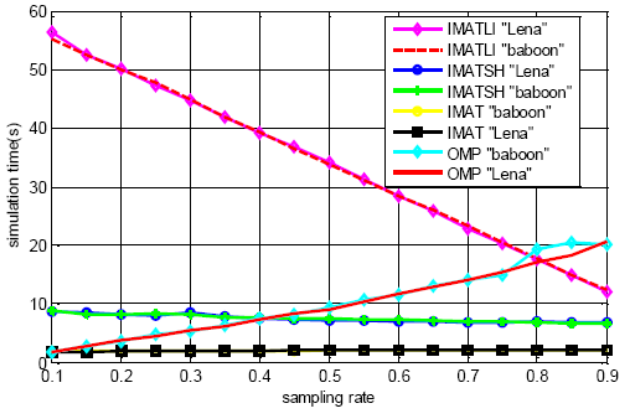


Fig. 3. simulation time of IMATLI, IMATSH, IMAT and OMP

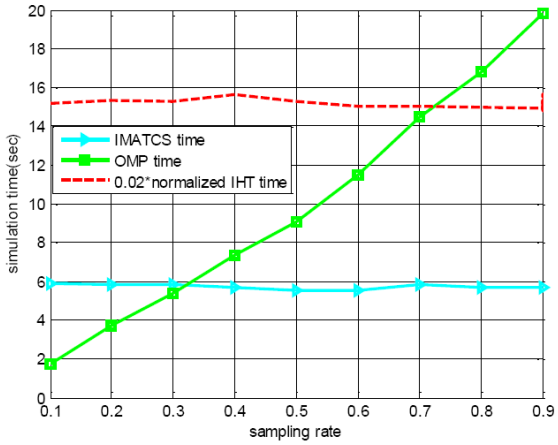


Fig. 5. simulation time of IMATCS, OMP and normalized IHT

is compared to some well known compressed sensing recovery methods such as OMP [5] and normalized IHT [8] in the case of natural image recovery. As the IHT method requires the signal to be k-sparse for efficient performance, we sparsify the image in the DCT domain up to 20 %. However, for the other two simulated methods (OMP and proposed IMATCS), the original (non-sparse) image is used. The efficiency of the recovery methods for various sampling rates are compared at Figure 4.

Having a look at Figure 4, we understand that the performance

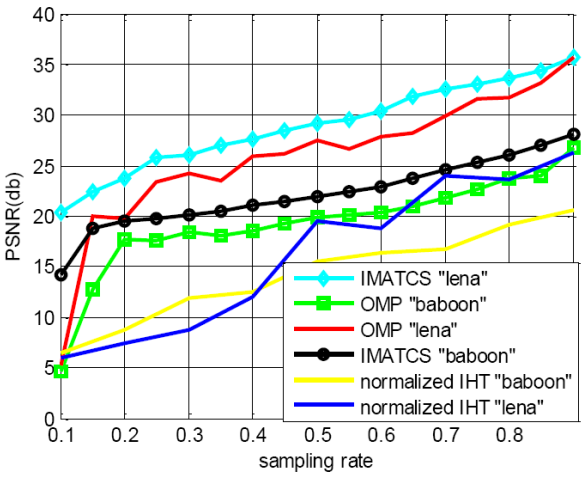


Fig. 4. recovery performance of IMATCS, OMP and normalized IHT

of IMATCS is similar to that of OMP and much better than that of normalized IHT for various sampling rates. To compare the complexities of the methods, the simulation time is shown in Figure 5.

According to Figure 5, the simulation time of normalized IHT is more than 50 times those of OMP and IMATCS. Furthermore, simulation time of OMP increases with the sampling rate while IMATCS has an approximately steady low simulation time. Consequently, the complexity of IMATCS is low and does not change for various sampling rates which can be an excellent characteristic from practical point of view, since a fixed and flexible implementation test bed can be used for various sampling rates.

V. Conclusion

In the case of IMATI method, the linear interpolation performs slightly better than sample and hold interpolation at the cost of more complexity. The IMATSH and IMATLI methods reconstruct the signal better than what IMAT does at the cost of more simulation time. Furthermore, the random sampling recovery techniques including IMAT and IMATI methods, outperform OMP (CS recovery technique) in both simplicity and recovery performance. Also, they exploit the spatial correlations in the 2-D image by taking 2-D DCT transform of the image and it is unnecessary to divide the whole image into small blocks as required in OMP. As a result, for the purpose of signal compression, one does not need a compressive matrix to take linear measurements of the signal coefficients and it is shown in this work that direct random sampling recovery (using IMAT and IMATI) performs better than compressive sampling recovery using OMP. However, when we are faced with an ill-posed system of equations (which inherently has the linear combinations of signal coefficients for example, in MRI imaging), compressed sensing recovery techniques are the only solutions. The simulation results indicate that the proposed CS recovery technique, IMATCS, outperforms IHT in both recovery performance and computational complexity without any need to have knowledge of the sparsity number. Moreover, IMATCS surpasses OMP in terms of recovery performance and convergence speed.

References

[1] F. Marvasti, *Nonuniform Sampling: Theory and Practice*, Springer, formerly Kluwer Academic/Plenum Publishers, 2001.

[2] F. Marvasti , A. Amini, F. Haddadi, et al, *A Unified Approach to Sparse Signal Precessing*, EURASIP Journal on Advances in Signal Processing , 2012.

[3] E. J. Candes, J. K. Romberg, and T. Tao, *Stable signal recovery from incomplete and inaccurate measurements*, Communications on Pure and Applied Mathematics , Vol.59, No.8, 2006.

[4] D. L. Donoho, *Compressed sensing*, IEEE Transaction on Information Theory , Vol.52, No.4, 2006.

[5] A. Tropp and A. C. Gilbert, *Signal recovery from partial information via orthogonal matching pursuit*, IEEE Transaction on Information Theory , Vol.53, No.12, 2007.

[6] T. Blumensath and M. E. Davies, *Iterative thresholding for sparse approximations*, Journal of Fourier Analysis and Applications , Vol.14, No.5, 2008.

[7] M. Fornasier and H. Rauhut, *Iterative Thresholding Algorithms*, Journal of Fourier Analysis and Applications , Vol.14, No.5, 2008.

[8] T. Blumensath and M. Davies *Normalized Iterative Hard Thresholding: Guaranteed Stability and Performance* IEEE Journal of Selected Topics in Signal Processing, Vol. 4, No. 2, 2010.

[9] R. Eghbali , A. Kazerooni , A. Rashidinejad and F. Marvasti, *Iterative method with adaptive thresholding for sparse signal reconstruction*, International Workshop on Sampling Theory and Applications (SampTA) , Singapore, May 2011.

[10] F. Marvasti, *An iterative method to compensate for the interpolation distortion*, IEEE Transactions on Acoustics, Speech and Signal Processing , Vol. 37, No. 10, 1989.

A Review of the Invertibility of Frame Multipliers

Peter Balazs
Acoustics Research Institute
Wohllebengasse 12-14, Vienna A-1040, Austria
peter.balazs@oeaw.ac.at

Diana T. Stoeva
Acoustics Research Institute
Wohllebengasse 12-14, Vienna A-1040, Austria
dstoeva@kfs.oeaw.ac.at

***Abstract*—In this paper we give a review of recent results on the invertibility of frame multipliers $M_{m,\Phi,\Psi}$. In particular, we give sufficient, necessary or equivalent conditions for the invertibility of such operators, depending on the properties of the sequences Ψ, Φ and m. We consider Bessel sequences, frames, and Riesz bases.**

I. Introduction

Certain mathematical objects appear in a lot of scientific disciplines, like physics [2], signal processing [16] and certainly mathematics [13]. In a general setting they can be described as frame multipliers, consisting of analysis, multiplication by a fixed sequence (called the symbol), and synthesis:

$$M_{m,\Phi,\Psi} f = \sum_k m_k \, \langle f, \psi_k \rangle \, \phi_k.$$

They are not only interesting mathematical objects [4], [7], [9], [12], [17], but also important for applications, for example for the realization of time-varying filters [8], [15], [24]. Therefore, for some applications it is important to find the inverse of a multiplier if it exists.

Here we collect results from [3], [20], [21], [23] about the invertibility of such operators.

II. Preliminaries and Notations

Throughout the paper, $\mathcal{H}$ denotes an infinite-dimensional Hilbert space, Φ (resp. Ψ) denotes a sequence $(\phi_n)_{n=1}^{\infty}$ (resp. $(\psi_n)_{n=1}^{\infty}$) with elements from $\mathcal{H}$, m denotes a complex scalar sequence $(m_n)_{n=1}^{\infty}$, $\overline{m} = (\overline{m}_n)_{n=1}^{\infty}$, and $m\Psi = (m_n\psi_n)_{n=1}^{\infty}$. The sequence m is called *semi-normalized* if $0 < \inf_n |m_n| \leq \sup_n |m_n| < \infty$. When the index set is omitted, $\mathbb{N}$ should be understood as the index set. A multiplier $M_{m,\Phi,\Psi}$ is the operator given by $M_{m,\Phi,\Psi} h = \sum_{n=1}^{\infty} m_n \langle h, \psi_n \rangle \phi_n$. If not stated otherwise, M denotes any one of the multipliers $M_{m,\Phi,\Psi}$ and $M_{m,\Psi,\Phi}$. The identity operator on $\mathcal{H}$ is denoted by $I_{\mathcal{H}}$. An operator $T : \mathcal{H} \to \mathcal{H}$ is called *invertible* if it is a bounded bijection. Depending on the sequences Φ, Ψ, and m, the multiplier $M_{m,\Phi,\Psi}$ might be well defined or not well defined, as well as invertible or not invertible. In [22] an extensive collection of examples for all those cases can be found.

Recall that Φ is called a *Bessel sequence* (in short, *Bessel*) *in $\mathcal{H}$ with bound B_Φ* if $B_\Phi > 0$ and $\sum |\langle h, \phi_n \rangle|^2 \leq B_\Phi \|h\|^2$ for every $h \in \mathcal{H}$. A Bessel sequence Φ in $\mathcal{H}$ with bound B_Φ is called a *frame for $\mathcal{H}$ with bounds A_Φ, B_Φ*, if $A_\Phi > 0$ and $A_\Phi \|h\|^2 \leq \sum |\langle h, \phi_n \rangle|^2$ for every $h \in \mathcal{H}$. For a given frame Φ for $\mathcal{H}$, its *frame operator* S_Φ is given by $S_\Phi h = \sum \langle h, \phi_n \rangle \phi_n$, $h \in \mathcal{H}$. When Φ is a frame for $\mathcal{H}$, there exist frames $\Phi^d = (\phi_n^d)$ satisfying $h = \sum \langle h, \phi_n^d \rangle \phi_n = \sum \langle h, \phi_n \rangle \phi_n^d$ for every $h \in \mathcal{H}$. Such frames Φ^d are called *dual frames of* Φ.

The sequence Φ is called a *Riesz basis for $\mathcal{H}$ with bounds* A_Φ, B_Φ, if Φ is complete in $\mathcal{H}$, $A_\Phi > 0$, and $A_\Phi \sum |c_n|^2 \leq \| \sum c_n \phi_n \|^2 \leq B_\Phi \sum |c_n|^2$, $\forall (c_n) \in \ell^2$. Every Riesz basis for $\mathcal{H}$ with bounds A, B is a frame for $\mathcal{H}$ with bounds A, B. For standard references for frame theory and related topics see [10], [11], [14].

The notion of frame multipliers is naturally connected to the one of weighted frames [6], [19]. Frames can also be used for the representation of operators [5]. In this setting multipliers are those operators that can be represented with diagonal matrices.

III. Sufficient and necessary conditions for invertibility of multipliers for Riesz bases

We start with the case of Riesz bases, where we can give equivalent conditions for the invertility of multipliers. We start with an easy result:

Proposition 3.1: [3] If Φ and Ψ are Riesz bases and m is semi-normalized, then $M_{m,\Phi,\Psi}$ is invertible and its inverse can be written as $M_{(1/m_n),\widetilde{\Psi},\widetilde{\Phi}}$, where $\widetilde{\Psi}$ and $\widetilde{\Phi}$ are the unique biorthogonal sequences to Ψ and Φ, respectively.

Even more, it can be shown, that if two of those three assumptions are assumed, the third one is equivalent to the invertibility of M:

Theorem 3.2: [20] Let Φ be a Riesz basis for $\mathcal{H}$. Then the following holds.

(i) If Ψ is a Riesz basis for $\mathcal{H}$, then M is invertible if and only if m is semi-normalized.
(ii) If m is semi-normalized, then M is invertible if and only if Ψ is a Riesz basis for $\mathcal{H}$.

More detailed, we can distinguish the cases of well-definedness, invertibility, injectivity, and surjectivity of multipliers for Riesz bases:

Proposition 3.3: [23] Let Φ be a Riesz basis for $\mathcal{H}$. The following equivalences hold.

(a) M is well defined if and only if $m\Psi$ is a Bessel sequence in $\mathcal{H}$.
(b) M is invertible if and only if $m\Psi$ is a Riesz basis for $\mathcal{H}$.
(c1) $M_{m,\Phi,\Psi}$ is injective if and only if $m\Psi$ is a complete Bessel sequence in $\mathcal{H}$.

(c2) $M_{m,\Psi,\Phi}$ is injective if and only if the synthesis operator $T_{m\Psi}$, given by $T_{m\Psi}(c_n) = \sum c_n m_n \psi_n$ for $(c_n) \in \ell^2$, is injective.

(d1) $M_{m,\Phi,\Psi}$ is surjective if and only if $\overline{m}\Psi$ is a Riesz basis for its closed linear span.

(d2) $M_{m,\Psi,\Phi}$ is surjective if and only if $m\Psi$ is a frame for $\mathcal{H}$.

As seen above, when at least one of Φ and Ψ is not a Riesz basis or m is not semi-normalized, then M does not need to be invertible. In such cases the multiplier M even does not need to be well defined (one can give simple examples, see e.g. [22]).

IV. Sufficient conditions for invertibility of multipliers for frames

In this section we consider the more general case, where it is assumed that Φ is a frame. We give four sufficient conditions for the invertibility of multipliers, give representations of the inverses as operator sums and give the corresponding n-term error:

Proposition 4.1: [20] Let Φ be a frame for $\mathcal{H}$. Assume that

$\mathcal{P}_1$: $\exists\, \mu \in [0, \frac{A_\Phi^2}{B_\Phi})$ such that $\sum |\langle h, \psi_n - \phi_n\rangle|^2 \leq \mu \|h\|^2$, $\forall h \in \mathcal{H}$.

For every positive (or negative) semi-normalized sequence m, satisfying

$$\frac{\sup_n |m_n|}{\inf_n |m_n|}\sqrt{\mu} < \frac{A_\Phi}{\sqrt{B_\Phi}},$$

it follows that Ψ is a frame for $\mathcal{H}$, M is invertible,

$$M^{-1} = \sum_{k=0}^{\infty} [S^{-1}_{(\sqrt{m_n}\phi_n)}(S_{(\sqrt{m_n}\phi_n)} - M)]^k S^{-1}_{(\sqrt{m_n}\phi_n)},$$

if $m_n > 0, \forall n \in \mathbb{N}$, and

$$M^{-1} = -\sum_{k=0}^{\infty} [S^{-1}_{(\sqrt{|m_n|}\phi_n)}(S_{(\sqrt{|m_n|}\phi_n)} + M)]^k S^{-1}_{(\sqrt{|m_n|}\phi_n)},$$

if $m_n < 0, \forall n \in \mathbb{N}$,
where the n-term error is bounded by the constant $\frac{\left(b\sqrt{\mu B_\Phi}\right)^{n+1}}{aA_\Phi - b\sqrt{\mu B_\Phi}} \left(\frac{1}{aA_\Phi}\right)^{n+1}$.

Proposition 4.2: [20] Let Φ be a frame for $\mathcal{H}$ and $\mathcal{P}_1$ hold. Let m satisfy

$$|m_n - 1| \leq \lambda < \frac{A_\Phi - \sqrt{\mu B_\Phi}}{B_\Phi + \sqrt{\mu B_\Phi}},\ \forall n \in \mathbb{N}, \tag{1}$$

for some λ. Then Ψ is a frame for $\mathcal{H}$, $M_{m,\Phi,\Phi}$ and M are invertible,

$$M^{-1}_{m,\Phi,\Phi} = \sum_{k=0}^{\infty} [S_\Phi^{-1}(S_\Phi - M_{m,\Phi,\Phi})]^k S_\Phi^{-1},$$

where the n-term error is bounded by $\frac{1}{A_\Phi - \lambda B_\Phi}\left(\frac{\lambda B_\Phi}{A_\Phi}\right)^{n+1}$, and

$$M^{-1} = \sum_{k=0}^{\infty} [M^{-1}_{m,\Phi,\Phi}(M_{m,\Phi,\Phi} - M)]^k M^{-1}_{m,\Phi,\Phi},$$

where the n-term error is bounded by the constant $\frac{1}{A_\Phi - \lambda B_\Phi - (\lambda+1)\sqrt{\mu B_\Phi}}\left(\frac{(\lambda+1)\sqrt{\mu B_\Phi}}{A_\Phi - \lambda B_\Phi}\right)^{n+1}$.

Proposition 4.3: [20] Let Φ be a frame for $\mathcal{H}$ and $\Phi^d = (\phi_n^d)$ be a dual frame of Φ. Let M denote any one of M_{m,Φ,Φ^d} and $M_{m,\Phi^d,\Phi}$, and m be such that $|m_n - 1| \leq \lambda < \frac{1}{\sqrt{B_\Phi B_{\Phi^d}}}$, $\forall n \in \mathbb{N}$, for some λ. Then M is invertible,

$$M^{-1}_{m,\Phi^d,\Phi} = \sum_{k=0}^{\infty} (M_{(1-m_n),\Phi^d,\Phi})^k,$$

and the n-term error is bounded by $\frac{\left(\lambda \cdot \sqrt{B_\Phi B_{\Phi^d}}\right)^{n+1}}{1 - \lambda \cdot \sqrt{B_\Phi B_{\Phi^d}}}$.

Proposition 4.4: [20] Let Φ be a frame for $\mathcal{H}$. Assume that

$\mathcal{P}_2$: $\exists\, \mu \in [0, \frac{1}{B_\Phi})$ such that $\sum |\langle h, m_n\psi_n - \phi_n^d\rangle|^2 \leq \mu \|h\|^2$, $\forall\ h \in \mathcal{H}$,

for some dual frame $\Phi^d = (\phi_n^d)$ of Φ. Let M denote any one of $M_{\overline{m},\Phi,\Psi}$ and $M_{m,\Psi,\Phi}$. Then $m\Psi$ is a frame for $\mathcal{H}$, M is invertible,

$$M^{-1} = \sum_{k=0}^{\infty} (I_\mathcal{H} - M)^k,$$

and the n-term error is bounded by $\frac{\left(\sqrt{\mu B_\Phi}\right)^{n+1}}{1 - \sqrt{\mu B_\Phi}}$.

V. Invertibility of Multipliers for equivalent frames

Two frames Φ and Ψ are called equivalent [1], [10], if there exists a bounded bijective operator G, such that $\psi_n = G\phi_n$ for every $n \in \mathbb{N}$.

Proposition 5.1: [20] Let Φ and Ψ be equivalent frames for $\mathcal{H}$. Let m be semi-normalized and satisfy one of the following three conditions:

- $m_n > 0$ for every $n \in \mathbb{N}$;
- $m_n < 0$ for every $n \in \mathbb{N}$; or
- there exists λ with $|m_n - 1| \leq \lambda < A_\Phi/B_\Phi$, $\forall n \in \mathbb{N}$.

Then M and $M_{m,\Phi,\Phi}$ are invertible, $M^{-1}_{m,\Phi,\Psi} = (G^{-1})^* M^{-1}_{m,\Phi,\Phi}$, and $M^{-1}_{m,\Psi,\Phi} = M^{-1}_{m,\Phi,\Phi} G^{-1}$.

VI. Necessary conditions for invertibility of multipliers for Bessel sequences

In this section we generalize the assumptions, considering the invertibility of multipliers for Bessel sequences.

Let one of the sequences Φ and Ψ be a Bessel sequence in $\mathcal{H}$. If the multiplier $M_{m,\Phi,\Psi}$ is invertible, then the other sequence does not need to be a Bessel sequence. The next statement contains necessary conditions for invertibility concerning the other sequence. In particular, it shows that if Φ and Ψ are Bessel sequences and m is bounded, then the multiplier $M_{m,\Phi,\Psi}$ can be invertible only if Φ and Ψ are frames for $\mathcal{H}$.

Proposition 6.1: [20] Let $M_{m,\Phi,\Psi}$ be invertible.

(i) If Ψ (resp. Φ) is a Bessel sequence in $\mathcal{H}$ with bound B, then $m\Phi$ (resp. $m\Psi$) satisfies the lower frame condition for $\mathcal{H}$ with bound $\frac{1}{B\,\|M^{-1}_{m,\Phi,\Psi}\|^2}$.

(ii) If Ψ (resp. Φ) and $m\Phi$ (resp. $m\Psi$) are Bessel sequences in $\mathcal{H}$, then they are frames for $\mathcal{H}$.

(iii) If Ψ (resp. Φ) is a Bessel sequence in $\mathcal{H}$ and $m \in \ell^\infty$, then Φ (resp. Ψ) satisfies the lower frame condition for $\mathcal{H}$.

(iv) If Ψ and Φ are Bessel sequences in $\mathcal{H}$ and $m \in \ell^\infty$, then Ψ, Φ, $m\Phi$, and $m\Psi$ are frames for $\mathcal{H}$.

VII. Unconditionally convergent invertible multipliers

The previous results contain conclusions for both of the multipliers $M_{m,\Phi,\Psi}$ and $M_{m,\Psi,\Phi}$ (under the same assumptions in each statement). This leads naturally to the question for a connection between the multipliers $M_{m,\Phi,\Psi}$ and $M_{m,\Psi,\Phi}$. Note that $M_{m,\Phi,\Psi}$ being invertible is not equivalent to $M_{m,\Psi,\Phi}$ being invertible, see Example 2.2 in [21]. The next statement gives an equivalence of the invertibility of those operators when unconditionally convergent multipliers $M_{m,\Phi,\Psi}$ and $M_{m,\Psi,\Phi}$ are considered.

Proposition 7.1: [21] For any Φ, Ψ and m, the following holds.

(i) Let $M_{m,\Phi,\Psi}$ be invertible and let $M_{\overline{m},\Psi,\Phi}$ be well defined. Then $M_{\overline{m},\Psi,\Phi}$ is invertible and $M^{-1}_{\overline{m},\Psi,\Phi} = (M^{-1}_{m,\Phi,\Psi})^*$.

(ii) $M_{m,\Phi,\Psi}$ is unconditionally convergent and invertible $\Leftrightarrow$ $M_{\overline{m},\Psi,\Phi}$ is unconditionally convergent and invertible.

Below we consider a necessary condition for certain classes of multipliers to be both unconditionally convergent and invertible. Among these multipliers we consider the cases when Gabor or Wavelet frames are used. Consider the condition

$\mathcal{P}_3$: $\exists$ (c_n) and (d_n) so that $M_{m,\Phi,\Psi}$ can be written as $M_{(1),(c_n\phi_n),(d_n\psi_n)}$, where the summands are kept and $(c_n\phi_n)$ and $(d_n\psi_n)$ are frames for $\mathcal{H}$.

Proposition 7.2: [21] Let Φ and Ψ be Gabor (or wavelet) systems and let m satisfy $\inf_n |m_n| > 0$. If $M_{m,\Phi,\Psi}$ is unconditionally convergent and invertible, then $\mathcal{P}_3$ holds.

Proposition 7.3: [21] Let Φ be minimal. If $M_{m,\Phi,\Psi}$ is unconditionally convergent and invertible, then $\mathcal{P}_3$ holds.

Proposition 7.4: If $M_{m,\Phi,\Phi}$ is unconditionally convergent and invertible, then $(\sqrt{m_n}\phi_n)$ is a frame for $\mathcal{H}$ (where $\sqrt{m_n}$ denotes one (any one) of the two complex square roots of m_n, $n \in \mathbb{N}$) and $\mathcal{P}_3$ holds.

VIII. Conclusion

In this paper we have considered the invertibility of frame multipliers and reviewed analytical results.

In the future we will investigate the numerics of the inversion of multipliers (using the LTFAT toolbox [18]) and classify the cases when the inverse of a multiplier is again a multiplier, i.e. generalizations of Prop. 3.1.

Acknowledgment

The work on this paper was supported by the Austrian Science Fund (FWF) START-project FLAME ('Frames and Linear Operators for Acoustical Modeling and Parameter Estimation'; Y 551-N13). The authors thank to Dominik Bayer for a proofreading.

References

[1] S. T. Ali, J. P. Antoine, and J. P. Gazeau, *Continuous frames in Hilbert space*, Ann. Phys. **222** (1993), no. 1, 1–37.

[2] ———, *Coherent States, Wavelets and Their Generalization*, Graduate Texts in Contemporary Physics, Springer New York, 2000.

[3] P. Balazs, *Basic definition and properties of Bessel multipliers*, J. Math. Anal. Appl. **325** (2007), no. 1, 571–585.

[4] ———, *Hilbert-Schmidt operators and frames - classification, best approximation by multipliers and algorithms*, Int. J. Wavelets Multiresolut. Inf. Process. **6** (2008), no. 2, 315 – 330.

[5] ———, *Matrix-representation of operators using frames*, Sampl. Theory Signal Image Process. **7** (2008), no. 1, 39–54.

[6] P. Balazs, J. P. Antoine, and A. Grybos, *Weighted and controlled frames: Mutual relationship and first numerical properties*, Int. J. Wavelets Multiresolut. Inf. Process. **8** (2010), no. 1, 109–132.

[7] P. Balazs, D. Bayer, and A. Rahimi, *Multipliers for continuous frames in Hilbert spaces*, J. Phys. A: Mathematical and Theoretical **45** (2012), 244023.

[8] P. Balazs, B. Laback, G. Eckel, and W. A. Deutsch, *Time-frequency sparsity by removing perceptually irrelevant components using a simple model of simultaneous masking*, IEEE Trans Audio Speech Lang Processing **18** (2010), no. 1, 34–49.

[9] J. Benedetto and G. Pfander, *Frame expansions for Gabor multipliers*, Appl. Comput. Harmon. Anal. **20** (2006), no. 1, 26–40.

[10] P. Casazza, *The art of frame theory*, Taiwanese J. Math. **4** (2000), no. 2, 129–202.

[11] O. Christensen, *An Introduction to Frames and Riesz Bases*, Birkhäuser, Boston, 2003.

[12] M. Dörfler and B. Torrésani, *Representation of operators in the time-frequency domain and generalized Gabor multipliers*, J. Fourier Anal. Appl. **16** (2010), no. 2, 261–293.

[13] H. G. Feichtinger and K. Nowak, *A first survey of Gabor multipliers*, ch. 5, pp. 99–128, Birkhäuser, Boston, 2003.

[14] C. Heil, *A Basis Theory Primer*, Birkhäuser, Boston, 2011.

[15] P. Majdak, P. Balazs, W. Kreuzer, and M. Dörfler, *A time-frequency method for increasing the signal-to-noise ratio in system identification with exponential sweeps*, Proceedings of the 36th International Conference on Acoustics, Speech and Signal Processing, ICASSP 2011, Prag, 2011.

[16] G. Matz and F. Hlawatsch, *Linear time-frequency filters: On-line algorithms and applications*, ch. 6 in 'Application in Time-Frequency Signal Processing', pp. 205–271, eds. A. Papandreou-Suppappola, Boca Raton (FL): CRC Press, 2002.

[17] A. Rahimi and P. Balazs, *Multipliers for p-Bessel sequences in Banach spaces*, Integral Equations Oper. Theory **68** (2010), no. 2, 193–205.

[18] P. Soendergaard, B. Torrésani, and P. Balazs, *The linear time frequency analysis toolbox*, Int. J. Wavelets Multiresolut. Inf. Process. **10** (2012), no. 4, 1250032.

[19] D. T. Stoeva and P. Balazs, *Weighted frames and frame multipliers*, Annual of the University of Architecture, Civil Engineering and Geodesy XLIII-XLIV 2004-2009 (Fasc. II Mathematics Mechanics), 2012, 33–42.

[20] ———, *Invertibility of multipliers*, Appl. Comput. Harmon. Anal. **33** (2012), no. 2, 292–299.

[21] ———, *Canonical forms of unconditionally convergent multipliers*, J. Math. Anal. Appl. **399** (2013), 252–259.

[22] ———, *Detailed characterization of unconditional convergence and invertibility of multipliers*, Sampl. Theory Signal Image Process **12** (2013), no. 2, to appear.

[23] ———, *Riesz bases multipliers*, Proceedings IWOTA 2011 (Manuel Cepedello Boiso, Håkan Hedenmalm, Marinus A. Kaashoek, Alfonso Montes-Rodríguez, and Sergei R. Treil, eds.), Operator Theory: Advances and Applications **236** (2013), 477–484, Springer Basel AG.

[24] D. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, Wiley-IEEE Press, 2006.

Hybrid Regularization and Sparse Reconstruction of Imaging Mass Spectrometry Data

Andreas Bartels*, Dennis Trede*,†, Theodore Alexandrov*,†,‡ and Peter Maass*,†

*Center for Industrial Mathematics, University of Bremen, Bibliothekstr. 1, 28359 Bremen, bartels@math.uni-bremen.de.
†Steinbeis Innovation Center SCiLS Research, Richard-Dehmel-Str. 69, 28211 Bremen, Germany.
‡MALDI Imaging Lab, University of Bremen, Leobener Str., NW2, 28359 Bremen, Germany.

***Abstract*—Imaging mass spectrometry (IMS) is a technique to visualize the molecular distributions from biological samples without the need of chemical labels or antibodies. The underlying data is taken from a mass spectrometer that ionizes the sample on spots on a grid of a certain size. Mathematical postprocessing methods has been investigated twice for better visualization but also for reducing the huge amount of data. We propose a first model that applies compressed sensing to reduce the number of measurements needed in IMS. At the same time we apply peak picking in spectra using the ℓ_1-norm and denoising on the m/z-images via the TV-norm which are both general procedures of mass spectrometry data postprocessing, but always done separately and not simultaneous. This is realized by using a hybrid regularization approach for a sparse reconstruction of both the spectra and the images. We show reconstruction results for a given rat brain dataset in spectral and spatial domain.**

I. Introduction

A. Mass spectrometry

Mass spectrometry is a technique of analytical chemistry for the determination of the elemental composition of a biological or chemical sample. The way this task is accomplished is through experimental measurement of the mass-to-charge ratio of gas-phase ions produced from molecules from the underlying analyte.

As an example for a mass spectrometer we will now shortly describe the main principles of the so-called matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) mass spectrometer. In MALDI mass spectrometry the sample or compound to be analyzed is dissolved in a so-called matrix with crystallized molecules. Next, the ionization of the sample is triggered by intense laser pulses over a short duration. The ions are then accelerated by an electrostatic field. Since the velocity of the ions depends on the mass-to-charge ratio it is possible to measure the time-of-flight (TOF) to find the mass-to-charge ratio.

Most applications of mass spectrometry can be found in biology and medicine. But generally, mass spectrometry is not limited to the analysis of organic molecules. In principle any ionizable element can be analyzed with this technique.

B. Imaging mass spectrometry

The imaging mass spectrometry (IMS) is a technique used in mass spectrometry to visualize the spatial distribution of e.g. proteins or other chemical compounds. Given a thin sample, usually a tissue section, in MALDI-IMS mass spectra at discrete spatial points across the sample surface are acquired independently, providing a so-called datacube or hyperspectral image, with a mass spectrum measured at each pixel [1], see Figure 1. A mass spectrum represents the relative abundances of ionizable molecules with various mass-to-charge ratios (m/z), ranging for MALDI-IMS from several hundred up to a few tens of thousands m/z. A channel of a MALDI datacube corresponding to an m/z-value is called an m/z- or molecular image and expresses the relative spatial abundances of a molecular ion with this m/z-value. MALDI-IMS data

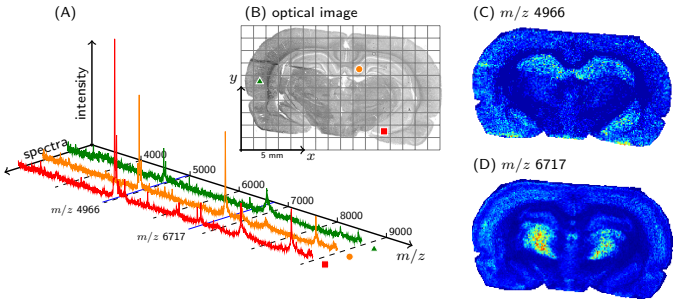


Fig. 1. Mass spectrometry data on an example of a rat brain tissue, taken from [2]. Each spot on the x, y-grid on the sample in (B) corresponds to one spectrum (A). Fixing an m/z-value yields to m/z-images representing the spatial distribution of the corresponding m/z-value, (C) and (D).

is large with a typical dataset comprising $10,000$ - $100,000$ spectra where an individual spectrum represents intensities measured at $10,000$ - $25,000$ m/z-bins.

In order to reduce the number of spectra required for reconstructing the hyperspectral IMS datacube we will make use of the compressed sensing (CS) idea: Instead of measuring spectra independently for each pixel we assume a setup that enables us to acquire some multiple sets of spectra at different points on the data which are then each summed up to a measurement-mean spectrum. Each measurement-mean spectrum then corresponds to one measurement and we would like to reconstruct the full dataset based on these measurements.

C. Compressed sensing and its applications to hyperspectral imaging

The combination of classical Shannon-Nyquist sampling and compression steps is one of the main ideas of CS. It turns

out that it is possible to represent or reconstruct given data with sampling rates much lower than the Nyquist rate [3], [4]. Mathematically spoken it means that given a signal, we do not need to acquire n periodic samples to get the discretized signal $x \in \mathbb{R}^n$. Instead, it suffices to take $m \ll n$ linear measurements $y_k \in \mathbb{R}$ using linear test functions $\varphi_k \in \mathbb{R}^n$, i.e. $y_k = \langle \varphi_k, x \rangle + z_k$ with some additive noise $z_k \in \mathbb{R}$. In matrix notation this reads

$$y = \Phi x + z,$$

where Φ is called the measurement matrix whose rows are filled with the linear functionals φ_k. Using the a-priori information that the signal x is S-sparse in a basis Ψ, i.e. $x = \Psi\lambda$ with $\|\lambda\|_{\ell_0} := |\mathrm{supp}(\lambda)| \leqslant S \ll n$, we are then interested in recovering the data x from only few taken measurements y. This can, e.g., be done with the basis pursuit approach, i.e. by solving the following convex optimization problem

$$\operatorname*{argmin}_{\lambda \in \mathbb{R}^n} \|\lambda\|_1 \quad \text{s.t.} \quad \|y - \Phi\Psi\lambda\|_2 \leqslant \varepsilon. \tag{1}$$

One of the many applications of CS can be found in hyperspectral imaging. A hardware realization of CS in that situation applying the single-pixel camera [5] has been studied in, for example, [6]. From the theoretical point of view mathematical models has been studied for CS reconstruction under certain priors [7]–[9]. Suppose that we have hyperspectral datacube $X \in \mathbb{R}^{n_x \times n_y \times c}$ whereas $n_x \times n_y$ denotes the spatial resolution of each image and c the number of channels. By concatenating each image as a vector we have $X \in \mathbb{R}^{n \times c}$ with $n := n_x \cdot n_y$. In the context of CS one aims to take $m \ll n$ measurements for each spectral channel $1 \leqslant j \leqslant c$ [8], [9] and to formulate a reconstruction strategy based on hyperspectral data priors. For example in [9] the authors assume the hyperspectral datacube to have low rank and piecewise constant channel images. Therefore the following convex optimization problem is presented

$$\operatorname*{argmin}_{\tilde{X} \in \mathbb{R}^{n \times c}} \|\tilde{X}\|_* + \lambda \sum_{j=1}^{c} \|\tilde{X}_j\|_{TV} \quad \text{s.t.} \quad \|Y - \Phi\tilde{X}\|_F \leqslant \varepsilon, \tag{2}$$

where $\|\cdot\|_*$ denotes the nuclear norm (i.e. the sum of the singular values), $\|\cdot\|_{TV}$ denotes the TV norm and the linear operator Φ is a measurement matrix as described above.

Another application of CS in hyperspectral imaging is the idea of calculating a compressed matrix factorization or a (blind) source separation of the data $X \in \mathbb{R}^{n \times c}$, i.e. $X = SH^T$, where $S \in \mathbb{R}^{n \times \rho}$ is a so called source matrix, $H \in \mathbb{R}^{c \times \rho}$ is a mixing matrix and ρ denotes the number of sources in the data which are supposed to be known. This model has been studied in the case of known mixing parameters H of the data X in [10] and with both matrices to be unknown in [7]. In case of the matrix H to be known and under the assumption that the columns of S are sparse or compressible in a basis Ψ, the problem being solved in [10] becomes

$$\operatorname*{argmin}_{\lambda \in \mathbb{R}^{\rho n}} \|\lambda\|_1 \quad \text{s.t.} \quad \|Y - \Phi\bar{H}\Psi\lambda\|_2 \leqslant \varepsilon, \tag{3}$$

where $\bar{H} = H \otimes I_n$ and $\otimes$ denotes the usual matrix Kronecker product and I_n the $n \times n$ identity matrix. The authors in [10] also studied the case where the ℓ_1-norm in (3) is replaced by the TV-norm with respect to the columns of S, i.e. $\sum_{j=1}^{\rho} \|S_j\|_{TV}$. However, as a result of (3) one has a decomposition of the data X as in two matrices S and H where the columns of S contain of the ρ most representative images of the hyperspectral datacube and those of H of the corresponding pseudo spectra.

In this work we investigate a hybrid reconstruction model for hyperspectral data similar to (2) and (3), but with special motivation for imaging mass spectrometry data $X \in \mathbb{R}_+^{n \times c}$ and formulated as a Tikhonov functional:

$$\operatorname*{argmin}_{\tilde{X} \in \mathbb{R}_+^{n \times c}} \frac{1}{2}\|Y - \mathcal{D}_{\Phi,\Psi}(\tilde{X})\|_F + \alpha \sum_{j=1}^{c} \|\tilde{X}_j\|_{TV} + \beta\|\tilde{X}\|_1. \tag{4}$$

Furthermore, we are interested in reconstructing the full dataset $X \in \mathbb{R}_+^{n \times c}$ while extracting its main features.

Since we know a-priori that mass spectra in IMS are typically nearly sparse or compressible we use the ℓ_1-norm as one regularization term. The second, i.e. the TV-term, comes from the fact that the m/z-images are supposed to have sparse image gradients.

II. Compressed sensing model for imaging mass spectrometry

A. Imaging mass spectrometry data

IMS data is a hyperspectral datacube $X \in \mathbb{R}_+^{n_x \times n_y \times c}$ with c channels and m/z-images $X_{(\cdot,\cdot,k)} \in \mathbb{R}_+^{n_x \times n_y}$ for $k = 1, \ldots, c$. By concatenating each image as a vector, the hyperspectral data becomes a matrix $X \in \mathbb{R}_+^{n \times c}$ where $n := n_x \cdot n_y$.

B. The compressed sensing process

A part of the measurement process in IMS consists of the ionization of the given sample. In MALDI-MS, for instance, the tissue is ionized by a laser beam, which shoots on each of the n pixel of a predefined grid. This yields n independently measured spectra. In order to reduce the number of spectra needed for the reconstruction we make use of CS [4], [11].

In the context of compressed sensing, each entry y_{ij} of the measurement vectors $y_i \in \mathbb{R}^c$ for $i = 1, \ldots, m$ and $j = 1, \ldots, c$ is the result of an inner product between the data $X \in \mathbb{R}_+^{n \times c}$ and a test function $\varphi_i \in \mathbb{R}^n$ with components φ_{ik}, i.e.

$$y_{ij} = \langle \varphi_i, X_{(\cdot,j)} \rangle. \tag{5}$$

The results y_i for $i = 1, \ldots, m$ are in our IMS context so called *measurement-mean spectra* since they are calculated by the mean intensities on each channel. This can be seen by rewriting (5) as

$$y_i^T = \varphi_i^T X = \sum_{k=1}^{n} \varphi_{ik} X_{(k,\cdot)}, \tag{6}$$

which directly shows that the measurement vectors y_i^T are linear combinations of the original spectra $X_{(k,\cdot)}$.

We are looking for a reconstruction of the data X based on m measurement-mean spectra, each measured by one linear function φ_i. In matrix form (5) becomes

$$Y = \Phi X \in \mathbb{R}_+^{m\times c}, \tag{7}$$

where $\Phi \in \mathbb{R}^{m\times n}$ is the measurement matrix. By incorporating noise $Z \in \mathbb{R}_+^{m\times c}$ that arises during the mass spectrometry measurement process, (7) becomes

$$Y = \Phi X + Z \in \mathbb{R}_+^{m\times c}, \tag{8}$$

at which $\|Z\|_F \leqslant \varepsilon$. By this we explicitly assume to have inherent Gaussian noise in the data and we will keep this for the rest of our analysis.

C. First assumption: compressible spectra

Within IMS data acquisition process for each pixel on the sample we gain a mass spectrum whose entries can be seen as positive real numbers, i.e. $X_{(k,\cdot)} \in \mathbb{R}_+^c, \quad k = 1,\dots,n$. As our first assumption we take into account that we know that these spectra are sparse or compressible in spectral domain. Therefore we assume that these spectra are well presented by a suitable choice of functions $\psi_i \in \mathbb{R}_+^c$ for $i = 1,\dots,c$. More precisely this means that there exists a matrix $\Psi \in \mathbb{R}_+^{c\times c}$ such that for each spectrum $X_{(k,\cdot)}$ there is a coefficient vector $\lambda_k \in \mathbb{R}_+^c$ with $\|\lambda_k\|_0 \ll c$, such that $X_{(k,\cdot)}^T = \Psi\lambda_k$. In matrix form this sparsity property can be written as

$$X^T = \Psi\Lambda, \tag{9}$$

where $\Lambda \in \mathbb{R}_+^{c\times n}$ is the coefficient matrix or *feature matrix*, see Figure 2. In light of the sparse spectra, our aim should be to minimize the columns $\Lambda_{(\cdot,i)}$ of Λ with respect to the l_0 "norm", i.e.

$$\sum_{i=1}^{n} \|\Lambda_{(\cdot,i)}\|_0. \tag{10}$$

Putting (8) and (9) together leads to

$$Y = \Phi\Lambda^T\Psi^T + Z. \tag{11}$$

D. Second assumption: sparse image gradients

By keeping one m/z-value $i_0 \in \{1,\dots,c\}$ fixed we get an m/z-image $X_{(\cdot,i_0)} \in \mathbb{R}_+^n$ (one column of the data X) that represents the spatial distribution of the fixed mass m_0 in the measured biological sample. Another a priori knowledge takes into account the sparsity of these m/z-images with respect to their gradient. Besides this, we are also aware of the large variance of noise variance in each m/z-image. To handle both, we want to make use of the total variation (TV) model introduced by Rudin, Osher and Fatemi [12]. So as a second statement, we want each m/z-image to be minimized with respect to its TV semi-norm. By taking into account the coefficient matrix Λ in (9), it arises to minimize its rows $\Lambda_{(i,\cdot)}$ for $i = 1,\dots,c$ since each of them corresponds to an m/z-image, i.e.

$$\sum_{i=1}^{c} \|\Lambda_{(i,\cdot)}\|_{TV}. \tag{12}$$

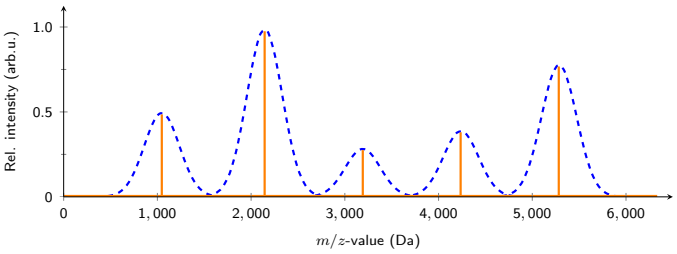


Fig. 2. Illustration of peak picking approach in mass spectrometry. Instead of finding a minimizer Λ and multiply it with a convolution operator Ψ, we aim to recover the features $\tilde{\Lambda}$. Dashed line (- - -): Reconstruction of the i-th spectrum, i.e. $\tilde{X}_{(\cdot,i)}^T = (\Psi\tilde{\Lambda})_{(\cdot,i)}$. Solid line (—): Only the main features of the i-th spectrum $\tilde{\Lambda}_{(\cdot,i)}$, i.e. the main peaks, are extracted.

As in the first assumption and also explained in Figure 2, we aim to extract the main features such as the main peaks in the data. For incorporating also the spatial domain information in each channel, we again only use the coefficient matrix in (12).

E. The final model

Putting it all together, we are now able to formulate our model for CS in IMS. Since minimizing with respect to the ℓ_0 "norm" is NP-hard, it is common to obviate this by replacing it with the ℓ_1-norm. Our minimization problem then finally becomes

$$\underset{\Lambda\in\mathbb{R}_+^{c\times n}}{\operatorname{argmin}} \frac{1}{2}\|Y - \Phi\Lambda^T\Psi^T\|_F^2 + \alpha\sum_{i=1}^{c}\|\Lambda_{(i,\cdot)}\|_{TV} + \beta\sum_{i=1}^{n}\|\Lambda_{(\cdot,i)}\|_1. \tag{13}$$

III. Numerical results

The algorithm we are using to solve (13) is based on the parallelizable primal-dual splitting algorithm presented in [13].

The test data $X \in \mathbb{R}_+^{n\times c}$ is made of a well-studied rat brain coronal section [2] (see Figure 1) which consists of $c = 2,000$ channels with m/z-images of spatial resolution 121×202 and therefore $n = 24,442$ pixel. The spectra were normalized using total ion count (TIC) normalization which is nothing else than a division by the ℓ_1-norm of each spectrum. Furthermore, they were baseline-corrected using the TopHat algorithm with a minimal baseline with set to 10%. We assume the mass spectra to be sparse with respect to shifted Gaussians [14]

$$\psi_k(x) = \frac{1}{\pi^{1/4}\sigma^{1/2}}\exp\left(-\frac{(x-k)^2}{2\sigma^2}\right). \tag{14}$$

In (14), the variance should be set data dependent [15]. The measurement matrix Φ is randomly filled with numbers from an i.i.d. Gaussian distribution with zero mean and variance one. For our results we have further set the regularization parameters in (13) by hand as $\alpha = 1.6$ and $\beta = 1.4$ and applied 100 iterations.

First we present the mean spectrum, i.e. the sum over all pixel spectra $\Lambda_{(i,\cdot)}$ for $i = 1,\dots,n$ of the rat brain data as well as the mean spectrum of the reconstructed datacube, see Figure 4. The reconstruction is based on 40% taken measurements. We can see is that the main peaks from the

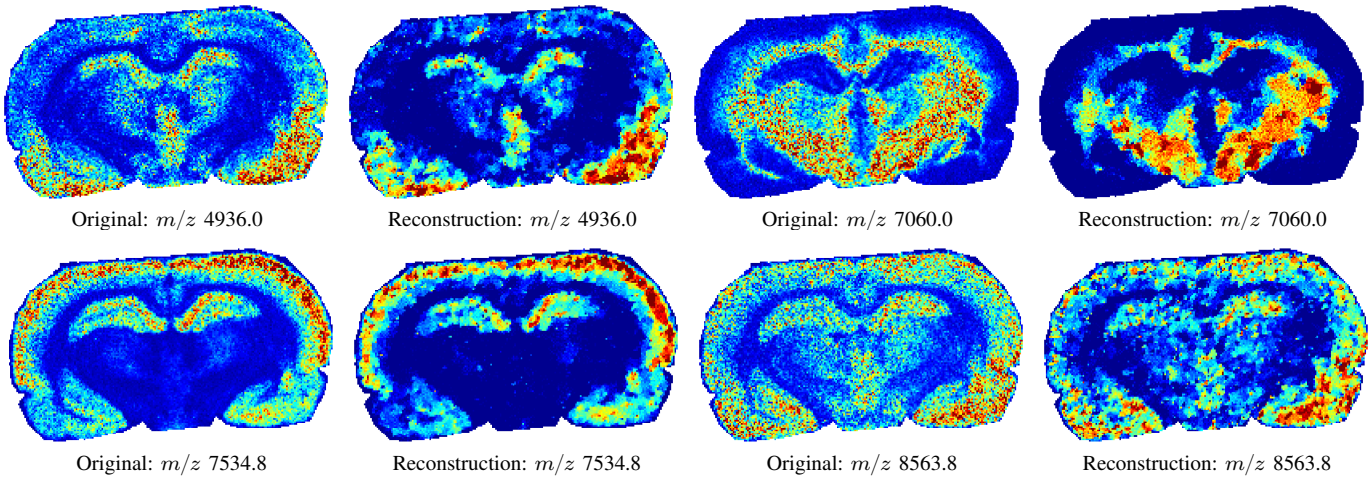


Fig. 3. Reconstruction results of four m/z-images based on 40% taken measurements of spectra.

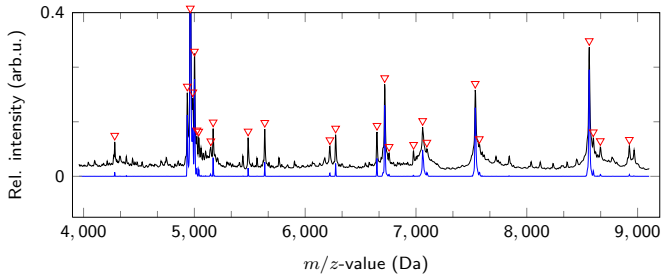


Fig. 4. Mean spectrum from the rat brain dataset (black) and its compressed reconstruction (blue), based on 40% taken measurements. As part of the reconstruction, also main peaks within the spectrum were detected (triangles).

original mean spectrum are recovered while the rest of the m/z-values are set to zero. At second we show four different m/z-images which all belong to a peak in the mean spectrum in Figure 4. We clearly see the influence of both the TV and the ℓ_1 penalty term. Where there are regions of high intensity pixels, total varations effects smoothing those ones while preserving the edges. In addition we see that due to ℓ_1 minimization other non-high intensity pixels were set to (nearly) zero.

IV. Conclusion

We have proposed a first CS model for imaging mass spectrometry. While reconstructing the data from fewer measurements we apply peak picking in mass spectra as well as TV-denoising on the m/z-images at the same time. Our results look promising and motivates for further research in this direction. Future work might be done by replacing the Gaussian noise model with a Poisson statistics approach which is supposed to be more suitable for MALDI-TOF spectrometry [15].

Acknowledgment

The authors would like to thank Michael Becker (Bruker Daltonik) for providing the rat brain dataset.

References

[1] R. M. Caprioli, T. B. Farmer, and J. Gile, "Molecular imaging of biological samples: localization of peptides and proteins using MALDI-TOF MS," *Anal Chem*, vol. 23, no. 69, pp. 4751–4760, 1997.

[2] T. Alexandrov and J. H. Kobarg, "Efficient spatial segmentation of large imaging mass spectrometry datasets with spatially aware clustering," *Bioinformatics*, vol. 27, no. ISMB 2011, pp. i230–i238, 2011.

[3] E. Candès, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Commun. Pure Appl. Math.*, vol. 59, no. 8, pp. 1207–1223, 2006.

[4] D. Donoho, "Compressed sensing," *IEEE Trans. Inform. Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[5] D. Takhar, J. Laska, M. Wakin, M. Duarte, D. Baron, S. Sarvotham, K. Kelly, and R. Baraniuk, "A new compressive imaging camera architecture using optical-domain compression," *Computational Imaging IV*, vol. 6065, pp. 43–52, 2006.

[6] T. Sun and K. F. Kelly, "Compressive sensing hyperspectral imager," *Comp. Optical Sensing and Imaging*, 2009.

[7] M. Golbabaee, S. Arberet, and P. Vandergheynst, "Distributed compressed sensing of hyperspectral images via blind source separation," *Presentation given at Asilomar conf. on signals, systems, and computers, Pacific Groove, CA, USA, November 7-10*, 2010.

[8] ——, "Multichannel compressed sensing via source separation for hyperspectral images," *Eusipco 2010, Aalborg, Denmark*, 2010.

[9] M. Golbabaee and P. Vandergheynst, "Joint trace/TV norm minimization: A new efficient approach for spectral compressive imaging," *IEEE Int. Conf. on Img. Proc (ICIP)*.

[10] M. Golbabaee, S. Arberet, and P. Vandergheynst, "Compressive source separation: Theory and methods for hyperspectral imaging," 2012, submitted.

[11] E. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inform. Theory*, vol. 52, no. 2, pp. 489–509, 2006.

[12] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D: Nonlinear Phenomena*, vol. 60, pp. 259–268, 1992.

[13] P. L. Combettes and J.-C. Pesquet, "Primal-dual splitting algorithm for solving inclusions with mixtures of composite, Lipschitzian, and parallel-sum type monotone operators," *Set-Valued and Variational Analysis*, vol. 20, no. 2, pp. 307–330, 2012.

[14] L. Denis, D. A. Lorenz, and D. Trede, "Greedy solution of ill-posed problems: error bounds and exact inversion," *Inverse Problems*, vol. 25, no. 11, pp. 1–24, 2009.

[15] T. Alexandrov, M. Becker, S.-O. Deininger, G. Ernst, L. Wehder, M. Grasmair, F. v. Eggling, H. Thile, and P. Maass, "Spatial segmentation of imaging mass spectrometry data with edge-preserving image denoising and clustering," *J. Proteome Research*, vol. 9, no. 12, pp. 6535–6546, 2010.

Level crossing sampling of strongly monoHölder functions

Brigitte Bidegaray-Fesquet
and Marianne Clausel
Univ. Grenoble Alpes, LJK, BP 53, 38041 Grenoble Cedex 9, FRANCE
Email: Brigitte.Bidegaray@imag.fr and Marianne.Clausel@imag.fr

***Abstract*—We address the problem of quantifying the number of samples that can be obtained through a level crossing sampling procedure for applications to mobile systems. We specially investigate the link between the smoothness properties of the signal and the number of samples, both from a theoretical and a numerical point of view.**

I. Introduction

An important issue in the design of mobile systems is to increase their autonomy and/or reduce their size and weight. This can be achieved by reducing their power consumption by processing signal with a smaller number of samples. For a large class of signals, especially sporadic signals, non-uniform sampling leads to a reduced number of samples, compared to a Nyquist sampling [7], [9], [10], [13]. A way to obtain such samples is to use specific system architectures (e.g. event-driven). These architectures take samples each time some specific event occurs, typically specific voltage levels are crossed. We can therefore design simple analog circuits, with low power consumption, to acquire information, possibly at high speed. Here we consider a system where amplitudes are selected thanks a M-bit asynchronous analog-to-digital converter (AADC) and 2^M levels are predefined in the voltage range.

In this article we want to understand on which signal characteristics the number of samples depend. An intuitive look at the problem indicates that the more the signal is oscillating locally the higher the number of samples is. The number of samples at the neighborhood of some point may then be related to the *local smoothness* of the signal, that is to its so–called *Hölder regularity*. To put in evidence this relationship, we consider toy models of signals whose smoothness properties are perfectly known at each point. We then perform numerical simulations and link the sample reduction rate with this regularity. The next step, which will be the purpose of a forthcoming paper, will then be to consider signals whose regularity may change from point to point such as multifractional or multifractal signals. We then intend to apply our results to biological signals such as EEG signals or fMRI data which are well-known to be both highly irregular and non stationary signals, and which provide interesting ranges of application for non-uniform signal processing.

II. Algorithm

In the event-driven systems, the signals are not sampled at totally arbitrary times. Indeed there are local clocks that measure the time elapsed since the previous sample was taken. Therefore we can consider that the samples are taken at some multiples of some basis time t_b. The mathematical algorithm that is used to mimic the AADC is the following:

- **Step 1:** generate uniform samples on $[0,1]$ with sampling interval t_b;
- **Step 2:** for each sample replace the amplitude by the value of the level just below;
- **Step 3:** decimate the samples so as to keep only one (the last) sample when consecutive samples have the same amplitude.

III. Mathematical interpretation

Up to some time re-scaling we suppose that the precision of the local clock that measures time delays is $t_b = 2^{-j}$, for $j \in \mathbb{N}$. At best we only know the function f by its samples at times $k2^{-j}$, $k \in \mathbb{Z}$. At scale 2^{-j}, we define the intervals

$$I_{j,k} = [k2^{-j}, (k+1)2^{-j}[.$$

A. Faber–Schauder hierarchical basis

We define the Faber–Schauder hierarchical basis as defined in [5]. Let V_j be the space of continuous functions, which are affine on intervals $I_{j,k}$, $k \in \mathbb{Z}$. We can uniquely define the linear interpolation f_j of f at scale 2^{-j} by $f_j(k2^{-j}) = f(k2^{-j})$, for all $k \in \mathbb{Z}$. Let $\varphi(x) = \max\{0, 1-|x|\}$. A natural basis for V_j is given by the functions $\varphi_{j,k} = \varphi(2^j \cdot - k)$, $k \in \mathbb{Z}$. In this basis, we have the unique representation

$$f_j = \sum_{k\in\mathbb{Z}} f(k2^{-j})\varphi_{j,k}.$$

B. Interpretation

We now suppose that f is compactly supported in $[0,1]$. In the previous notations we will only need $k = 0, \dots, 2^j - 1$. We assume that levels are uniformly spaced by some quantum 2^{-M}. The level crossing algorithm can be described as follows.

Step 1: approximation in V_j. We only know $f_j(k2^{-j}) = f(k2^{-j})$.

Step 2: level crossing. We denote $\lfloor x \rfloor$ the integer part of x, namely $\lfloor x \rfloor = \inf\{n \in \mathbb{N},\, x \leq n\}$. We then define

$$\tilde{f}_j = \sum_{k \in \mathbb{Z}} 2^{-M} \left\lfloor 2^M f(k2^{-j}) \right\rfloor \varphi_{j,k}.$$

Step 3: decimation of $\tilde{f}_j$. We only keep a subsequence of $k = 0, \ldots, 2^j - 1$, defined by induction: $k_0 = 0$ and

$$k_{i+1} = \min\{k \geq 1 + k_i / \left\lfloor 2^M f(k2^{-j}) \right\rfloor \neq \left\lfloor 2^M f(k_i 2^{-j}) \right\rfloor\}.$$

To be able to reconstruct $\tilde{f}_j$, we only store the couples $(\delta t_i, a_i)$ where $\delta t_i = (k_i - k_{i-1})2^{-j}$, $i \geq 1$ is the delay since the last sample, and $a_i = 2^{-M} \left\lfloor 2^M f(k_i 2^{-j}) \right\rfloor$ is the amplitude of the sample. There is no approximation in Step 3, we only do not store redundant data.

IV. Mathematical properties

We now want to illustrate through numerical experiments that the properties of our algorithm can be related to smoothness properties of the sampled signal.

A. MonoHölderian functions

Definition 1 (Hölder space $\mathcal{C}^\alpha$): Let $\alpha \in (0,1)$. The function f belongs to the Hölder space $\mathcal{C}^\alpha([0,1])$ if there exists a constant C such that for all $(x,y) \in [0,1]^2$,

$$|f(x) - f(y)| \leq C|x-y|^\alpha \,.$$

The following definition has been introduced in [12].

Definition 2 (Anti-Hölderian functions): Let $\alpha \in (0,1)$. The function f is said to be uniformly anti-Hölderian with exponent α, if there exists a constant C such that for all $(x,y) \in [0,1]^2$,

$$\sup_{(u,v) \in [x,y]^2} |f(u) - f(v)| \geq C|x-y|^\alpha \,.$$

The set of uniformly anti-Hölderian functions is denoted $I^\alpha([0,1])$.

Definition 3: Let $\alpha \in (0,1)$. If the function f both belongs to $\mathcal{C}^\alpha([0,1])$ and $I^\alpha([0,1])$ then f is said to be monoHölderian with exponent α.

B. Approximation properties

As already mentioned, only step 1 and 2 lead to approximations. If $f \in \mathcal{C}^\alpha([0,1])$, it is well–known [6], [11] that there exists a constant C (which depends on f but not on the scale j) such that

$$\|f - f_j\|_{L^\infty} \leq C 2^{-j\alpha},$$

whereas, if the function f is assumed to be uniformly monoHölderian, one deduces from [4] that there exists a constant C (which depends on f but not on the scale j) such that for any $\epsilon > 0$

$$\|f - f_j\|_{L^\infty} \geq C j^{-(2\alpha+\epsilon)} 2^{-j\alpha}.$$

Note that the last condition is much weaker than uniform anti–Hölderianity (see [4]) since it involves the modulus of continuity of $f - f_j$ on the whole interval $[0,1]$, whereas oscillations of uniformly anti-Hölderian functions can be bounded from below at any point. The approximation made at step 2 clearly does not depend on the regularity of function f, and we have

$$\|f_j - \tilde{f}_j\|_{L^\infty} \leq 2^{-M}.$$

C. Theoretical number of samples in the case of a monotonous function

If f is a monoHölderian function with exponent α, by definition there exists $C_1, C_2 > 0$ and for any scale $j \geq 0$ and $0 \leq k \leq 2^j - 1$

$$C_1 2^{-j\alpha} \leq \sup_{(u,v) \in [k/2^j, (k+1)/2^j]^2} |f(u) - f(v)| \leq C_2 2^{-j\alpha}.$$

If the function is additionally supposed to be monotonous, we have further that

$$\sup_{(u,v) \in [k/2^j, (k+1)/2^j]^2} |f(u) - f(v)| = |f((k+1)/2^j) - f(k/2^j)|.$$

Hence

$$\begin{aligned} C_1 2^{j(1-\alpha)} &\leq |f(1) - f(0)| \\ &= \sum_{k=0}^{2^j-1} |f(\frac{k+1}{2^j}) - f(\frac{k}{2^j})| \leq C_2 2^{j(1-\alpha)}. \end{aligned}$$

Such a signal crosses equi-spaced levels with quantum 2^{-M} at most $C2^{M+(1-\alpha)j}$ times. With our algorithm, we also take at most 2^j samples (since we use the initial samples). For large values of M (or small values of α), we indeed keep almost all the 2^j samples. Otherwise we can expect some reduction of the number of samples. For $C = 1$, the threshold is $M \simeq \alpha j$. Observe that the proof is based on the fact, that in the monotonous, we can estimate in a very simple way the oscillations

$$\sup_{(u,v) \in [k/2^j, (k+1)/2^j]^2} |f(u) - f(v)|$$

of the function. Of course in the general case, the situation can be much more complicated. Nevertheless, generic results in the sense of prevalence as stated in [2] are expected to hold. In what follows, we illustrate through numerical simulations what may happens.

D. Numerical simulations

1) Test functions: We test level crossing on two toy models: sample paths of fractional Brownian motion B_H and the Weierstrass function W_H. Here $H \in]0,1[$ is called the Hurst index. In each of these two cases, the smoothness properties of the function are well–known and related to the Hurst index.

The fractional Brownian motion (fBm) B_H is the unique Gaussian H–self-similar process with stationary increments. It can be defined from its covariance function

$$\mathbb{E}\big[B_H(x) B_H(y)\big] = \tfrac{1}{2}\left(|x|^{2H} + |y|^{2H} - |x-y|^{2H}\right)$$

for all $(x,y) \in [0,1]^2$. For $H = 1/2$, we recover the classical Brownian motion. We recall that the sample paths of fBm are well-known to be almost surely continuous. Further, the Hurst index H of fBm is directly related to the roughness of its sample paths. More precisely, almost surely,

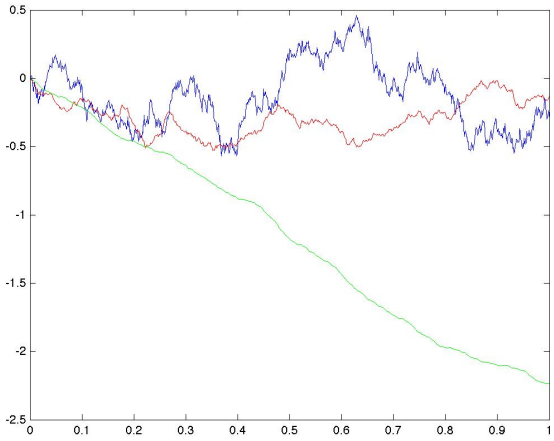


Fig. 1. Three realizations of fractional Brownian motions for $H = 0.5$ (blue plot), $H = 0.7$ (red plot), $H = 0.9$ (green plot)

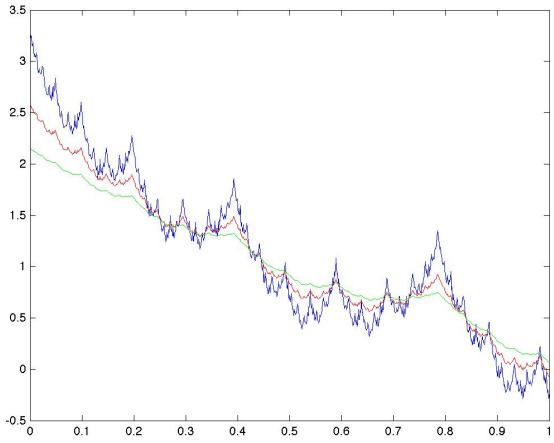


Fig. 2. The Weierstrass function for $H = 0.5$ (blue plot), $H = 0.7$ (red plot), $H = 0.9$ (green plot)

$B_H \in \mathcal{C}^{H-\varepsilon}([0,1]) \cap I^{H+\varepsilon}([0,1])$ (classical law of the iterated logarithm). Roughly speaking, a.s. for all $(x,y) \in [0,1]^2$, $\sup_{(u,v)\in[x,y]^2} |B_H(u) - B_H(v)| \sim |x-y|^H$. Figure 1 presents three realizations of sample paths of fractional Brownian motions for $H = 0.5$, $H = 0.7$, $H = 0.9$ and 1024 samples ($j = 10$).

We also use the Weierstrass function defined as

$$W_H(x) = \sum_{j=0}^{\infty} 2^{-jH} \cos(2^j x).$$

The Weierstrass function W_H is a classical example of monoHölderian function with exponent H as proved in [8]. Hence for all $(x,y) \in [0,1]^2$, $\sup_{(u,v)\in[x,y]^2} |W_H(u) - W_H(v)| \sim |x-y|^H$. Figure 2 present some graphs of Weierstrass functions for $H = 0.5$, $H = 0.7$, $H = 0.9$ and 1024 samples ($j = 10$).

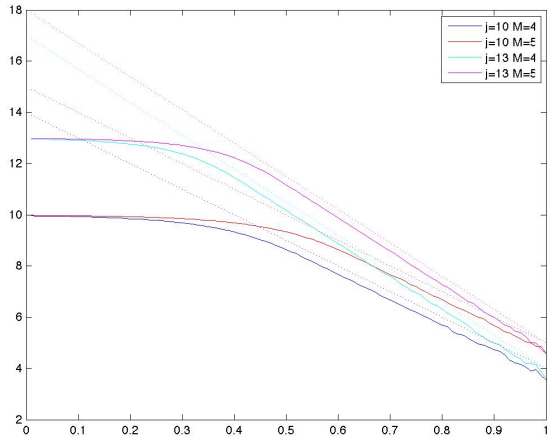


Fig. 3. Average number n of samples in terms of the Hurst number in the log scale ($\log_2(n)$ is represented on the y-axis). Four cases are plotted (solid lines) corresponding to $j = 10$ and 13 and $M = 4$ and 5. The dotted lines correspond to the worst case $M + (1-H)j$.

	$M = 4$	$M = 5$
$j = 10$	0.4	0.5
$j = 13$	~ 0.3	~ 0.4

TABLE I
"CRITICAL" VALUES OF THE HURST NUMBER.

2) Test cases: The tests are performed within the SPASS Matlab toolbox [1]. To generate fractional Brownian motions, we make use of the `genFBMJFC.m` function [3].

We use two values of j (10 and 13) and two values of M (4 and 5). These small values of M are sufficient for most mobile applications. Our output is the number of samples after decimation (Step 3). For the fractional Brownian motion, we perform 1000 realizations and average the number of samples obtained for each realization to obtain an average number n.

We perform this for values of the Hurst number H in the $]0,1[$ range and obtain the curves in Figure 3. We also plot the number of samples computed in the worst case (monotonous function i.e. maximum total variation) for $C = 1$: $2^{M+(1-H)j}$. The plots are given in the semi-log scale: $\log_2(n)$ and $M + (1-H)j$. This allows to distinguish the two regimes below some value of the Hurst number $H \sim M/j$ the algorithm more or less keeps all the original samples, above this value the decimation is efficient and yields a significant reduction of the number of samples.

For the different curves these "critical" values of H are given in Table I.

We perform the same tests on the Weierstrass function. The plots associated to fBm are much more regular because there are obtained by an averaging procedure. We expect that the critical value of M holds in an asymptotical way. Our results are then expected to improve when j tends to infinity.

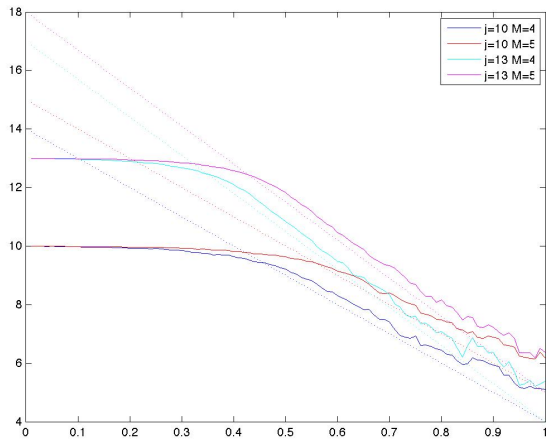


Fig. 4. Number of samples in terms of the Hurst number in the log scale ($\log_2(n)$ is represented on the y-axis). Four cases are plotted (solid lines) corresponding to $j = 10$ and 13 and $M = 4$ and 5. The dotted lines correspond to the worst case $M + (1 - H)j$.

V. Conclusion

We have first shown numerically that there is strong relationship between the smoothness properties of a signal and the number of samples that can be obtained by the crossing level algorithm presented in this paper. We also proved this result in the case of monotonous monoHölderian functions. We intent to address this problem in more general cases. It will be the purpose of a forthcoming paper.

Acknowledgment

This work has been supported by the MathSTIC Project OASIS of the Grenoble University. The authors want to thank Jean-François Coeurjolly for fruitful discussions.

References

[1] B. Bidégaray-Fesquet and L. Fesquet, *Signal Processing for ASynchronous Systems toolbox,* Matlab toolbox, 2011.
[2] M. Clausel and S. Nicolay, *Some prevalent results about strongly monoHölder functions,* Nonlinearity, **23**(9), 2101–2116, 2010.
[3] J.-F. Coeurjolly, *Simulation and identification of the fractional Brownian motion: a bibliographical and comparative study,* Journal of Statistical Software, **5**(7), 1–53, 2000.
[4] M. Clausel, S. Nicolay, *A wavelet characterization for the upper global Hölder index.*, to appear in J. Fourier Anal. Appl.2013.
[5] A. Cohen, *Wavelet Methods in Numerical analysis,* Handbook of Numerical Analysis, Vol. VII, Elsevier, 2000.
[6] R.A. DeVore, B. Jawerth, B.J. Lucier, *Image Compression Through Wavelet Transform Coding*, IEEE Trans. Inf. Theory **38**(2):719–746, 1992.
[7] K. Guan and A.C. Singer, *Opportunistic sampling by level-crossing,* In IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2007, pages 1513–1516, Honolulu, Hawai'i, USA, April 2007. IEEE.
[8] G.H. Hardy, *Weierstrass's non differentiable function,* Trans. Amer. Math. Soc., **17**, pp.301–325 (1916).
[9] J.W. Mark and T.D. Todd, *A nonuniform sampling approach to data compression,* IEEE Trans. Commun., **29**(1):24–32, 1981.
[10] F.A. Marvasti, *Nonuniform Sampling. Theory and Practice,* Information Technology: Transmission, Processing and Storage. Springer, 2001.
[11] Y. Meyer, *Ondelettes et opérateurs*, Hermann, 1990.
[12] C. Tricot, *Courbes et dimension fractale*, Springer (1992).
[13] N. Sayiner, H.V. Sorensen, and T.R. Viswanathan, *A level-crossing sampling scheme for A/D conversion,* IEEE Trans. Circ. Syst. II, **43**(4):335–339, 1996.

MAP Estimators for Self-Similar Sparse Stochastic Models

Emrah Bostan, Julien Fageot, Ulugbek S. Kamilov and Michael Unser

Biomedical Imaging Group, EPFL, Lausanne, Switzerland

***Abstract*—We consider the reconstruction of multi-dimensional signals from noisy samples. The problem is formulated within the framework of the theory of continuous-domain sparse stochastic processes. In particular, we study the fractional Laplacian as the whitening operator specifying the correlation structure of the model. We then derive a class of MAP estimators where the priors are confined to the family of infinitely divisible distributions. Finally, we provide simulations where the derived estimators are compared against total-variation (TV) denoising.**

***Index Terms*—Innovation models, fractional Laplacian, fractals, invariance, self-similarity, sparse stochastic processes, MAP estimation.**

I. Introduction

Consider the signal denoising problem where the goal is to estimate the unknown signal $\mathbf{s} \in \mathbb{R}^K$ from the noisy measurement

$$\mathbf{y} = \mathbf{s} + \mathbf{n}, \tag{1}$$

where the vector $\mathbf{n} \in \mathbb{R}^K$ represents the noise that is assumed to be i.i.d. Gaussian with variance σ^2.

We consider the statistical formulation of the denoising problem based on the prior knowledge of the distribution of the signal and concentrate on MAP estimators. To that end, we first specify a *continuous-domain* signal model by using the theory of sparse stochastic processes [1]. The model has two fundamental elements: an *innovation process* governing the sparsity pattern and the *whitening operator* determining the correlation structure of the underlying signal.

The contribution of this work is to extend our previous line of work [2], [3] by using fractional-order Laplacians $(-\Delta)^{\gamma/2}$ with $\gamma > 0$ as our whitening operator. The unique feature of these operators is their invariance to translation, scaling, and rotation [4]. They also have been associated with $1/\|\omega\|^{\gamma}$-type power spectrum that appears in natural images [5], [6]. In this prospective, the derived estimators are suitable for removing noise from fractal-like images. We perform simulations and show that the derived estimators can improve upon TV denoising for particular images.

II. Mathematical foundations

We assume that the underlying signal $\mathbf{s}$ is the discretized version of a stochastic process $s(\mathbf{r})$ in $\mathbb{R}^d$ that is defined as the solution of the stochastic differential equation

$$\mathrm{L}s = w, \tag{2}$$

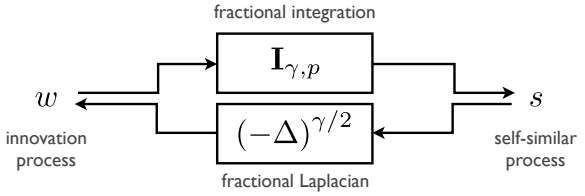


Fig. 1. Continuous-domain innovation model.

where w is a continuous-domain white noise that is *not necessarily Gaussian* and L is a suitable differential operator. What (2) implies is that the formal solution (if it exists) is given by $s = \mathrm{L}^{-1}w$. Therefore, the correlation structure of s is determined by the mixing operator L^{-1}, while its sparsity pattern is characterized by w that we shall call the *innovation process* (see Section II-A).

In the sequel, we restrict L to be in the subclass of *fractional Laplacians* (Section II-B). Our goal is to define a general class of *self-similar processes* (Section II-C) as illustrated in Figure 1.

A. Innovation processes

We define continuous-domain innovation processes in the framework of generalized functions of Schwartz [7]. In the one-dimensional setting, they are the weak derivative of the family of Lévy processes. As a member of the family of generalized stochastic processes, w is a random generalized function that is observed through scalar-products with test functions φ in the space $\mathcal{S}$ of smooth and rapidly decreasing functions. Hence, for fixed φ, the linear observation $\langle w, \varphi \rangle$ is a real random variable.

The innovation process w is a stationary stochastic process with independent value at every point. Its statistical properties are characterized by its characteristic functional (the infinite-dimensional generalization of the characteristic function of random variables) $\widehat{\mathscr{P}}_w(\varphi) = \mathbb{E}[e^{j\langle w,\varphi\rangle}]$. The characteristic functional of w has the general form

$$\widehat{\mathscr{P}}_w(\varphi) = \exp\left(\int_{\mathbb{R}^d} f(\varphi(\mathbf{r}))d\mathbf{r}\right), \tag{3}$$

where $f(\cdot)$ is called the Lévy exponent of w. The set of admissible Lévy exponents, and thus of innovation processes, is described in [7].

It is important to note that Lévy exponents are also in one-to-one correspondence with the so-called infinitely divisible

(i.d.) distributions. Indeed, the characteristic functions of i.d. random variables are precisely of the form $\mathrm{e}^{\mathrm{f}(\omega)}$ [8]. Equivalently, an innovation process is characterized by its canonical pdf defined by $p_{\mathrm{i.d.}} = \mathcal{F}^{-1}\left\{\mathrm{e}^{\mathrm{f}(\omega)}\right\}$, where $\mathcal{F}$ denotes the Fourier transform.

B. Fractional Laplacian operators and their inverses

As mentioned earlier, we choose L to be a member of the class of fractional Laplacian operators $(-\Delta)^{\gamma/2}$ for $\gamma > 0$. These isotropic differential operators defined in Fourier domain by

$$\mathcal{F}\left\{(-\Delta)^{\gamma/2}\varphi\right\}(\boldsymbol{\omega}) = |\boldsymbol{\omega}|^{\gamma}\mathcal{F}\{\varphi\}(\boldsymbol{\omega}),$$

where $\varphi \in \mathcal{S}$.

The fractional Laplacian is a linear, self-adjoint, and continuous operator with translation-, rotation-, and scaling-invariance properties. Its inverse operator is the Riesz potential I_{γ} for $\gamma < d$ and is extended for all non-integer $\gamma > d$ in Sun and Unser [4]. It has been also observed that the natural translation-invariant inverse I_{γ} of the fractional Laplacian operator can be unstable.

In accordance with this previous work, we define $\mathrm{I}_{\gamma,p}$ as the unique corrected version of the inverse mapping from $\mathcal{S}$ to the space L^p of functions with finite p-norm $\left(\int_{\mathbb{R}^d}|f(\mathbf{r})|^p d\mathbf{r}\right)^{1/p}$. The L^p stability comes with the cost of losing the translation-invariance for the inverse operator.

C. Self-similar processes

We now would like to define the process s. As we consider processes s such that $(-\Delta)^{\gamma/2}s = w$ is an innovation process, one formally writes

$$\langle s, \varphi\rangle = \langle \mathrm{I}_{\gamma,p}w, \varphi\rangle = \langle w, \mathrm{I}^*_{\gamma,p}\varphi\rangle,$$

where $\mathrm{I}_{\gamma,p}$ is the corrected inverse operator of $(-\Delta)^{\gamma/2}$ defined above.

To satisfy the admissibility conditions required between the Lévy exponent $f(\cdot)$ of w and the stability property of the inverse operator [1], one needs $p = 1$ for the Laplace innovations and $p = 2$ for the Gaussian ones. The characteristic functional of s is then given by

$$\widehat{\mathscr{P}_s}(\varphi) = \exp\left(\int_{\mathbb{R}^d} f\left(\mathrm{I}^*_{\gamma,p}\varphi(\mathbf{r})\right) d\mathbf{r}\right). \tag{4}$$

The resulting process s is called self-similar (in a stochastic sense) because an application of a similarity transformation such as scaling does not change its statistical behavior (up to some possible renormalization).

III. MAP ESTIMATION

After explaining that the process s is mathematically well-defined, we now concentrate on developing practical algorithms.

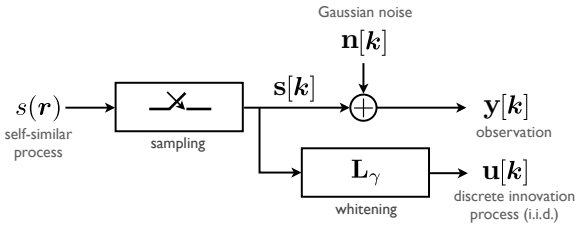


Fig. 2. Observation model.

A. Discrete innovation model

The discrete counterpart of innovation model (2) is obtained by introducing the discrete version $\mathbf{L}$ of the operator L [9]. Since we are only given the sampled version of s in real-world applications, one can think of formulating the discrete innovation model by applying $\mathbf{L}$ to the sampled process $\boldsymbol{s}[\boldsymbol{k}] = s(\mathbf{r} = \boldsymbol{k})$ for $\boldsymbol{k}$ being in a suitable discrete space Ω. In the case of fractional Laplacian operator, $\mathbf{L}_{\gamma}$ is efficiently implemented in Fourier domain via FFT operation. In effect, we define the discretized version of (2) as

$$\mathbf{u} = \mathbf{L}_{\gamma}\mathbf{s} \tag{5}$$

where $\mathbf{u}$ is called the discrete innovation process whose first-order pdf p_U is proven to be an infinitely divisible distribution [3].

As shown in [3], it is equivalent to define the discrete counterpart $(-\Delta)_d^{\gamma/2}$ of the operator $(-\Delta)^{\gamma/2}$ such that $\mathbf{u}[\boldsymbol{k}] = \left\{(-\Delta)_d^{\gamma/2}s\right\}(\mathbf{r} = \boldsymbol{k})$. In other terms, we have

$$\mathbf{u}[\boldsymbol{k}] = (\beta_{\gamma,p} * w)(\mathbf{r} = \boldsymbol{k}),$$

where $\beta_{\gamma,p} \in L^1$ is a polyharmonic B-spline and is the impulse response of the operator $(-\Delta)_d^{\gamma/2}\mathbf{I}_{\gamma,p}$. We note that the primary statistical features of $\mathbf{u}$ is related to the continuous-domain innovation process w.

Proposition 1. *If $p_{\mathrm{i.d.}} = \mathcal{F}^{-1}\left\{\mathrm{e}^{\mathrm{f}(\omega)}\right\}$ is symmetric α-stable (in particular Gaussian case), then the same is true for p_U. If $p_{\mathrm{i.d.}}$ is symmetric, unimodal with exponential decay, then the same is true for p_U.*

B. MAP estimation

We now formulate the MAP estimators for the denoising problem given in (1) under the assumption that the components $(\mathbf{u}[\boldsymbol{k}])_{k\in\Omega}$ are i.i.d. random variables. We then get the posterior distribution $p_{S|Y}$ from the Bayes' rule

$$\begin{aligned} p_{S|Y}(\mathbf{s}|\mathbf{y}) &\propto p_N(\mathbf{y}-\mathbf{s})p_U(\mathbf{u}) \\ &\propto \exp\left(-\frac{||\mathbf{y}-\mathbf{s}||_2^2}{2\sigma^2}\right)\prod_{\boldsymbol{k}\in\Omega} p_U(\mathbf{L}_{\gamma}\mathbf{s}[\boldsymbol{k}]). \end{aligned}$$

We define the *potential function* $\Phi_U(x) = -\log p_U(x)$. Then, the MAP estimator $\mathbf{s}_{\mathrm{MAP}} = \arg\max_{\mathbf{s}} p_{S|Y}(\mathbf{s}|\mathbf{y})$ is given by,

$$\begin{aligned} \mathbf{s}_{\mathrm{MAP}} = \quad & \arg\min_{\mathbf{s}} \quad \frac{1}{2}\|\mathbf{y}-\mathbf{s}\|_2^2 + \sigma^2\sum_{\boldsymbol{k}\in\Omega}\Phi_U\left(\mathbf{u}[\boldsymbol{k}]\right) \\ & \text{subject to} \quad \mathbf{u} = \mathbf{L}_{\gamma}\mathbf{s}. \end{aligned} \tag{6}$$

By using the previous definitions and the inverse Fourier transform, we arrive at

$$\Phi_U(x) = -\log\left(\int_{\mathbb{R}} \exp\left(\int_{\mathbb{R}^d} f(\beta_{\gamma,p}(\mathbf{r})\omega)d\mathbf{r} + j\omega x\right)\frac{d\omega}{2\pi}\right). \quad (7)$$

We now characterize the asymptotical form of Φ_U using the Lévy exponent of the underlying innovation process.

Theorem 1. *There exist constants A_1, A_2, A_3, B_1, B_2 and B_3 depending on the parameter of the considered innovation such that*

- *If $f(\omega) = -\sigma_0^2\omega^2$ (Gaussian case),*

$$\Phi_U(x) = A_1x^2 + B_1$$

- *If $f(\omega) = -s|\omega|^\alpha$ (α-stable case),*

$$\Phi_U(x) \sim A_2\log(|x|) + B_2$$

- *If $f(\omega) = \log\left(\frac{\lambda^2}{\lambda^2+\omega^2}\right)$ (Laplace case),*

$$\Phi_U(x) \sim A_3|x| + B_3$$

where $f \sim g$ denotes that $f - g \to 0$.

Since the computation of the exact potential function (7) is challenging in the case of Laplacian innovation, we use its simplified asymptotic form $\Phi_U(x) = A_3|x| + B_3$. Note that the constants in Theorem 1 are irrelevant for the optimization task.

IV. Numerical Examples

We perform a simple simulation that compares the estimation performance of different estimators specified by our formalism. Particulary, we concentrate on denoising of a natural texture-type and a biological image that are shown in Figure 3. We consider two innovation processes (Gaussian and Laplacian). We also consider two different whitening operators: fractional Laplacian and the discrete gradient. For the latter case, we note that one obtains Tikhonov and TV denoising.

In the experiments, the noise-free images are degraded with various levels of AWGN where the noise variance σ^2 is specified to match some given input SNR. For denoising, we use FISTA [10] for 250 iterations without any stopping criteria. The multiplicative factors are optimized for all the estimators by using an oracle to obtain the highest-possible SNR. This optimization is done in a joint way for the γ parameter of the fractional Laplacian operator. The denoising results (output SNR in dB) are reported in Table I.

The results reported in Table I illustrate that the self-similarity assumption is well-suited for the particular images considered. For the clouds, it is coherent with the fact that the self-similar processes present a fractal-type statistical behavior. Moreover, the stem cells image is seemed to be appropriate for our model as corroborated by the results. For both images, it is observed that the performance of the self-similar models outperform TV denoising.

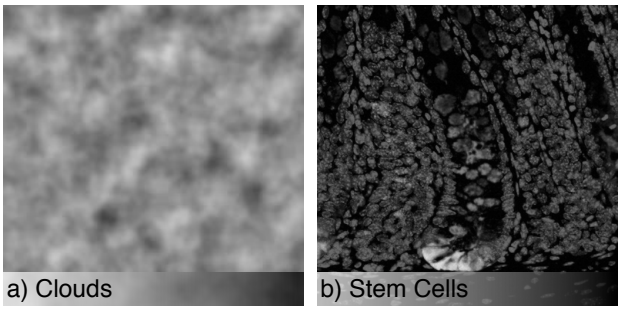


Fig. 3. Images used in the experiments.

TABLE I
Denoising Performance of different MAP estimators.

Input SNR (dB)	0	10	20	30
Estimator	Clouds			
Gaussian (discrete gradient)	20.44	24.68	29.93	35.31
Gaussian (fractional Laplacian)	**21.01**	**25.70**	**31.41**	36.45
Laplace (discrete gradient)	19.77	24.03	29.29	34.93
Laplace (fractional Laplacian)	20.22	25.29	31.16	**36.75**
Estimator	Stem cells			
Gaussian (discrete gradient)	11.16	15.85	22.13	30.40
Gaussian (fractional Laplacian)	**11.57**	**16.82**	**23.51**	**31.32**
Laplace (discrete gradient)	11.12	16.09	22.63	30.74
Laplace (fractional Laplacian)	11.20	16.70	**23.52**	**31.30**

V. Conclusion

The purpose of this work has been to drive MAP estimators that are suitable for reconstructing self-similar multi-dimensional signals from noisy samples. Our experiments showed that these estimators can outperform TV denoising for certain type of images.

VI. Acknowledgement

This work was supported by the European Commission under Grant ERC-2010-AdG 267439-FUN-SP.

References

[1] M. Unser, P. D. Tafti, and Q. Sun, "A unified formulation of Gaussian vs. sparse stochastic processes–Part I: Continuous-domain theory," *arXiv:1108.6150v1*.

[2] E. Bostan, U. Kamilov, and M. Unser, "Reconstruction of biomedical images and sparse stochastic modeling," in *Proceedings of the 9^{th} IEEE International Symposium on Biomedical Imaging: From Nano to Macro (ISBI'12)*, Barcelona, Spain, May 2-5, 2012, pp. 880–883.

[3] E. Bostan, U. S. Kamilov, M. Nilchian, and M. Unser, "Sparse stochastic processes and discretization of linear inverse problems," to appear in IEEE Transactions on Image Processing.

[4] Q. Sun and M. Unser, "Left-inverses of fractional Laplacian and sparse stochastic processes," *Advances in Computational Mathematics*, vol. 36, no. 3, pp. 399–441, April 2012.

[5] B. B. Mandelbrot, *The Fractal Geometry of Nature*. W. H. Freeman, 1983.

[6] J. Huang and D. Mumford, "Statistics of natural images and models," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Fort Collins, CO, 23-25 June 1999, pp. 637–663.

[7] I. Gelfand and N. Y. Vilenkin, *Generalized Functions. Vol. 4. Applications of Harmonic Analysis.* New York, USA: Academic Press, 1964.

[8] K. Sato, *Lévy Processes and Infinitely Divisible Distributions.* Cambridge, 1994.

[9] M. Unser, P. D. Tafti, A. Amini, and H. Kirshner, "A unified formulation of Gaussian vs. sparse stochastic processes–Part II: Discrete-domain theory," *arXiv:1108.6152v1*.

[10] A. Beck and M. Tebouelle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM*, vol. 2, no. 2, pp. 183–202, 2009.

From variable density sampling to continuous sampling using Markov chains

Nicolas Chauffert, Philippe Ciuciu
CEA, NeuroSpin center,
INRIA Saclay, PARIETAL Team
145, F-91191 Gif-sur-Yvette, France
Email: firstname.lastname@inria.fr

Pierre Weiss
ITAV, USR 3505
Toulouse, France
Email: pierre.weiss@itav-recherche.fr

Fabrice Gamboa
Université de Toulouse; CNRS
IMT-UMR5219
F-31062 Toulouse, France
Email: fabrice.gamboa@math.univ-toulouse.fr

***Abstract*—Since its discovery over the last decade, Compressed Sensing (CS) has been successfully applied to Magnetic Resonance Imaging (MRI). It has been shown to be a powerful way to reduce scanning time without sacrificing image quality. MR images are actually strongly compressible in a wavelet basis, the latter being largely incoherent with the k-space or spatial Fourier domain where acquisition is performed. Nevertheless, since its first application to MRI [1], the theoretical justification of actual k-space sampling strategies is questionable. Indeed, the vast majority of k-space sampling distributions have been heuristically designed (e.g., variable density) or driven by experimental feasibility considerations (e.g., random radial or spiral sampling to achieve smoothness k-space trajectory). In this paper, we try to reconcile very recent CS results with the MRI specificities (magnetic field gradients) by enforcing the measurements, i.e. samples of k-space, to fit continuous trajectories. To this end, we propose random walk continuous sampling based on Markov chains and we compare the reconstruction quality of this scheme to the state-of-the art.**

I. Introduction

Compressed Sensing [2], [3] is a theoretical framework which gives guarantees to recover sparse signals (signals reprensented by few non-zero coefficients in a given basis) from a limited number of linear projections. In some applications, the measurement basis is fixed and the projections should be selected amongst a fixed set. For instance, in MRI, the signal is sparse in the wavelet basis, and the sampling is performed in the spatial (2D or 3D) Fourier basis (called k-space). Possible measurements are then projections on the lines of matrix $A = F^*\Psi$, where F^* and Ψ denote the Fourier and inverse wavelet transform, respectively.

Recent results [4], [5] give bounds on the number of measurement m needed to exactly recover s-sparse signals in $\mathbb{C}^n$ or $\mathbb{R}^n$ in the framework of bounded orthogonal systems. The authors have shown that for a given s-sparse signal, the number of measurements needed to ensure its perfect recovery is $O(s\log(n))$. This methodology, called *variable density sampling*, involves an independent and identically distributed (iid) random drawing and has already given promising results in reconstruction simulations [1], [6]. Nevertheless, in real MRI, such sampling patterns cannot be implemented, because of the limited speed of magnetic field gradient commutation. Hardware constraints require at least continuity of the sampling trajectory, which is not satisfied by two-dimensional iid sampling. In this paper, we introduce a new Markovian sampling scheme to enforce continuity. Our approach relies on the following reconstruction condition introduced by Juditski, Karzan and Nemirovki [7]:

Theorem 1 ([7])**.** *If A satisfies:*

$$\gamma(A) = \min_{Y\in\mathbb{R}^{n\times m}} \|I_n - Y^T A\|_\infty < \frac{1}{2s}.$$

All s-sparse signals $x \in \mathbb{R}^n$ are recovered exactly by solving:

$$\operatorname*{argmin}_{A_m w = A_m x} \|w\|_1 \tag{1}$$

which can be seen as an alternative to the *mutual coherence* [3]. We will show that this criterion makes it possible to obtain theoritical guarantees on the number of measurements necessary to reconstruct s sparse signals, using variable density sampling or markovian sampling. Unfortunately the bounds we obtain are in $O(s^2)$. This phenomenon is due to the *quadratic bottleneck* described in [4]. We are currently trying to obtain $O(s)$ results using different proof strategies.

Notation

A signal $x \in \mathbb{R}^n$ is said to be s-sparse if it has at most s non-zero coefficients. x is measured through the acquisition system represented by a matrix A_0. Downsampling the measurements consists of deriving a matrix A composed of m lines of A_0 and observing $y = Ax \in \mathbb{R}^m$.

II. Theoretical result

A. Independent Sampling

We aim at finding $A_m \in \mathbb{R}^{m\times n}$ composed of m rows of A, and $Y_m \in \mathbb{R}^{m\times n}$ such that $\|I_n - Y_m^T A_m\|_\infty < \frac{1}{2s}$, for a given positive integer s. Following [8], we set $\Theta_i = \frac{a_i a_i^T}{\pi_i}$ and use the decomposition $I_n = A^T A = \sum_{i=1}^n \pi_i \Theta_i$. We consider a sequence of m random i.i.d. matrices $Z_1, \dots, Z_m$, taking value Θ_i with probability π_i. We set $\pi_i = \|a_i\|_\infty^2 / L$, where $L = \sum_{i=1}^n \|a_i\|_\infty^2$, so that $\|Z_l\|_\infty$ is equal to L. Let us denote $W_m = \frac{1}{m}\sum_{l=1}^m Z_l$. Then W_m may be written as $Y_m^T A_m$.

Lemma 1. $\forall t > 0$

$$\mathbb{P}(\|I_n - W_m\|_\infty > t) \le n(n+1)\exp\Big(-\frac{mt^2}{2L^2 + 2Lt/3}\Big). \tag{2}$$

Proof: Bernstein's concentration inequality [9] states that if $X_1, \dots, X_m$ are independent zero-mean random variables such that for all i, $|X_i| \leq \alpha$ and $\sigma^2 = \sum_i \mathbb{E}\left(X_i^2\right) < \infty$, then $\forall t > 0$

$$\mathbb{P}\left(|\sum_{i=1}^{m} X_i| > t\right) \leq 2\exp\left(-\frac{t^2}{2(\sigma^2 + \alpha t/3)}\right).$$

For $1 \leq a, b \leq n$, let $M^{(a,b)}$ denote the (a,b)th entry of a matrix $M \in \mathbb{R}^{n\times n}$. The random variable $(I_n - Z_l)^{(a,b)}$ is centered since $\sum_{i=1}^{n} \pi_i \Theta_i = I_n$. Moreover, $|(I_n - Z_l)^{(a,b)}| \leq L$. Applying Bernstein's inequality to the sequence $\frac{1}{m}\left((I_n - Z_l)^{(a,b)}\right)_{1\leq l\leq m}$ gives

$$\mathbb{P}\left(|(I_n - W_m)^{(a,b))}| > t\right) \leq 2\exp\left(-\frac{mt^2}{2L^2 + 2Lt/3}\right).$$

Finally, using a union bound and the symmetry property of matrix $(I_n - W_m)$, we get:

$$\mathbb{P}\left(\|I_n - W_m\|_\infty > t\right) \leq \sum_{1\leq a\leq b\leq n} \mathbb{P}\left(|I_n - W_m|^{(a,b)} > t\right).$$

Since $\mathbb{P}\left(|I_n - W_m|^{(a,b)} > t\right)$ is independent of (a,b), we obtain Eq. (2). ∎

Remark 1. *Setting $t = 4L\sqrt{\frac{2\ln(2n^2)}{m}}$ in lemma 1, the bound given by Juditsky et al. in [8] is $\mathbb{P}\left(\|I_n - W_m\|_\infty \geq t\right) \leq \frac{1}{2}$. This bound is obtained by upper-bounding the mean of $\|I_n - W_m\|_\infty$ and using Markov inequality. Setting the same t value in Eq.* (2)*, and assuming $t \leq L$, we obtain $\mathbb{P}\left(\|I_n - W_m\|_\infty \geq t\right) \leq \frac{1}{2n^4}$. This huge difference comes from inability of Markov inequality to capture large deviations behaviors.*

From lemma 1, we can derive the immediate following result by setting $t = 1/2s$:

Proposition 1. *Let A_m be a measurement matrix designed by drawing m lines of A under the distribution π. Then, with probability $1-\eta$, if*

$$m \geqslant 5L^2 s^2 \log(n^2/\eta), \tag{3}$$

every s-sparse signal x is the unique solution of the ℓ_1 problem:

$$\underset{A_m w = A_m x}{\operatorname{argmin}} \|w\|_1$$

B. Markovian sampling

Sampling patterns obtained using the strategy presented in Section II-A are not usable for many practical devices. A common constraint met on many hardwares (e.g. MRI) is the proximity of successive measurements. A simple way to model dependence between successive samples consists of introducing a Markov chain $X_1 \dots X_m$ on the set $\{1, \dots, n\}$ that represents locations of possible measurements. The transition probability to go from location i to location j is positive if and only if sampling i and j successively is possible. We denote $W_m = \frac{1}{m}\sum_{l=1}^{m} \Theta_{X_l}$.

In order to use a concentration inequality, W_m should satisfy $\mathbb{E}\left(W_m\right) = I_n$. We thus need (i) to set the stationary distribution of the Markov chain to π and (ii) to set up the chain with its stationnary distribution π. These two conditions ensure that the marginal distribution of the chain is π_i at any time. The issue of designing such a chain is widely studied in the frame of Markov chain Monte Carlo (MCMC) algorithms.

A simple way to build up the transition matrix $P = (P_{ij})_{1\leq i,j\leq n}$ is the Metropolis algorithm [10]. Let us now recall a concentration inequality for finite-state Markov chains [11].

Theorem 2. *Let (P, π) be an irreductible and reversible Markov chain on a finite set G of size n. Let $f : G \to \mathbb{R}$ be such that $\sum_{i=1}^{n} \pi_i f_i = 0$, $\|f\|_\infty \leq 1$ and $0 < \sum_{i=1}^{n} f_i^2 \pi_i \leq b^2$. Then, for any initial distribution q, any positive integer m and all $0 < t \leq 1$,*

$$\mathbb{P}\Big(\frac{1}{m}\sum_{i=1}^{m} f(X_i) \geq t\Big) \leq e^{\frac{\epsilon(P)}{5}} N_q \exp\Big(-\frac{mt^2\epsilon(P)}{4b^2(1 + h(5t/b^2))}\Big)$$

where $N_q = (\sum_{i=1}^{n}(\frac{q_i}{\pi_i})^2\pi_i)^{1/2}$, $\beta_1(P)$ is the second largest eigenvalue of P, and $\epsilon(P) = 1 - \beta_1(P)$ is the spectral gap of the chain. Finally h is given by $h(x) = \frac{1}{2}(\sqrt{1+x} - (1 - x/2))$.

Using this theorem, we can guarantee the following control of the term $\|I_n - W_m\|_\infty$:

Lemma 2. *$\forall\ 0 < t \leq 1$,*

$$\mathbb{P}\left(\|I_n - W_m\|_\infty \geq t\right) \leq n(n+1)e^{\frac{\epsilon(P)}{5}} \exp\Big(-\frac{mt^2\epsilon(P)}{12L^2}\Big). \tag{4}$$

Proof: By applying Theorem 2 to a function f and then to its opposite $-f$, we get:

$$\mathbb{P}\Big(\Big|\frac{1}{m}\sum_{i=1}^{m} f(X_i)\Big| \geq t\Big) \leq 2e^{\frac{\epsilon(P)}{5}} N_q \exp\Big(-\frac{mt^2\epsilon(P)}{4b^2(1 + h(5t/b^2))}\Big).$$

Then we set $f(X_i) = (I_n - \Theta_{X_i})^{(a,b)}/(1+L)$. The Markov chain is constructed such that $\sum_{i=1}^{n} \pi_i f(X_i) = 0$. Since we have $\|f\|_\infty \leq 1$, $b = 1$, and since $t \leqslant 1$, $1 + h(5t) < 3$. Moreover, since the initial distribution is π, $q_i = \pi_i, \forall i$ and thus $N_q = 1$. Again, resorting to a union bound (II-A) enables us to extend the result for the (a,b)th entry to the whole infinite norm of the $n \times n$ matrix $I_n - W_m$ (4). ∎

Then we can quantify the number of measurements needed to ensure exact recovery:

Proposition 2. *Let A_m be a measurement matrix designed by drawing m lines of A under the Markovian process described above. Then, with probability $1-\eta$, if*

$$m \geqslant \frac{12L^2}{\epsilon(P)} s^2 \log(2n^2/\eta), \tag{5}$$

every s-sparse signal x is the unique solution of the ℓ_1 problem:

$$\underset{A_m w = A_m x}{\operatorname{argmin}} \|w\|_1$$

Remark 2. *The spectral gap $\epsilon(P)$ takes its value between 0 and 1 and describes the mixing properties of the Markov chain. The closer the spectral gap to 1, the fastest the convergence to the mean.*

Remark 3. *All the results above can be extended to the complex case using a slightly different proof.*

III. Results and discussion

In order to cover a larger domain of k-space, we consider the following chain: $P^{(\alpha)} = (1-\alpha)P + \alpha\tilde{P}$, where $\tilde{P}$ corresponds to an independent drawing $\tilde{P}_{ij} = \pi_j, \forall i, j$. This chain has π as invariant distribution, and fulfills the continuity property while enabling a jump with probability of α.

Weyl's Theorem [12] ensures that $\epsilon(P^{(\alpha)}) > \alpha$. This bound is useful because of the dependence of $\epsilon(P)$ with respect to the problem dimension, which would have weakened condition (5).

Sampling scheme obtained by these methods are composed of $1/\alpha$-average length random walks on the k-space. All our experiments consist of reconstructing a two-dimensional image from a sampled k-space by solving an ℓ_1 minimization problem. Constrained ℓ_1 minimization (Eq. (1)) is performed using the Douglas-Rachford algorithm [13]. In each case, only twenty percent of the Fourier coefficients are kept, which corresponds to an acceleration factor of $r = 5$. Since the schemes are obtained by a random process, we run each experiment 10 times independently, and compared the mean value of the reconstruction results in terms of *Peak Signal-to-Noise Ratio* (PSNR).

In Fig. 1, it is shown that the image reconstruction quality degrades when α decreases. These results can be explained by the spatial confinement of the continuous parts of a given Markov chain, except for large values of α. There seems to be a compromise between the number of discontinuities of the chain (linked to the hardware constraints in MRI) and the k-space coverage. Nevertheless, accurate reconstruction results can be observed with reasonable average mean length of connected subparts ($\alpha = 0.01$ or 0.001).

The mixing properties of the chain (through its spectral gap) seem to have a strong impact on the quality of the scheme, as shown in Proposition 2. Unfortunately, the spectral gap is strongly related to the problem dimension n and can tend to zero if n goes to infinity. This proves to be a theoretic limitation of this method. Nevertheless, we obtained reliable reconstruction results which cannot be explained by the proposed theory. Since the design process is based on randomness, we can even expose a specific scheme which provides accurate reconstruction results instead of considering the mean behavior (Fig. 2). We currently aim at deriving a stronger result on the number of measurements needed, involving a $O(s)$ bound. Meanwhile, we are developing second order chains which can ensure more regularity of the trajectories and for which we have already observed good reconstruction results (Fig. 3).

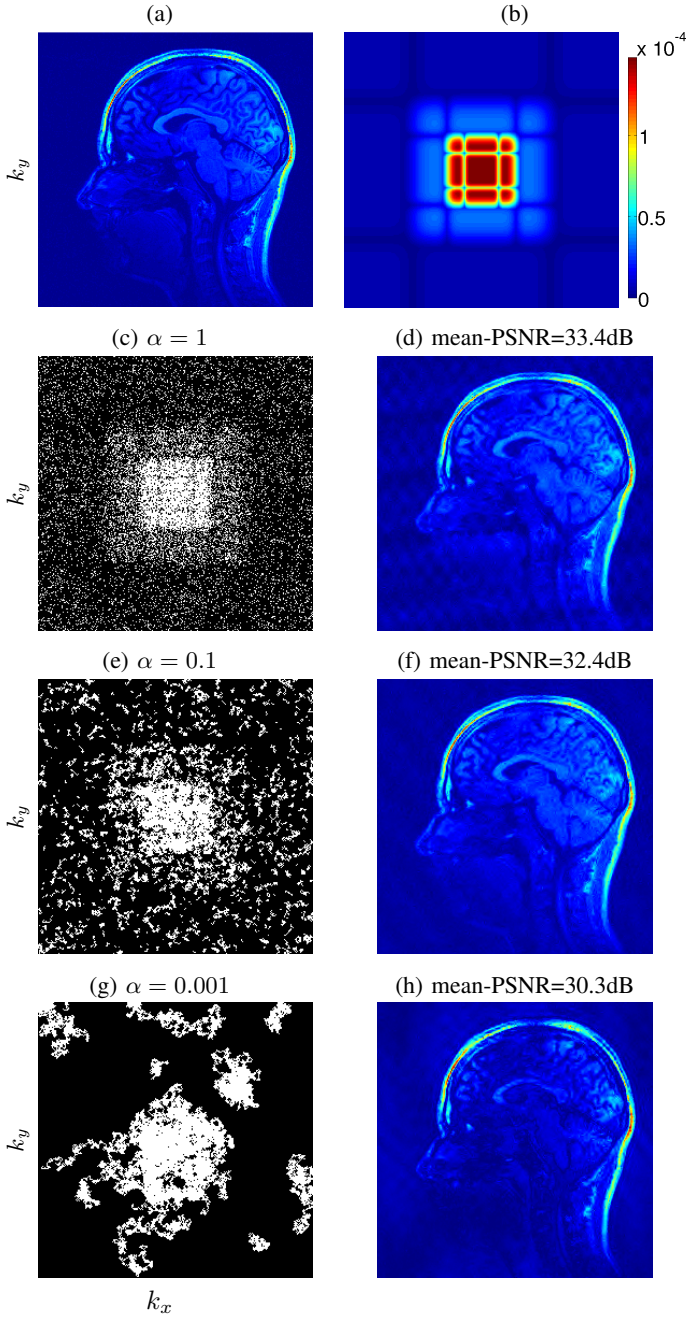


Fig. 1: **First line:** reference image used in our experiments (a) and π distribution (b). **Lines 2 to 4, left:** different sampling patterns (with an acceleration factor $r = 5$). **right:** reconstruction results. From line 2 to bottom: independent drawing from distribution π (c), corresponding to $\alpha = 1$. (e) (resp (g)) represents a sampling scheme designed with the presented markovian process with transition matrix $P^{(\alpha)}$ for $\alpha = 0.1$) (resp. $\alpha = 0.001$).

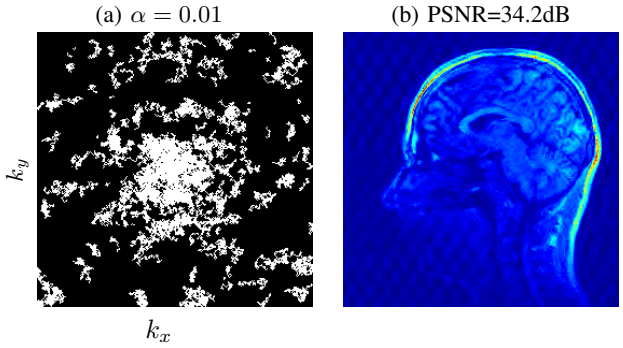


Fig. 2: Sampling scheme obtained setting $\alpha = 0.01$ and $r = 5$ (a) and its corresponding reconstructed image (b).

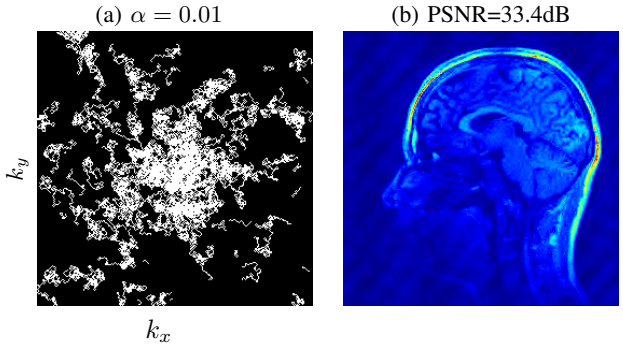


Fig. 3: Preliminary results for second order Markov chain: sampling scheme obtained setting $\alpha = 0.01$ and $r = 5$ (a) and its corresponding reconstructed image (b).

IV. Conclusion

We proposed a novel approach combining compressed sensing and Markov chains to design continuous sampling trajectories, required for MRI applications. Our work may easily be extended to a 3D framework by considering a different neighbourhood of each k-space location. Existing continuous trajectories in CS-MRI only exploit 1D or 2D randomness for 2D or 3D k-space sampling, respectively. In the latter case, the points are randomly drawn in the plane defined by the partition and phase encoding directions so as to maintain continuous sampling in the orthogonal readout direction (frequency encoding). Here, the novelty relies both on the use of randomness in all k-space dimensions, and the establishment of compressed sensing results for continuous trajectories, based on a concentration result for Markov chains.

Acknowledgements

We thank Jérémie Bigot for the time he dedicated to our questions and his helpful remarks. The authors would like to thank the CIMI Excellence Laboratory for inviting Philippe Ciuciu on an excellence researcher position during winter 2013.

References

[1] M. Lustig, D. L. Donoho, and J. M. Pauly, "Sparse MRI: The application of compressed sensing for rapid MR imaging," *Magn. Reson. Med.*, vol. 58, no. 6, pp. 1182–1195, Dec. 2007.

[2] E. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 489–509, 2006.

[3] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[4] H. Rauhut, "Compressive Sensing and Structured Random Matrices," in *Theoretical Foundations and Numerical Methods for Sparse Recovery*, M. Fornasier, Ed., vol. 9 of *Radon Series Comp. Appl. Math.*, pp. 1–92. deGruyter, 2010.

[5] E. J. Candès and Y. Plan, "A probabilistic and ripless theory of compressed sensing," *IEEE Trans. Inf. Theory*, vol. 57, no. 11, pp. 7235–7254, 2011.

[6] G. Puy, P. Vandergheynst, and Y. Wiaux, "On variable density compressive sampling," *IEEE Signal Processing Letters*, vol. 18, no. 10, pp. 595–598, 2011.

[7] A. Juditsky and A. Nemirovski, "On verifiable sufficient conditions for sparse signal recovery via ℓ_1 minimization," *Mathematical Programming Ser. B*, vol. 127, pp. 89–122, 2011.

[8] A. Juditsky, F.K. Karzan, and A. Nemirovski, "On low rank matrix approximations with applications to synthesis problem in compressed sensing," *SIAM J. on Matrix Analysis and Applications*, vol. 32, no. 3, pp. 1019–1029, 2011.

[9] M. Ledoux, "The Concentration of Measure Phenomenon," *Amer. Mathematical Society*, vol. 89, 2001.

[10] W. K. Hastings, "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika*, vol. 57, no. 1, pp. 97–109, Apr. 1970.

[11] P. Lezaud, "Chernoff-type bound for finite Markov chains," *Annals of Applied Probability*, vol. 8, no. 3, pp. 849–867, 1998.

[12] R. Horn and C. Johnson, *Topics in matrix analysis*, Cambridge University Press, Cambridge, 1991.

[13] P L Combettes and J.-C Pesquet, "Proximal Splitting Methods in Signal Processing," in *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, pp. 185–212. Springer, 2011.

A Comparison of Reconstruction Methods for Compressed Sensing of the Photoplethysmogram

Nicholas J. Conn
Microsystems Engineering
Rochester Institute of Technology
Rochester, New York 14623
Email: nicholasjconn@ieee.org

David A. Borkholder, PhD
Microsystems Engineering
Rochester Institute of Technology
Rochester, New York 14623
Email: david.borkholder@rit.edu

***Abstract*—Compressed sensing has the possibility to significantly decrease the power consumption of wireless medical devices. The photoplethysmogram (PPG) is a device which can greatly benefit from compressed sensing due to the large amount of power needed to capture data. The aim of this paper is to determine if the least absolute shrinkage and selection operator (LASSO) optimization algorithm is the best approach for reconstructing a compressively sampled PPG across varying physiological states. The results show that LASSO reconstruction approaches, but does not surpass, the reliability of constrained optimization.**

I. Introduction

Compressed sensing is the process of sampling a sparse signal at a rate significantly lower the Nyquist rate [1]. The Nyquist rate is defined as twice the highest frequency of the measured signal. Compressed sensing involves reconstructing a sample vector of length N from a set of M random measurements, where M is much less than N. These M measurements are captured using a randomly generated sensing matrix, which is used to reconstruct the sample vector by estimating the coefficients in sparse domain of the signal being measured.

In 2006, the seminal papers on compressed sensing were published by Candes and Donoho [2]–[5]. These papers discuss the mathematical principles behind compressed sensing. They have shown that a signal has a very high probability of being exactly reconstructed when the signal has a known sparse domain and is measured using a sensing matrix that is incoherent to the basis functions of the signal's sparse domain.

Previous work has shown the success of the least absolute shrinkage and selection operator (LASSO) optimization algorithm for reconstruction of compressively sensed physiological signals [6]. More specifically, the work from Beheti introduces the use of a weighted LASSO technique for reconstructing a compressively sensed photoplethysmogram (PPG), but does not compare it to constrained ℓ_1 norm reconstruction [7], [8].

LASSO allows for a balance between minimizing the norm of the sparse domain and maintaining measurement accuracy by adhering to the constraints imposed by the measurement vector. This trade-off is determined by the LASSO penalty parameter, which is a constant that must be chosen before implementation [9]. For this reason, LASSO is a popular reconstruction method for the compressed sensing of signals which contain measurement noise.

The PPG is used in many medical devices to attain blood oxygenation levels, heart rate, and other cardiac intervals. To estimate the blood oxygenation level (SpO2), the ratio between the root mean square (RMS) of PPG signals captured at two separate wavelengths can be used [8]. This paper quantitatively shows how different LASSO penalty parameters affect important aspects of the PPG such as the root mean square (RMS) and temporal information change when compared to constrained optimization. The purpose of this paper is to determine if LASSO provides any benefit over constrained optimization for reconstructing the PPG across a range of physiological states.

Section II includes a mathematical introduction to compressed sensing, an explanation of different reconstruction methods, and the metrics used for determining the accuracy of reconstruction. In Section III, results are presented and analyzed. Finally, Section IV provides a short summary and concluding remarks.

II. Theory

A. Overview of Compressed Sensing

Compressed sensing is the process of utilizing only M measurements to reconstruct a discrete signal of length N, where $M \ll N$. The level of compression achieved can be represented by the under sampling ratio (USR), which is N divided by M. For compressed sensing to work, the sample vector $\vec{\mathbf{x}}$ must be sparse in some domain [10]. By multiplying $\vec{\mathbf{x}}$ by the discrete cosine matrix $\mathbf{\Psi}$, the PPG is shown to be sparse in the discrete cosine domain (Fig. 1). This process is shown in (1).

$$\mathbf{\Psi} \cdot \vec{\mathbf{x}} = \vec{\mathbf{s}} \tag{1}$$

Given that $\vec{\mathbf{x}} \in \mathbb{R}^N$ is sparse it is multiplied by a sensing matrix, $\mathbf{A} \in \mathbb{R}^{M \times N}$. This results in the measurement vector $\vec{\mathbf{y}} \in \mathbb{R}^M$, as shown in (2).

$$\mathbf{A} \cdot \vec{\mathbf{x}} = \vec{\mathbf{y}} \tag{2}$$

The only requirement for the sensing matrix is that it is incoherent to the basis functions of the sparse domain, $\mathbf{\Psi}$. Normally, this is achievable by populating $\mathbf{A}$ with values generated from a random distribution [10]. The sensing matrix

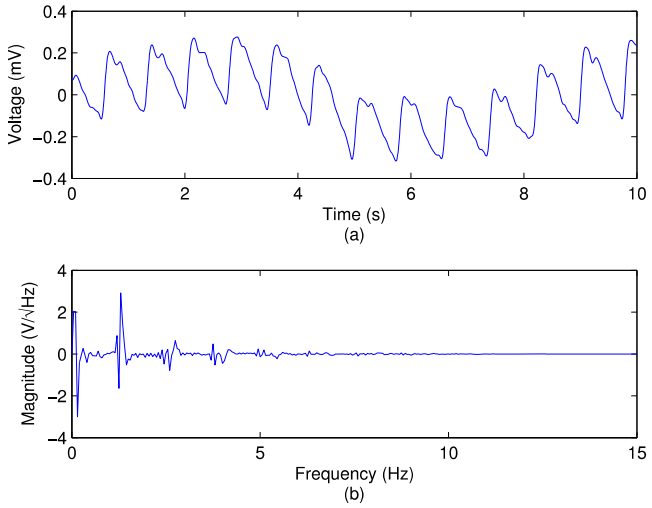


Fig. 1. The sparsity of a 10 second (N = 1600) long PPG signal (a) is shown in the discrete cosine domain (b).

used in this research is generated using an identity matrix, which is an orthogonal set of impulse functions. The process used for this is shown in (3), where random rows are removed from an identity matrix to create $\mathbf{A}$. This structure was chosen so that not all samples in $\vec{\mathbf{x}}$ are used to generate $\vec{\mathbf{y}}$.

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ \cancel{0} & \cancel{1} & \cancel{0} & \cancel{0} & \cancel{0} & \cancel{0} \\ \cancel{0} & \cancel{0} & \cancel{1} & \cancel{0} & \cancel{0} & \cancel{0} \\ 0 & 0 & 0 & 1 & 0 & 0 \\ \cancel{0} & \cancel{0} & \cancel{0} & \cancel{0} & \cancel{1} & \cancel{0} \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad (3)$$

The process shown in (2), which uses both $\mathbf{A}$ and $\vec{\mathbf{x}}$ to generate $\vec{\mathbf{y}}$, forms an under-determined linear system which is used as constraints during reconstruction.

B. Reconstruction

Typically, the reconstruction process for compressed sensing can be defined as a convex optimization problem [2], [10]. Given the under-determined linear system in (2) and the matrix transform that projects $\vec{\mathbf{x}}$ onto a sparse basis shown in (1), the optimization problem is defined as

$$\min_{\hat{\mathbf{x}} \in \mathbb{R}^N} \|\mathbf{\Psi} \cdot \hat{\mathbf{x}}\|_1 \text{ subject to } \vec{\mathbf{y}} = \mathbf{A} \cdot \hat{\mathbf{x}} \ . \quad (4)$$

This will minimize the ℓ_1 norm of the sparse coefficients, while exactly adhering to constraints imposed by the measurement vector and sensing matrix. This method, also called constrained optimization, typically does not perform well when in applications where measurement noise is present [9], [10].

One method for accurately reconstructing a signal in the presence of noise is least absolute shrinkage and selection operator (LASSO) based optimization, shown in (5) [6]. By adjusting the penalty parameter λ, the amount of deviation from (2) can be specified, allowing for a more robust overall reconstruction for signals that contain noise [6], [9].

$$\min_{\hat{\mathbf{x}} \in \mathbb{R}^N} \{\lambda \|\mathbf{\Psi} \cdot \hat{\mathbf{x}}\|_1 + \|\vec{\mathbf{y}} - \mathbf{A} \cdot \hat{\mathbf{x}}\|_2^2\} \quad (5)$$

This approach can be expanded further by incorporating a weighting vector for the sparse domain. The weighting vector is generated using *a priori* information and can greatly increase the accuracy of the reconstructed waveform. This method can be used on any signal that has a typical or characteristic set of coefficients in its sparse domain.

The results in this paper use the discrete cosine domain as the sparse domain of the PPG. The weighting vector $\vec{\mathbf{w}}$ for the sparse domain used is generated in (6) from the average sparse domain coefficients for the signal being captured. The average sparse domain $\bar{\mathbf{s}}$ can be found by averaging a set of training signals together.

The larger the weighting coefficient, the smaller the reconstructed sparse domain coefficient will be. The maximum possible value of $\vec{\mathbf{w}}$ is determined by σ, a small constant. When the average sparse coefficient is zero, the maximum weighting is $1/\sigma$. By integrating the weighting vector into constrained ℓ_1 norm optimization, (7) is defined.

$$[w]_i = \frac{1}{[\bar{s}]_i + \sigma} \qquad i = 0, 1, \cdots, N-1 \quad (6)$$

$$\min_{\hat{\mathbf{x}} \in \mathbb{R}^N} \| \langle \vec{\mathbf{w}}, (\mathbf{\Psi} \cdot \hat{\mathbf{x}}) \rangle \|_1 \text{ subject to } \vec{\mathbf{y}} = \mathbf{A} \cdot \hat{\mathbf{x}} \quad (7)$$

Similarly, a weighted LASSO minimization is formed in (8).

$$\min_{\hat{\mathbf{x}} \in \mathbb{R}^N} \{\lambda \| \langle \vec{\mathbf{w}}, (\mathbf{\Psi} \cdot \hat{\mathbf{x}}) \rangle \|_1 + \|\vec{\mathbf{y}} - \mathbf{A} \cdot \hat{\mathbf{x}}\|_2^2\} \quad (8)$$

The results presented in Section III use the weighting vector shown in (6) for both LASSO based reconstruction and constrained ℓ_1 norm optimization.

C. Error Metrics

Two very important measurements which can be extracted from the PPG are the blood oxygenation level (SpO2), and when used in conjunction with a synchronized ECG, the pulse transit time. For this reason, two different metrics are used to compare the accuracy of the reconstructed sample vector to the original Nyquist sample vector, the change in RMS and the change in the location of the PPG foot.

In order to show how reconstruction errors affect the estimated SpO2 levels, the percent change in RMS value is used. This percentage is determined by normalizing the change in RMS to the original signal's RMS value, as shown in (9).

$$RMS_{diff} = \frac{|RMS(\vec{\mathbf{x}}) - RMS(\hat{\mathbf{x}})|}{RMS(\vec{\mathbf{x}})} \quad (9)$$

Even small errors in the reconstructed frequency domain can result in peak deformation and other temporal changes that will not be apparent when only using (9). In fact, maintaining

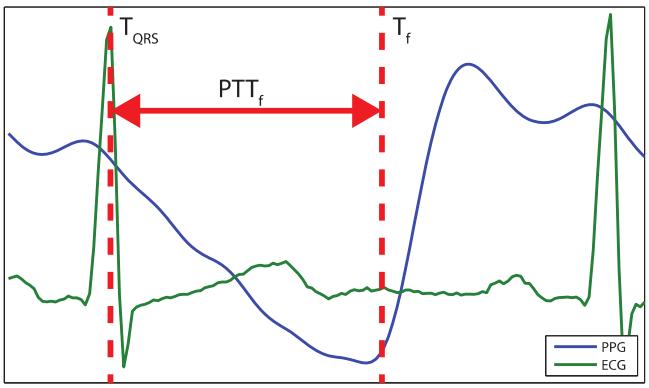


Fig. 2. By locating the foot of the PTT (T_f) and the QRS peak on the ECG (T_{QRS}), the pulse transit time (PTT_f) can be found.

the temporal accuracy of physiological signals is required for many applications.

An important temporal feature of the PPG is the location of the foot in each beat. By determining the interval between the ECG QRS peak and the PPG foot, the pulse transit time (PTT) can be estimated. Fig. 2 illustrates how the PTT is estimated for a given beat and (10) shows how the normalized PTT is calculated. A signal's PTT is calculated by averaging the PTT for each beat. It is important to note that the ECG was not compressively sampled since it is used as a baseline to determine how specific locations of the reconstruction differs from the original signal.

$$PTT_f = \frac{|(T_f - T_{QRS}) - (\widehat{T}_f - T_{QRS})|}{T_f - T_{QRS}} \qquad (10)$$

Finally, in order to show how the penalty parameter affects the time of reconstruction, the time it took to reconstruct 60 seconds of data was captured directly in MATLAB.

III. Results and Discussion

The physiological data sets used were attained under informed consent in a protocol approved by the Rochester Institute of Technology Institutional Review Board for Protection of Human Subjects. The Biopac MP36 (Biopack Systems, Inc., Goleta, CA) was used to capture synchronized ECG and ear PPG at a sample rate of 50 kHz.

One minute measurements were captured at eleven different activity levels to test how well each reconstruction method performs across a range of physiological states. Before each measurement was analyzed, it was decimated to a sample rate of 160 Hz to more closely match the Nyquist sampling rate found in physical systems. Each minute sample was split into five sample vectors with an N of 1920 (a window size of 12 seconds) for compressed sensing.

To further increase the reliability of the results provided, each metric is averaged across eleven activity levels and four different random sensing matrices generated using (3). The following tests were performed in MATLAB (R2012a) using CVX, a package for specifying and solving convex optimization problems [11], [12].

The results in Fig. 3 show that larger penalty parameters correspond to a higher RMS error for a wide range of USRs. A more detailed look at how the penalty parameter affects the RMS accuracy for a specific USR is shown in Fig. 4. This more clearly shows how different LASSO penalty parameters perform when compared to the constrained ℓ_1 norm reconstruction shown in (4).

When using constrained ℓ_1 norm reconstruction, the average RMS percent difference is 8% with a standard deviation of 1.7%. For penalty parameters below 0.003, the RMS percent difference is less than 10% with an average standard deviation of approximately 2.1%.

As discussed in Section II, analyzing the temporal accuracy of a reconstructed physiological signal is very important. For a USR of 16, the PTT foot error results are shown in Fig. 5 and show a higher error rate than the percent RMS difference; this is due to the fact that the algorithm used to detect the foot of the PPG is not perfectly robust in the presence of noise.

As the penalty parameter decreases, the standard deviation and error rate approach that of the constrained ℓ_1 norm reconstruction. This is the same trend shown in the RMS error

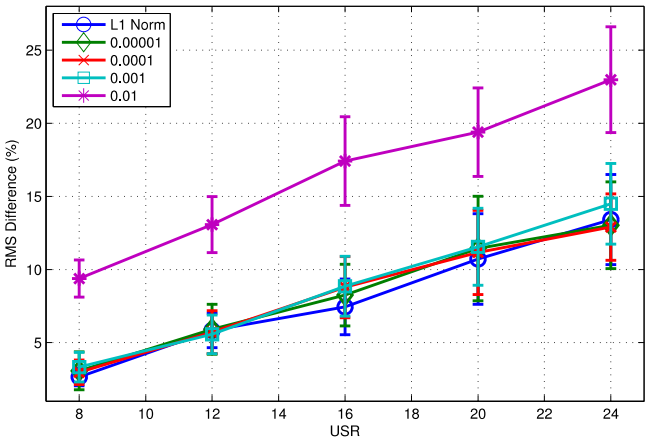


Fig. 3. The difference in RMS for different LASSO penalty parameters decreases as the penalty parameter decreases.

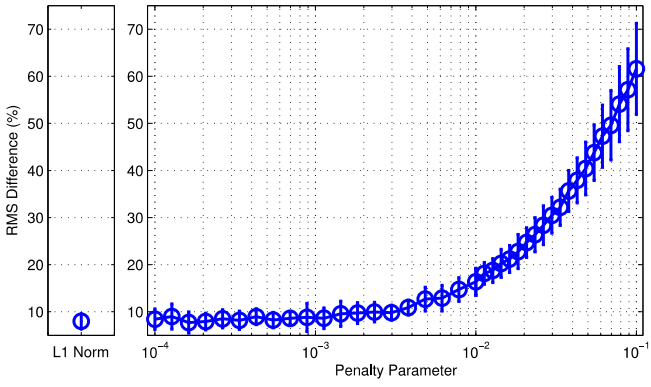


Fig. 4. As the penalty parameters decreases at a USR of 16, the RMS difference approaches the rate and variance of the constrained ℓ_1 norm error (on the left).

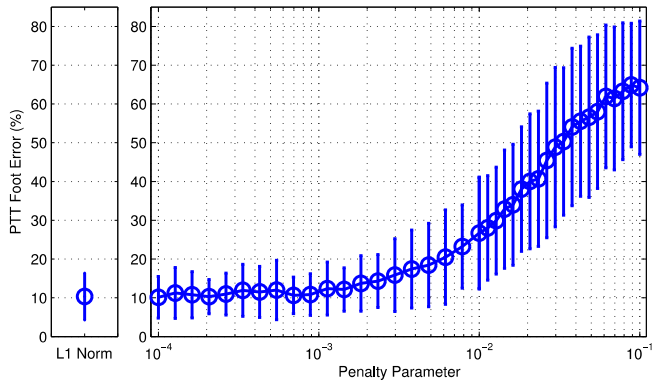


Fig. 5. While the standard deviation of the PTT_f is higher than that of the RMS difference, the general trend is the same at a USR of 16. As the penalty parameter decreases, the accuracy increases.

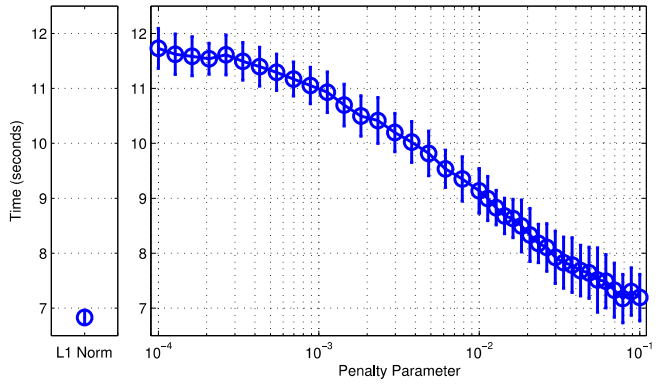


Fig. 6. Lower penalty parameters correspond to significantly higher reconstruction times when compared to constrained ℓ_1 norm reconstruction at a USR of 16.

results. For penalty parameters below 0.001 the average error rate is 11% with a standard deviation of 6%. While the smaller penalty parameters perform just as well as the ℓ_1 norm method for both of these metrics, they do not perform any better. The ℓ_1 norm method has an error rate of 10.35% with a standard deviation of 6%.

Finally, the average reconstruction time for different penalty parameters for a USR of 16 and a window size of 12 seconds is shown in Fig. 6. As the penalty parameter becomes smaller, the reconstruction time increases. While the overall increase in reconstruction time is small, approximately 5 seconds, it can become significant for larger window sizes.

When noise distorts the measurement vector $\vec{\mathbf{y}}$ in (2), LASSO typically allows for a decrease in reconstruction variability compared to constrained ℓ_1 norm reconstruction [5], [9], [10]. The results presented herein utilize low noise physiologic signals, which may explain why LASSO based reconstruction performs as well, but not better than, ℓ_1 norm reconstruction.

IV. Conclusion

These results shown that a compressively sampled PPG can be accurately reconstructed using both weighted LASSO reconstruction and weighted constrained ℓ_1 norm reconstruction across a range of physiological states (activity levels). Based on a Nyquist sample rate of 160 Hz, compressed sensing of the PPG can be achieved with a USR of 16 while maintaining an overall error rate of approximately 10%, this corresponds to an average sample rate of 10 Hz.

LASSO based reconstruction with penalty parameters below 0.001, is just as reliable and accurate as constrained ℓ_1 norm reconstruction. Given that it is also slower than the constrained ℓ_1 norm reconstruction, LASSO offers no quantitative befit for the compressed sensing of the PPG.

Future research should compare LASSO based reconstruction to constrained ℓ_1 norm reconstruction, on a physical compressed sensing system, by measuring and reconstructing physiological signals which contain noise that distort the measurement vector. Additionally, the affect different sparse domains have on the accuracy of reconstruction for the PPG should be investigated.

Acknowledgment

The authors would like to thank Jeff Lillie and Cody Cziesler for their technical support and for gathering the subject data used for this publication. They would also like to thank National Semiconductor for their support.

References

[1] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-time signal processing (2nd ed.).* Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1999.

[2] D. Donoho, "Compressed sensing," *Information Theory, IEEE Transactions on*, vol. 52, no. 4, pp. 1289 –1306, april 2006.

[3] D. Donoho and Y. Tsaig, "Recent advances in sparsity-driven signal recovery," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, vol. 5, march 2005, pp. v/713 – v/716 Vol. 5.

[4] E. Candes and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?" *Information Theory, IEEE Transactions on*, vol. 52, no. 12, pp. 5406 –5425, dec. 2006.

[5] E. J. Candes, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on Pure and Applied Mathematics*, vol. 59, no. 8, pp. 1207–1223, 2006. [Online]. Available: http://dx.doi.org/10.1002/cpa.20124

[6] A. Dixon, E. Allstot, A. Chen, D. Gangopadhyay, and D. Allstot, "Compressed sensing reconstruction: Comparative study with applications to ecg bio-signals," in *Circuits and Systems (ISCAS), 2011 IEEE International Symposium on*, may 2011, pp. 805 –808.

[7] P. Baheti and H. Garudadri, "An ultra low power pulse oximeter sensor based on compressed sensing," in *Wearable and Implantable Body Sensor Networks, 2009. BSN 2009. Sixth International Workshop on*, june 2009, pp. 144 –148.

[8] P. K. Baheti, H. Garudadri, and S. Majumdar, "Blood oxygen estimation from compressively sensed photoplethysmograph," in *Wireless Health 2010*, ser. WH '10. New York, NY, USA: ACM, 2010, pp. 10–14. [Online]. Available: http://doi.acm.org/10.1145/1921081.1921084

[9] D. Angelosante, G. Giannakis, and E. Grossi, "Compressed sensing of time-varying signals," in *Digital Signal Processing, 2009 16th International Conference on*, july 2009, pp. 1 –8.

[10] E. Candes and M. Wakin, "An introduction to compressive sampling," *Signal Processing Magazine, IEEE*, vol. 25, no. 2, pp. 21 –30, march 2008.

[11] CVX Research, Inc., "CVX: Matlab software for disciplined convex programming, version 2.0 beta," http://cvxr.com/cvx, Sep. 2012.

[12] M. Grant and S. Boyd, "Graph implementations for nonsmooth convex programs," in *Recent Advances in Learning and Control*, ser. Lecture Notes in Control and Information Sciences, V. Blondel, S. Boyd, and H. Kimura, Eds. Springer-Verlag Limited, 2008, pp. 95–110, http://stanford.edu/~boyd/graph_dcp.html.

Generalized sampling in U-invariant subspaces

H. R. Fernández–Morales
Departamento de Matemáticas,
Universidad Carlos III de Madrid
Madrid, España
Email: hfernand@math.uc3m.es

A. G. García
Departamento de Matemáticas,
Universidad Carlos III de Madrid
Madrid, España
Email: agarcia@math.uc3m.es

M. A. Hernández–Medina
Departamento de Matemática Aplicada,
E.T.S.I.T., U.P.M.,
Madrid, España
Email: miguelangel.hernandez.medina@upm.es

***Abstract*—In this work we carry out some results in sampling theory for U-invariant subspaces of a separable Hilbert space $\mathcal{H}$, also called atomic subspaces:**

$$\mathcal{A}_a = \big\{ \sum_{n\in\mathbb{Z}} a_n U^n a : \{a_n\} \in \ell^2(\mathbb{Z}) \big\},$$

where U is an unitary operator on $\mathcal{H}$ and a is a fixed vector in $\mathcal{H}$. These spaces are a generalization of the well-known shift-invariant subspaces in $L^2(\mathbb{R})$; here the space $L^2(\mathbb{R})$ is replaced by $\mathcal{H}$, and the shift operator by U. Having as data the samples of some related operators, we derive frame expansions allowing the recovery of the elements in $\mathcal{A}_a$. Moreover, we include a frame perturbation-type result whenever the samples are affected with a jitter error.

I. Introduction

Our work is motivated by the generalized sampling problem in shift-invariant subspaces of $L^2(\mathbb{R})$. Namely, assume that our functions (signals) belong to some shift-invariant space of the form:

$$V_\varphi^2 := \overline{\operatorname{span}}_{L^2(\mathbb{R})} \big\{ \varphi(t-n),\ n\in\mathbb{Z} \big\},$$

where the generator function φ belongs to $L^2(\mathbb{R})$ and the sequence $\{\varphi(t-n)\}_{n\in\mathbb{Z}}$ is a Riesz sequence for $L^2(\mathbb{R})$. Thus, the shift-invariant space V_φ^2 can be described as

$$V_\varphi^2 = \Big\{ \sum_{n\in\mathbb{Z}} \alpha_n\ \varphi(t-n)\ :\ \{\alpha_n\} \in \ell^2(\mathbb{Z}) \Big\}. \tag{1}$$

On the other hand, in many common situations the available data are samples of some filtered versions $f * \mathsf{h}_j$ of the signal f itself, where the average function h_j reflects the characteristics of the adquisition device.

Suppose that s convolution systems (linear time-invariant systems or filters in engineering jargon) $\mathcal{L}_j f = f * \mathsf{h}_j$, $j = 1, 2, \dots, s$, are defined on V_φ^2. Assume also that the sequence of samples $\{(\mathcal{L}_j f)(kr)\}_{k\in\mathbb{Z};\, j=1,2,\dots,s}$, where $r \in \mathbb{N}$, is available for any f in V_φ^2.

Mathematically, the generalized sampling problem consists of the stable recovery of any $f \in V_\varphi^2$ from the above sequence of samples, i.e., to obtain sampling formulas in V_φ^2 having the form

$$f(t) = \sum_{j=1}^{s} \sum_{k\in\mathbb{Z}} \big(\mathcal{L}_j f\big)(kr)\, S_j(t-kr), \quad t\in\mathbb{R}, \tag{2}$$

such that the sequence of reconstruction functions $\{S_j(\cdot - kr)\}_{k\in\mathbb{Z};\, j=1,2,\dots,s}$ is a frame for the shift-invariant space V_φ^2 (see, for instance, [3], [5], [6], [7], [9], [10], [15], [16], [17]).

In the present work we provide a generalization of the above problem in the following sense: Let $\{U^t\}_{t\in\mathbb{R}}$ denote a continuous *group of unitary operators* in $\mathcal{H}$ containing our unitary operator U (see Section C) below). For a fixed $a \in \mathcal{H}$, we consider the subspace of $\mathcal{H}$ given by

$$\mathcal{A}_a := \overline{\operatorname{span}} \big\{ U^n a,\ n\in\mathbb{Z} \big\}.$$

In case that the sequence $\{U^n a\}_{n\in\mathbb{Z}}$ is a Riesz sequence in $\mathcal{H}$ (see, for instance, a necessary and sufficient condition in [13]) we have

$$\mathcal{A}_a = \Big\{ \sum_{n\in\mathbb{Z}} \alpha_n U^n a : \{\alpha_n\} \in \ell^2(\mathbb{Z}) \Big\}.$$

On the other hand, for $b_j \in \mathcal{H},\ j = 1, 2 \dots, s$ we consider the linear operators $x \in \mathcal{H} \mapsto \mathcal{L}_j x \in C(\mathbb{R})$ defined on $\mathbb{R}$ as

$$\big(\mathcal{L}_j x\big)(t) := \langle x, U^t b_j \rangle_{\mathcal{H}}, \quad t\in\mathbb{R}. \tag{3}$$

These operators $\mathcal{L}_j$ can be seen as a generalization of the previous convolution systems.

II. Goals and procedure

Given $b_j \in \mathcal{A}_a,\ j = 1, 2 \dots, s$, our aim is to recover any $x \in \mathcal{A}_a$, in a stable way, by means of the sequence of generalized samples

$$\big\{ \big(\mathcal{L}_j x\big)(kr) \big\}_{k\in\mathbb{Z};\, j=1,2,\dots,s},$$

obtained from (3) (here r denotes a fixed number in $\mathbb{N}$). In order to do this we only deal with the discrete group $\{U^n\}_{n\in\mathbb{Z}}$ completely determined by U, but we might be in presence of a time jitter error, and then, the study of the continuous group of unitary operators $\{U^t\}_{t\in\mathbb{R}}$ becomes essential. Having as data a perturbed sequence of samples

$$\big\{ \big(\mathcal{L}_j x\big)(kr + \epsilon_{kj}) \big\}_{k\in\mathbb{Z};\, j=1,2,\dots,s},$$

with errors $\epsilon_{kj} \in \mathbb{R}$, again we want to recover $x \in \mathcal{A}_a$.

In order to attack these problems we have proceeded in the following steps:

(a) The study of when the sequence $\big\{U^{kr} b_j\big\}_{k\in\mathbb{Z};\, j=1,2,\dots,s}$ is a complete system, a Bessel sequence, a frame or a Riesz basis for $\mathcal{A}_a$.

(b) In the frame case, search for a family of dual frames of the form $\big\{U^{kr} c_j\big\}_{k\in\mathbb{Z};\, j=1,2,\dots,s}$, where $c_j \in \mathcal{A}_a,\ j =$

$1, 2 \ldots, s$, allowing to recover any $x \in \mathcal{A}_a$ by means of the sampling formula

$$x = \sum_{k\in\mathbb{Z}} \sum_{j=1}^{s} \big(\mathcal{L}_j x\big)(kr)\, U^{kr} c_j \quad \text{in } \mathcal{H}\,. \tag{4}$$

(c) Using the standard perturbation theory of frames (see Ref. [4]) and the group of unitary operators theory [2], [18], to find a condition on the error sequence $\{\epsilon_{kj}\}$ allowing the recovery of any $x \in \mathcal{A}_a$ by means of a sampling expansion as

$$x = \sum_{j=1}^{s} \sum_{k\in\mathbb{Z}} \big(\mathcal{L}_j x\big)(kr + \epsilon_{kj})\, C^{\boldsymbol{\epsilon}}_{k,j} \quad \text{in } \mathcal{H}\,, \tag{5}$$

where the sequence $\{C^{\boldsymbol{\epsilon}}_{k,j}\}_{k\in\mathbb{Z};\, j=1,2,\ldots,s}$ is a frame for $\mathcal{A}_a$.

At stages (a) and (b) we have used some borrowed ideas from [13]; mainly related to the stationary properties of a sequence of the form $\{U^n b\}_{n\in\mathbb{Z}}$, $b \in \mathcal{H}$, and the spectral measure associated with the (auto)-covariance function of b.

III. Main results

A. The study of the sequence $\big\{U^{kr} b_j\big\}_{k\in\mathbb{Z};\, j=1,2,\ldots,s}$

If for every $j = 1, 2, \ldots s$ the spectral measure in the integral representation of the (cross)-covariance function of the sequences $\{U^k a\}_{k\in\mathbb{Z}}$, $\{U^k b_j\}_{k\in\mathbb{Z}}$ has no singular part, we have the following representation

$$\langle U^k a, U^{nr} b_j\rangle = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{i(k-rn)\theta}\phi_{a,b_j}(e^{i\theta})d\theta.$$

where ϕ_{a,b_j} stands for the cross spectral density of the stationary correlated sequences $\{U^k a\}_{k\in\mathbb{Z}}$ and $\{U^k b_j\}_{k\in\mathbb{Z}}$. Consider the $s \times 1$ matrices of functions defined on the torus $\mathbb{T} := \{e^{i\theta} \,:\, \theta \in [-\pi, \pi)\}$

$$\Phi_{a,b}(e^{i\theta}) := \begin{pmatrix} \phi_{a,b_1}(e^{i\theta}) \\ \phi_{a,b_2}(e^{i\theta}) \\ \vdots \\ \phi_{a,b_s}(e^{i\theta}) \end{pmatrix},$$

and

$$\Psi^l_{a,b}(e^{i\theta}) := (D_r S^{-l}\Phi_{a,b})(e^{i\theta})\,, \quad l = 0, 1, \ldots, r-1\,,$$

where $D_r : L^2(\mathbb{T}) \to L^2(\mathbb{T})$ denotes the decimation operator

$$\sum_{k\in\mathbb{Z}} a_k e^{ik\theta} \overset{D_r}{\longmapsto} \sum_{k\in\mathbb{Z}} a_{rk} e^{ik\theta}$$

and $S : L^2(\mathbb{T}) \to L^2(\mathbb{T})$ denotes the (left) shift operator

$$\sum_{k\in\mathbb{Z}} a_k e^{ik\theta} \overset{S}{\longmapsto} \sum_{k\in\mathbb{Z}} a_{k+1} e^{ik\theta}\,.$$

Finally, defining the $s \times r$ matrix of functions on the torus $\mathbb{T}$

$$\Psi_{a,b}(e^{i\theta}) := \big(\Psi^0_{a,b}(e^{i\theta})\ \Psi^1_{a,b}(e^{i\theta})\ \ldots \Psi^{r-1}_{a,b}(e^{i\theta})\big)\,, \tag{6}$$

and its related constants,

$$\begin{aligned} A_\Psi &:= \operatorname*{ess\,inf}_{\zeta\in\mathbb{T}}\ \lambda_{\min}\big[\Psi^*_{a,b}(\zeta)\Psi_{a,b}(\zeta)\big]; \\ B_\Psi &:= \operatorname*{ess\,sup}_{\zeta\in\mathbb{T}}\ \lambda_{\max}\big[\Psi^*_{a,b}(\zeta)\Psi_{a,b}(\zeta)\big] \end{aligned} \tag{7}$$

we have the following result:

Theorem 3.1: Let b_j be in $\mathcal{A}_a$ for $j = 1, 2, \ldots, s$ and let $\Psi_{a,b}$ be the associated matrix given in (6) and its related constants (7). Then, the following results hold:

i) The sequence $\big\{U^{rk} b_j\big\}_{k\in\mathbb{Z};\, j=1,2,\ldots s}$ is a complete system in $\mathcal{A}_a$ if and only the rank of the matrix $\Psi_{a,b}(\zeta)$ is r a.e. ζ in $\mathbb{T}$.
ii) The sequence $\big\{U^{rk} b_j\big\}_{k\in\mathbb{Z};\, j=1,2,\ldots s}$ is a Bessel sequence for $\mathcal{A}_a$ if and only the constant $B_\Psi < \infty$.
iii) The sequence $\big\{U^{rk} b_j\big\}_{k\in\mathbb{Z};\, j=1,2,\ldots s}$ is a frame for $\mathcal{A}_a$ if and only if constants A_Ψ and B_Ψ satisfy $0 < A_\Psi \le B_\Psi < \infty$. In this case, A_Ψ and B_Ψ are the optimal frame bounds for $\big\{U^{rk} b_j\big\}_{k\in\mathbb{Z};\, j=1,2,\ldots s}$.
iv) The sequence $\big\{U^{rk} b_j\big\}_{k\in\mathbb{Z};\, j=1,2,\ldots s}$ is a Riesz basis for $\mathcal{A}_a$ if and only if it is a frame and $s = r$.

B. The frame expansion

Define the $r \times s$ matrix Γ of functions on $\mathbb{T}$ as

$$\Gamma(e^{i\theta}) := \sum_{k\in\mathbb{Z}} \Gamma_k e^{ik\theta} = [\Psi^*_{a,b}(e^{i\theta})\Psi_{a,b}(e^{i\theta})]^{-1}\Psi^*_{a,b}(e^{i\theta}). \tag{8}$$

Note that $\Psi^\dagger_{a,b}(e^{i\theta}) := [\Psi^*_{a,b}(e^{i\theta})\Psi_{a,b}(e^{i\theta})]^{-1}\Psi^*_{a,b}(e^{i\theta})$ stands for the Moore-Penrose left-inverse. In case that condition iii) in Theorem 3.1 is satisfied, we can define,

$$\widetilde{a}_n := \begin{pmatrix} U^{nr} a \\ U^{nr+1} a \\ \vdots \\ U^{nr+r-1} a \end{pmatrix}$$

and

$$\begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_s \end{pmatrix} := \sum_{k\in\mathbb{Z}} \Gamma_k^\top \widetilde{a}_k\,.$$

Note that, under condition iii) in Theorem 3.1, the matrix $\Gamma(e^{i\theta})$ has entries in $L^\infty(\mathbb{T})$.

Then, the sequences $\{U^{kr} c_j\}_{k\in\mathbb{Z};\, j=1,2,\ldots,s}$ and $\{U^{kr} b_j\}_{k\in\mathbb{Z};\, j=1,2,\ldots s}$ are a pair of dual frames for $\mathcal{A}_a$. Hence we obtain the following recovery formula in $\mathcal{A}_a$: For any $x \in \mathcal{A}_a$, the expansion

$$x = \sum_{j=1}^{s} \sum_{k\in\mathbb{Z}} \langle x, U^{kr} b_j\rangle U^{kr} c_j \quad \text{in } \mathcal{H}$$

holds.

The analysis done provides a whole family of dual frames; in fact, everything works if we choose in (8) a matrix of the form

$$\Gamma_{\mathbb{U}}(e^{i\theta}) := \Psi^\dagger_{a,b}(e^{i\theta}) + \mathbb{U}(e^{i\theta})\big[\mathbb{I}_s - \Psi_{a,b}(e^{i\theta})\Psi^\dagger_{a,b}(e^{i\theta})\big]\,,$$

where $\mathbb{U}(e^{i\theta})$ denotes any $r \times s$ matrix with entries in $L^\infty(\mathbb{T})$, and $\Psi^\dagger_{a,b}$ the Moore-Penrose left pseudo-inverse.

Notice that if $s = r$, $\Psi^\dagger_{a,b} = \Psi^{-1}_{a,b}$ which implies that Γ is unique and we are in presence of a pair of dual Riesz basis.

Remark: In Theorem 3.1 we have assumed that b_j belongs to $\mathcal{A}_a$ for each $j = 1, 2, \ldots, s$ since we want the sequence $\{U^{rk} b_j\}_{k\in\mathbb{Z};\, j=1,2,\ldots s}$ to be contained in $\mathcal{A}_a$. In case that some $b_j \notin \mathcal{A}_a$, the sequence $\{U^{rk} b_j\}_{k\in\mathbb{Z};\, j=1,2,\ldots s}$ is not necessarily contained in $\mathcal{A}_a$. However, whenever $0 < A_\Psi \leq B_\Psi < \infty$, the inequalities

$$A_\Psi \|x\|^2 \leq \sum_{j=1}^{s} \sum_{k\in\mathbb{Z}} |\langle x, U^{rk} b_j\rangle|^2 \leq B_\Psi \|x\|^2 \quad \text{for all } x \in \mathcal{A}_a$$

hold, and conversely. Hence, the sequence $\{U^{rk} b_j\}_{k\in\mathbb{Z};\, j=1,2,\ldots s}$ is a pseudo-frame for $\mathcal{A}_a$ (see Refs. [11], [12]).

Denoting by $P_{\mathcal{A}_a}$ the orthogonal projection onto $\mathcal{A}_a$, since for each $x \in \mathcal{A}_a$ we have

$$\langle x, U^{rk} b_j\rangle = \big\langle x, P_{\mathcal{A}_a}\big(U^{rk} b_j\big)\big\rangle\,,\ k \in \mathbb{Z} \text{ and } j = 1, 2, \ldots, s\,,$$

and, as a consequence, Theorem 3.1 can be reformulated in terms $\big\{P_{\mathcal{A}_a}\big(U^{rk} b_j\big)\big\}_{k\in\mathbb{Z};\, j=1,2,\ldots s}$, a sequence in $\mathcal{A}_a$.

C. The study of the time jitter error

In Sections A) and B) it is not strictly necessary to have a group of unitary operators $\{U^t\}_{t\in\mathbb{R}}$ to obtain the announced results. However, in order to deal with the time-jitter error this formalism becomes essential in our approach.

Let $\{U^t\}_{t\in\mathbb{R}}$ denote a continuous *group of unitary operators* in $\mathcal{H}$ containing our unitary operator U, i.e., say for instance $U := U^1$. Recall that $\{U^t\}_{t\in\mathbb{R}}$ is a family of unitary operators in $\mathcal{H}$ satisfying (see Ref. [2, vol. 2; p. 29]):

1) $U^t\, U^{t'} = U^{t+t'}$,
2) $U^0 = I_{\mathcal{H}}$,
3) $\langle U^t x, y\rangle_{\mathcal{H}}$ is a continuous function of t for any $x, y \in \mathcal{H}$.

Note that $(U^t)^{-1} = U^{-t}$, and since $(U^t)^* = (U^t)^{-1}$, we have $(U^t)^* = U^{-t}$.

Classical Stone's theorem [14] assures us the existence of a self-adjoint operator T (possibly unbounded) such that $U^t \equiv \mathrm{e}^{\mathrm{i}tT}$. This self-adjoint operator T, defined on the dense domain of $\mathcal{H}$

$$D_T := \big\{x \in \mathcal{H} \text{ such that } \int_{-\infty}^{\infty} w^2\, d\|E_w x\|^2 < \infty\big\}\,,$$

admits the *spectral representation* $T = \int_{-\infty}^{\infty} w\, dE_w$ which means:

$$\langle Tx, y\rangle = \int_{-\infty}^{\infty} w\, d\langle E_w x, y\rangle \quad \text{for any } x \in D_T \text{ and } y \in \mathcal{H}\,,$$

where $\{E_w\}_{w\in\mathbb{R}}$ is the corresponding *resolution of the identity*, i.e., a one-parameter family of projection operators E_w in $\mathcal{H}$ such that

1) $E_{-\infty} := \lim_{w\to-\infty} E_w = O_{\mathcal{H}}$, $\quad E_\infty := \lim_{w\to\infty} E_w = I_{\mathcal{H}}$,
2) $E_{w^-} = E_w$ for every $-\infty < w < \infty$,
3) $E_u\, E_v = E_w$ where $w = \min\{u, v\}$.

Recall that $\|E_w x\|^2$ and $\langle E_w x, y\rangle$, as functions of w, have bounded variation and define, respectively, a positive and a complex Borel measure on $\mathbb{R}$.

Furthermore, for any $x \in D_T$ we have that $\lim_{t\to 0} \dfrac{U^t x - x}{t} = \mathrm{i}\, Tx$ and the operator T is said to be the *infinitesimal generator* of the group $\{U^t\}_{t\in\mathbb{R}}$. For each $x \in D_T$, $U^t x$ is a continuous differentiable function of t. Notice that, whenever the self-adjoint operator T is bounded, $D_T = \mathcal{H}$ and $\mathrm{e}^{\mathrm{i}tT}$ can be defined as the usual exponential series; in any case, $U^t \equiv \mathrm{e}^{\mathrm{i}tT}$ means that

$$\langle U^t x, y\rangle = \int_{-\infty}^{\infty} \mathrm{e}^{\mathrm{i}wt} d\langle E_w x, y\rangle\,, \quad t \in \mathbb{R}\,,$$

where $x \in D_T$ and $y \in \mathcal{H}$.

The following result on frame perturbation, which proof can be found in [4, p. 354] has been used:

Lemma 3.2: Let $\{x_n\}_{n=1}^\infty$ be a frame for the Hilbert space $\mathcal{H}$ with frame bounds A, B, and let $\{y_n\}_{n=1}^\infty$ be a sequence in $\mathcal{H}$. If there exists a constant $R < A$ such that

$$\sum_{n=1}^{\infty} |\langle x_n - y_n, x\rangle|^2 \leq R\|x\|^2 \ \text{ for each } x \in \mathcal{H}\,,$$

then the sequence $\{y_n\}_{n=1}^\infty$ is also a frame for $\mathcal{H}$ with bounds $A\big(1 - \sqrt{R/A}\big)^2$ and $B\big(1 + \sqrt{R/B}\big)^2$. If $\{x_n\}_{n=1}^\infty$ is a Riesz basis, then $\{y_n\}_{n=1}^\infty$ is a Riesz basis.

Thus, we have the following result:

Theorem 3.3: Assume that for some $b_j \in D_T$, i.e., $\int_{-\infty}^{\infty} w^2 d\|E_w b_j\|^2 < \infty$ for each $1 \leq j \leq r$, the sequence $\{U^{kr} b_j\}_{k\in\mathbb{Z};\, j=1,2,\ldots,r}$ is a Riesz basis for $\mathcal{A}_a$ with Riesz bounds $0 < A_\Psi \leq B_\Psi < \infty$. For a sequence $\boldsymbol{\epsilon} := \{\epsilon_{kj}\}_{k\in\mathbb{Z},\, j=1,2,\ldots,r}$ of errors, let R be the constant given by

$$R := \|\boldsymbol{\epsilon}\|^2 \max_{j=1,2,\ldots,r} \Big\{ \int_{-\infty}^{\infty} w^2 d\|E_w b_j\|^2 \Big\}\,,$$

where $\|\boldsymbol{\epsilon}\|$ denotes the ℓ^2_s-norm of the sequence $\boldsymbol{\epsilon}$.

If $R < A_\Psi$, then the sequence $\{U^{kr+\epsilon_{kj}} b_j\}_{k\in\mathbb{Z};\, j=1,2,\ldots,r}$ is a Riesz sequence in $\mathcal{H}$ with Riesz bounds $A_\Psi\big(1 - \sqrt{R/A_\Psi}\big)^2$ and $B_\Psi\big(1 + \sqrt{R/B_\Psi}\big)^2$.

Next, we deal with the problem of the recovery of any $x \in \mathcal{A}_a$ in a stable way from the perturbed sequence

$$\{(\mathcal{L}_j x)(kr + \epsilon_{kj})\}_{k\in\mathbb{Z};\, j=1,2,\ldots,s}\,,$$

where $\boldsymbol{\epsilon} := \{\epsilon_{kj}\}_{k\in\mathbb{Z};\, j=1,2,\ldots,s}$ denotes a sequence of real errors.

Taking into account the $L^2(0,1)$ functions

$$g_j(w) := \sum_{k\in\mathbb{Z}} \langle a, U^k b_j\rangle_{\mathcal{H}}\, \mathrm{e}^{2\pi\mathrm{i}kw}\,,\ j = 1, 2, \ldots, s\,, \quad (9)$$

we can define the $s \times r$ matrix

$$\mathbb{G}(w) := \Big[g_j\Big(w + \frac{k-1}{r}\Big)\Big]_{\substack{j=1,2,\ldots,s\\ k=1,2,\ldots,r}}$$

and its related the constants $\alpha_{\mathbb{G}}$ and $\beta_{\mathbb{G}}$ are given by

$$\alpha_{\mathbb{G}} := \operatorname*{ess\,inf}_{w\in(0,1/r)} \lambda_{\min}[\mathbb{G}^*(w)\mathbb{G}(w)],$$
$$\beta_{\mathbb{G}} := \operatorname*{ess\,sup}_{w\in(0,1/r)} \lambda_{\max}[\mathbb{G}^*(w)\mathbb{G}(w)]\,.$$

It is worth to mention that in [9] was proved that the sequence $\{\overline{g_j(w)}\,\mathrm{e}^{2\pi \mathrm{i} rnw}\}_{n\in\mathbb{Z};\,j=1,2,\ldots,s}$ is a frame for $L^2(0,1)$ if and only if $0 < \alpha_{\mathbb{G}} \le \beta_{\mathbb{G}} < \infty$. The idea is to consider the sequence $\left\{g_{m,j}(w)\,\mathrm{e}^{2\pi \mathrm{i} rmw}\right\}_{m\in\mathbb{Z};\,j=1,2,\ldots s}$ as a perturbation of the above frame in $L^2(0,1)$, where

$$g_{m,j}(w) := \sum_{k\in\mathbb{Z}} \langle a, U^{k+\epsilon_{mj}} b_j\rangle_{\mathcal{H}}\,\mathrm{e}^{2\pi \mathrm{i} kw}\,,\ j=1,2,\ldots,s\,.$$

For $|\gamma| < 1/2$, define the functions,

$$M_{a,b_j}(\gamma) := \sum_{k\in\mathbb{Z}} \max_{t\in[-\gamma,\gamma]} |\langle a, U^{k+t} b_j\rangle - \langle a, U^k b_j\rangle|\,,$$

and

$$N_{a,b_j}(\gamma) := \max_{k=0,1,\ldots,r-1} \sum_{m\in\mathbb{Z}} \max_{t\in[-\gamma,\gamma]} |\langle a, U^{rm+k+t} b_j\rangle - \langle a, U^{rm+k} b_j\rangle|\,.$$

Notice that $N_{a,b_j}(\gamma) \le M_{a,b_j}(\gamma)$ and for $r=1$ the equality holds. Moreover, assuming that the continuous functions $\varphi_j(t) := \langle a, U^t b_j\rangle$, $j=1,2,\ldots,s$, satisfy a decay condition as $\varphi_j(t) = O\big(|t|^{-(1+\eta_j)}\big)$ when $|t|\to\infty$ for some $\eta_j > 0$, we deduce that the functions $N_{a,b_j}(\gamma)$ and $M_{a,b_j}(\gamma)$ are continuous near to 0.

Theorem 3.4: Assume that for the functions g_j, $j = 1,2,\ldots,s$, given in (9) we have $0 < \alpha_{\mathbb{G}} \le \beta_{\mathbb{G}} < \infty$. For an error sequence $\boldsymbol{\epsilon} := \{\epsilon_{mj}\}_{m\in\mathbb{Z};\,j=1,\ldots,s}$, define the constant $\gamma_j := \sup_{m\in\mathbb{Z}} |\epsilon_{mj}|$ for each $j=1,2,\ldots,s$. Then the condition $\sum_{j=1}^s M_{a,b_j}(\gamma_j) N_{a,b_j}(\gamma_j) < \alpha_{\mathbb{G}}/r$ implies that there exists a frame $\{C^{\boldsymbol{\epsilon}}_{m,j}\}_{m\in\mathbb{Z};\,j=1,2,\ldots,s}$ for $\mathcal{A}_a$ such that, for any $x\in\mathcal{A}_a$, the sampling expansion

$$x = \sum_{j=1}^{s}\sum_{m\in\mathbb{Z}} \langle x, U^{rm+\epsilon_{mj}} b_j\rangle_{\mathcal{H}}\, C^{\boldsymbol{\epsilon}}_{m,j} \quad \text{in } \mathcal{H}\,, \qquad (10)$$

holds. Moreover, when $r = s$ the sequence $\{C^{\boldsymbol{\epsilon}}_{m,j}\}_{m\in\mathbb{Z};\,j=1,2,\ldots,s}$ is a Riesz basis for $\mathcal{A}_a$, and the interpolation property $\left\langle C^{\boldsymbol{\epsilon}}_{n,j}, U^{rm+\epsilon_{ml}} b_l\right\rangle_{\mathcal{H}} = \delta_{j,l}\,\delta_{n,m}$ holds.

Sampling formula (10) is useless from a practical point of view: it is impossible to determine the involved frame $\{C^{\boldsymbol{\epsilon}}_{m,j}\}_{m\in\mathbb{Z};\,j=1,2,\ldots,s}$. As a consequence, in order to recover $x\in\mathcal{A}_a$ from the sequence of inner products $\left\{\langle x, U^{rm+\epsilon_{mj}} b_j\rangle_{\mathcal{H}}\right\}_{m\in\mathbb{Z};\,j=1,2,\ldots,s}$ we could implement a frame algorithm in $\ell^2(\mathbb{Z})$. Another possibility is given in the recent Ref. [1].

IV. Conclusion

By way of conclusion we may say that we have obtained a complete characterization of the sequence $\left\{U^{kr} b_j\right\}_{k\in\mathbb{Z};\,j=1,2,\ldots,s}$ in $\mathcal{A}_a$, where $b_j\in\mathcal{A}_a$, $1\le j\le s$. We have found a necessary and sufficient condition ensuring that it is a complete system, a Bessel sequence, a frame or a Riesz basis for $\mathcal{A}_a$.

In the case that this sequence is a frame for $\mathcal{A}_a$ we can give an explicit family of dual frames allowing to recover any $x\in\mathcal{A}_a$ by means of a sampling formula like (4).

Concerning the perturbation framework, we have found a condition related to the ℓ^2-norm of $\boldsymbol{\epsilon} = \{\epsilon_{kj}\}_{k\in\mathbb{Z};\,j=1,2,\ldots,s}$ and the $\max_{j=1,2,\ldots,s}\left\{\int_{-\infty}^{\infty} w^2 d\|E_w b_j\|^2\right\}$ such that the sequence $\left\{U^{kr+\epsilon_{kj}} b_j\right\}_{k\in\mathbb{Z};\,j=1,2,\ldots,s}$ is a Riesz sequence in $\mathcal{H}$ and we have obtained a sampling expansion allowing us to recover any $x\in\mathcal{A}_a$ in a stable way from the perturbed sequence of samples $\{(\mathcal{L}_j x)(kr+\epsilon_{kj})\}_{k\in\mathbb{Z};\,j=1,2,\ldots,s}$.

Acknowledgment

This work has been supported by the grant MTM2009–08345 from the Spanish *Ministerio de Ciencia e Innovación* (MICINN).

References

[1] B. Adcock and A. C. Hansen. Stable reconstructions in Hilbert spaces and the resolution of the Gibbs phenomenon. *Appl. Comput. Harmon. Anal.*, 32:357–388, 2012.

[2] N. I. Akhiezer and I. M. Glazman. Theory of linear operators in Hilbert space. Dover Publications, New York, 1993.

[3] A. Aldroubi and K. Gröchenig. Non-uniform sampling and reconstruction in shift-invariant spaces. *SIAM Rev.*, 43:585–620, 2001.

[4] O. Christensen. An Introduction to Frames and Riesz Bases. Birkhäuser, Boston, 2003.

[5] O. Christensen and Y. C. Eldar. Oblique dual frames and shift-invariant spaces. *Appl. Comput. Harmon. Anal.*, 17(1):48–68, 2004.

[6] O. Christensen and Y. C. Eldar. Generalized shift-invariant systems and frames for subspaces. *Appl. Comput. Harmon. Anal.*, 11(3):299–313, 2005.

[7] H. R. Fernández-Morales, A. G. García and G. Pérez-Villalón. Generalized sampling in $L^2(\mathbb{R}^d)$ shift-invariant subspaces with multiple stable generators. Multiscale Signal Analysis and Modeling, Lecture Notes in Electrical Engineering, Springer, New York, 2012.

[8] H. R. Fernández-Morales, A. G. García, M. A. Hernández-Medina and M. J. Muñoz-Bouzo. Generalized sampling: from shift-invariant to U-invariant spaces. Submitted 2013.

[9] A. G. García and G. Pérez-Villalón. Dual frames in $L^2(0,1)$ connected with generalized sampling in shift-invariant spaces. *Appl. Comput. Harmon. Anal.*, 20(3):422–433, 2006.

[10] A. G. García, M. A. Hernández-Medina and G. Pérez-Villalón. Generalized sampling in shift-invariant spaces with multiple stable generators. *J. Math. Anal. Appl.*, 337:69–84, 2008.

[11] S. Li and H. Ogawa. Pseudo-Duals of frames with applications. *Appl. Comput. Harmon. Anal.*, 11:289–304, 2001.

[12] S. Li and H. Ogawa. Pseudoframes for subspaces with applications. *J. Fourier Anal. Appl.*, 10(4):409–431, 2004.

[13] V. Pohl and H. Boche. U-invariant sampling and reconstruction in atomic spaces with multiple generators. IEEE Trans. Signal Process., **60**(7), 3506–3519, 2012.

[14] M. H. Stone. On one-parameter unitary groups in Hilbert spaces. Ann. Math., 33(3):643–648, 1932.

[15] W. Sun and X. Zhou. Average sampling in shift-invariant subspaces with symmetric averaging functions. *J. Math. Anal. Appl.*, 287:279–295, 2003.

[16] M. Unser and A. Aldroubi. A general sampling theory for non ideal acquisition devices. *IEEE Trans. Signal Process.*, 42(11):2915–2925, 1994.

[17] G. G. Walter. A sampling theorem for wavelet subspaces. *IEEE Trans. Inform. Theory*, 38:881–884, 1992.

[18] J. Weidmann. Linear Operators in Hilbert Spaces Springer, New York, 1980.

Iterative Hard Thresholding with Near Optimal Projection for Signal Recovery

Raja Giryes and Michael Elad
Computer Science Department
Technion - IIT 32000, Haifa, ISRAEL
Email: [raja,elad]@cs.technion.ac.il

***Abstract*—Recovering signals that have sparse representations under a given dictionary from a set of linear measurements got much attention in the recent decade. However, most of the work has focused on recovering the signal's representation, forcing the dictionary to be incoherent and with no linear dependencies between small sets of its columns. A series of recent papers show that such dependencies can be allowed by aiming at recovering the signal itself. However, most of these contributions focus on the analysis framework. One exception to these is the work reported in [1], proposing a variant of the CoSaMP for the synthesis model, and showing that signal recovery is possible even in high-coherence cases. In the theoretical study of this technique the existence of an efficient near optimal projection scheme is assumed. In this paper we extend the above work, showing that under very similar assumptions, a variant of IHT can recover the signal in cases where regular IHT fails.**

I. Introduction

Recovering a sparse signal from a given set of linear measurements has been a major subject of research in recent years. In the basic setup, an unknown signal $\mathbf{x}_0 \in \mathbb{R}^d$ passes through a given linear transformation $\mathbf{M} \in \mathbb{R}^{m\times d}$ with an additive noise $\mathbf{e} \in \mathbb{R}^m$ providing a set of linear measurements $\mathbf{y} = \mathbf{M}\mathbf{x}_0 + \mathbf{e}$. The signal $\mathbf{x}_0$ is assumed to have a k-sparse representation $\boldsymbol{\alpha}_0 \in \mathbb{R}^n$ under a given dictionary $\mathbf{D} \in \mathbb{R}^{d\times n}$, i.e. $\mathbf{x}_0 = \mathbf{D}\boldsymbol{\alpha}_0$, $\|\boldsymbol{\alpha}_0\|_0 \leq k$ and $k \ll d$, where $\|\cdot\|_0$ is the "ℓ_0-norm" that counts the number of non-zero entries in a vector. The sparsity prior results with the following minimization problem

$$\min_{\boldsymbol{\alpha}} \|\mathbf{y} - \mathbf{M}\mathbf{D}\boldsymbol{\alpha}\|_2 \quad s.t. \quad \|\boldsymbol{\alpha}\|_0 \leq k, \tag{1}$$

in which we pursue the representation $\boldsymbol{\alpha}$ in order to recover the original signal $\mathbf{x}_0$ from $\mathbf{y}$. Given a reconstructed representation $\hat{\boldsymbol{\alpha}}$, the estimation for the signal is simply given by $\hat{\mathbf{x}} = \mathbf{D}\hat{\boldsymbol{\alpha}}$.

Solving (1) is a NP-hard problem and many approximation techniques has been proposed for it [2]. One of these is the iterative hard thresholding (IHT) algorithm [3]. This approach, summarized in Algorithm 1, recovers the representation in an iterative way using two repeating steps: (i) Gradient step: moving in the optimal gradient direction for minimizing $\|\mathbf{y} - \mathbf{M}\mathbf{D}\boldsymbol{\alpha}\|_2$; (ii) Projection step: ensuring that the representation estimate is k-sparse. The operator $\mathrm{supp}(\cdot, k)$ returns the support of the largest k elements in a given vector and the subscript T for a vector/matrix means taking the entries/columns corresponding to the indices in T.

In order to evaluate the performance of IHT, the restricted isometry property (RIP) [4] of the matrix $\mathbf{MD}$ is used. A matrix $\mathbf{A} \in \mathbb{R}^{d\times n}$ satisfies the RIP with a constant δ_k if for any k sparse vector $\boldsymbol{\alpha} \in \mathbb{R}^n$

$$(1-\delta_k)\|\boldsymbol{\alpha}\|_2^2 \leq \|\mathbf{M}\mathbf{D}\boldsymbol{\alpha}\|_2^2 \leq (1+\delta_k)\|\boldsymbol{\alpha}\|_2^2 . \tag{2}$$

With this definition in hand it has been shown that if $\delta_{2k} \leq 1/4$ or $\delta_{3k} \leq 1/\sqrt{3}$ then IHT recovers the representation stably, i.e.,

$$\|\hat{\boldsymbol{\alpha}}_{\text{IHT}} - \boldsymbol{\alpha}_0\|_2 \leq c_{\text{IHT}} \|\mathbf{e}\|_2 , \tag{3}$$

where $c_{\text{IHT}} > 2$ is a function of δ_{2k} and δ_{3k} [3], [5], [6]. Note that with no prior on the noise distribution only a stable recovery is guaranteed with no noise reduction effect. The latter can be achieved by adding an assumption on the noise distribution [7]. This work deals only with the former case where $\mathbf{e}$ is an adversarial bounded noise.

Algorithm 1 Iterative hard thresholding (IHT)

Require: $k, \mathbf{M}, \mathbf{D}, \mathbf{y}$ where $\mathbf{y} = \mathbf{M}\mathbf{D}\boldsymbol{\alpha}_0 + \mathbf{e}$, k is the cardinality of $\boldsymbol{\alpha}_0$ and $\mathbf{e}$ is an additive noise.
Ensure: $\hat{\boldsymbol{\alpha}}_{\text{IHT}}$: k-sparse approximation of $\boldsymbol{\alpha}_0$.
Initialize representation $\hat{\boldsymbol{\alpha}}^0 = \mathbf{0}$ and set $t = 0$.
while halting criterion is not satisfied **do**
$t = t + 1$.
Perform a gradient step: $\boldsymbol{\alpha}_g = \hat{\boldsymbol{\alpha}}^{t-1} + \mu^t \mathbf{M}\mathbf{D}^*(\mathbf{y} - \mathbf{M}\mathbf{D}\hat{\boldsymbol{\alpha}}^{t-1})$
Find a new support: $T^t = \mathrm{supp}(\boldsymbol{\alpha}_g, k)$
Calculate a new representation: $\hat{\boldsymbol{\alpha}}^t = (\boldsymbol{\alpha}_g)_{T^t}$.
end while
Form the final solution $\hat{\boldsymbol{\alpha}}_{\text{IHT}} = \hat{\boldsymbol{\alpha}}^t$.

Note that in the case where $\mathbf{D}$ contains k correlated columns we have $\delta_k \geq 1$. Then the above recovery conditions fail and (3) does not hold. The reason for this is that in the presence of linear dependencies between a small group of columns from $\mathbf{D}$, the representation is no longer unique [8] and the solution of (1) is no longer stable [4]. Though the recovery of the representation is not achievable in the presence of correlations within $\mathbf{D}$, we should keep in mind that our task is to estimate the signal and not the representation. Recovering the wrong support of $\boldsymbol{\alpha}$, but one that is closely related to the original signal may suffice for our needs.

This key point is contained in the union of subspaces literature [9], [10], [11]. However, it has been pointed out more clearly in a series of contributions for the analysis framework [12], [13], [14], [15], [16], assuming a different sparse model. As such, correlations in the analysis dictionary were found to pose no problem and it has been demonstrated that such are even an advantage [14], [15], [16].

The analysis results serve as a clue that the same may happen in the synthesis model when the signal is the objective. In particular, the condition in [12] are presented in terms of the $\mathbf{D}$-RIP, which is a property of the measurement matrix $\mathbf{M}$ for the synthesis model. However, as indicated in [15], the results in [12] essentially hold true for signals emerging from the analysis model.

The work reported in [1] is very different from all the above, in addressing the synthesis model, providing signal recovery guarantees using the $\mathbf{D}$-RIP. This work presents a modified version of CoSaMP, Signal space CoSaMP (SSCoSaMP), that aims at recovering the signal, showing empirically that unlike the regular CoSaMP, the modified version gets a good recovery even in the presence of linear dependencies in $\mathbf{D}$. The authors of [1] use a similar proof technique to the one in [15] that was derived for the analysis CoSaMP (ACoSaMP). Just like [15], the work in [1] relies on the availability of near-optimal projection (this property will be defined clearly in the

next section). Another recent paper that exploits the $\mathbf{D}$-RIP in the context of the synthesis model is the one reported in [17], studying the basic synthesis ℓ_0-minimization problem.

In this work we continue with the same assumption as in [1] – the existence of a near optimal projection scheme[1] – and use the $\mathbf{D}$-RIP too. In Section II we present notations, and the definitions of the $\mathbf{D}$-RIP and the near-optimality of a projection. In Section III we introduce the signal space IHT (SSIHT) method for signal recovery and in Section IV we propose theoretical guarantees for it, relying on ideas taken from [15]. The SSIHT emerges from IHT as SSCoSaMP emerges from CoSaMP. The novelty of this work is by its theoretical study which relies on [15] and differs from [1]. Note that the proof technique used here can be adopted to develop new theoretical results for SSCoSaMP that differ from those in [1] and resemble those of ACoSaMP [15]. Section V presents some numerical results showing the advantage of SSIHT over IHT for the task of signal recovery.

II. Preliminaries

We start with the definition of the $\mathbf{D}$-RIP. As indicated in [12], many types of random matrices satisfy this property with a small δ_k[2].

Definition 2.1: A matrix $\mathbf{M}$ obeys the $\mathbf{D}$-RIP with a constant $\delta_k^{\mathbf{D}}$, if $\delta_k^{\mathbf{D}}$ is the smallest constant that satisfies

$$(1-\delta_k^{\mathbf{D}})\left\|\mathbf{z}\right\|_2^2 \leq \left\|\mathbf{M}\mathbf{z}\right\|_2^2 \leq (1+\delta_k^{\mathbf{D}})\left\|\mathbf{z}\right\|_2^2 \tag{4}$$

for any $\mathbf{z}\in\mathbb{R}^d$ such that $\mathbf{z}=\mathbf{D}\boldsymbol{\alpha}$ and $\left\|\boldsymbol{\alpha}\right\|_0 \leq k$.

Another definition we need is the one of a near optimal projection. In SSIHT we face the following problem: Given a general vector $\mathbf{z}\in\mathbb{R}^d$, we seek the closest vector to it, in the ℓ_2-norm sense, that has a k-sparse representation. Note that given a support set T, the closest vector is computed simply by using an orthogonal projection $\mathbf{P}_T=\mathbf{D}_T\mathbf{D}_T^{\dagger}$ onto it. Thus, the problem of finding the closest vector turns into the problem of finding its support, using the scheme

$$\mathcal{S}_k^*(\mathbf{z}) = \underset{T,|T|\leq k}{\operatorname{argmin}} \left\|\mathbf{z}-\mathbf{P}_T\mathbf{z}\right\|_2^2, \tag{5}$$

where the closest vector with k-sparse representation for $\mathbf{z}$ is simply given by $\mathbf{P}_{\mathcal{S}_k^*(\mathbf{z})}\mathbf{z}$. We should remark that for the task of projecting a given representation vector to the same domain (k-sparse vectors), a simple hard thresholding as done in IHT gives the ideal solution. However, finding the optimal support in the signal case seems to be a NP-hard problem as its equivalent form in analysis context is known to be so [18]. Thus an approximation procedure is needed. For this purpose we introduce the definition of a near-optimal projection [15].

Definition 2.2: A procedure $\hat{\mathcal{S}}_k$ implies a near-optimal projection $\mathbf{P}_{\hat{\mathcal{S}}_k(\cdot)}$ with a constant C_k if for any $\mathbf{z}\in\mathbb{R}^d$

$$\left\|\mathbf{z}-\mathbf{P}_{\hat{\mathcal{S}}_k(\mathbf{z})}\mathbf{z}\right\|_2^2 \leq C_k\left\|\mathbf{z}-\mathbf{P}_{\mathcal{S}_k^*(\mathbf{z})}\mathbf{z}\right\|_2^2. \tag{6}$$

In [1], a slightly different definition was used:

Definition 2.3: A procedure $\hat{\mathcal{S}}_k$ implies a near-optimal projection $\mathbf{P}_{\hat{\mathcal{S}}_k(\cdot)}$ with constants $C_{k,1}$ and $C_{k,2}$ if for any $\mathbf{z}\in\mathbb{R}^d$

$$\begin{aligned}&\left\|(\mathbf{P}_{\mathcal{S}_k^*(\mathbf{z})}-\mathbf{P}_{\hat{\mathcal{S}}_k(\mathbf{z})})\mathbf{z}\right\|_2 \\ &\leq \min\left\{C_{k,1}\left\|\mathbf{P}_{\mathcal{S}_k^*(\mathbf{z})}\mathbf{z}\right\|_2, C_{k,2}\left\|\mathbf{z}-\mathbf{P}_{\mathcal{S}_k^*(\mathbf{z})}\mathbf{z}\right\|_2\right\}.\end{aligned} \tag{7}$$

Having these definitions we recall the problem we aim at solving:

Definition 2.4 (Problem $\mathcal{P}$): Consider a measurement vector $\mathbf{y}\in\mathbb{R}^m$ such that $\mathbf{y}=\mathbf{M}\mathbf{x}_0+\mathbf{e}$ where $\mathbf{x}_0\in\mathbb{R}^d$ has a k-sparse representation under $\mathbf{D}$, $\mathbf{M}\in\mathbb{R}^{m\times d}$ is a degradation operator and $\mathbf{e}\in\mathbb{R}^m$ is a bounded additive noise. The largest singular value of $\mathbf{M}$ is $\sigma_{\mathbf{M}}$ and its $\mathbf{D}$-RIP constant is δ_k. The dictionary $\mathbf{D}\in\mathbb{R}^{p\times d}$ is given and fixed. A procedure $\hat{\mathcal{S}}_k$ is assumed to be available. Our task is to recover $\mathbf{x}_0$ from $\mathbf{y}$. The recovery result is denoted by $\hat{\mathbf{x}}$.

The following guarantee has been proposed in [1] for SSCoSaMP.

Theorem 2.5 (Theorem 2.1 in [1]): Consider the problem $\mathcal{P}$ and assume $\hat{\mathcal{S}}_k$ implies a near optimal projection with constants $C_{k,1}$ and $C_{k,2}$. After t iterations of SSCoSaMP, its signal estimate $\hat{\mathbf{x}}^t$ obeys

$$\left\|\hat{\mathbf{x}}^t-\mathbf{x}_0\right\|_2 \leq c_1\left\|\hat{\mathbf{x}}^{t-1}-\mathbf{x}_0\right\|_2 + c_2\left\|\mathbf{e}\right\|_2, \tag{8}$$

where $c_1=((2+C_{k,1})\delta_{4k}+C_{k,1})(2+C_{k,2})\sqrt{\frac{1+\delta_{4k}}{1-\delta_{4k}}}$ and $c_2=\frac{(2+C_{k,2})((2+C_{k,1})(1+\delta_{4k})+2)}{\sqrt{1-\delta_{4k}}}$.

Assuming $C_{k,1}=0.1$ and $C_{k,2}=1$ like in Corollary 2.1 in [1], a condition for $c_1<1$ is $\delta_{4k}<0.096$ which guarantees that after a finite number of iterations we have

$$\left\|\hat{\mathbf{x}}_{\text{SSCoSaMP}}-\mathbf{x}_0\right\|_2 \leq c_{\text{SSCoSaMP}}\left\|\mathbf{e}\right\|_2, \tag{9}$$

where c_{SSCoSaMP} is a function of c_1, c_2 and δ_{4k}. The bound in (9) implies a stable recovery of SSCoSaMP.

In this paper we show that under similar assumptions on the near optimality constant C_k of definition 2.2 and the maximal singular value of $\mathbf{M}$, $\sigma_{\mathbf{M}}$, the condition $\delta_{2k}<0.289$ guarantees a stable signal reconstruction for SSIHT. Note that in the condition of SSCoSaMP, two near optimality constants are involved. The second one is related to C_k as both of them measure the projection error and it is easy to show that they obey the inequality $(C_{k,2}-1)^2\leq C_k\leq(1+C_{k,2})^2$. The first constant $C_{k,1}$ measures the energy kept in the projection. This constant's relation to the other two depends on the initial norm of the projected signal. Since there is no direct relation between C_k and $C_{k,1}$, it is natural that another constant of the system would appear in our recovery conditions and indeed $\sigma_{\mathbf{M}}$ takes this role.

The existence of a general near-optimal projection scheme for any given dictionary is still an open problem and is left for future work. It is likely that there are non-trivial examples for which an efficient procedure exists as has been shown in [15] for the analysis case. In practice, any sparse recovery algorithm can be used in order to determine the support for the projection scheme. In this work we use a simple thresholding rule: For a given signal $\mathbf{z}$ it chooses the support to be the largest entries in $\mathbf{D}^*\mathbf{z}$. We show empirically that with this scheme we recover signals using SSIHT that cannot be recovered using the regular IHT. Note that thresholding does not have any known (near) optimality guarantee except for unitary operators.

III. Signal Space Iterative Hard Thresholding

SSIHT is presented in Algorithm 2. Its main difference from the regular IHT is the projection scheme. As IHT works in the representation domain, its projection is performed also there and as mentioned in the previous section, the projection is optimal in this case. For SSIHT that works in the signal domain no general projection procedure with an optimality guarantee is known.

The stopping criterion and the step size can be selected in the same way as in the regular IHT [19]. For the step size we consider three options: (i) Constant step-size selection $\mu^t=\mu$ in all iterations; (ii) Optimal changing step-size selection μ^t in each iteration by minimizing $\left\|\mathbf{y}-\mathbf{M}\hat{\mathbf{x}}^t\right\|_2$; and (iii) Adaptive changing step-size

[1] Our projection definition follows the one in [15], which is slightly different from the one used in [1].

[2] In this paper we shall use the brief notation δ_k to denote both RIP and $\mathbf{D}$-RIP, and the meaning should be understood from the context.

selection that has a closed-form solution and uses

$$\mu^t := \underset{\mu}{\operatorname{argmin}} \left\|\mathbf{y} - \mathbf{M}(\hat{\mathbf{x}}^{t-1} + \mu \mathbf{P}_{\tilde{T}}\mathbf{M}^*(\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^{t-1}))\right\|_2^2, \quad (10)$$

where $\tilde{T} = \hat{T}^{t-1} \cup \hat{\mathcal{S}}_k(M^*(\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^{t-1}))$. More details appears in [15], [19]. In our theoretical study we analyze the first two options. In the experimental part we use the third one as it works better than the first, and approximates the second that has no closed-form solution.

Algorithm 2 Signal space iterative hard thresholding (SSIHT)

Require: $k, \mathbf{M}, \mathbf{D}, \mathbf{y}$ where $\mathbf{y} = \mathbf{M}\mathbf{D}\boldsymbol{\alpha}_0 + \mathbf{e}$, k is the cardinality of $\boldsymbol{\alpha}_0$ and $\mathbf{e}$ is an additive noise.

Ensure: $\hat{\mathbf{x}}_{\text{SSIHT}}$: k-sparse approximation of $\mathbf{x}_0$.

Initialize estimate $\hat{\mathbf{x}}^0 = \mathbf{0}$ and set $t = 0$.

while halting criterion is not satisfied **do**

$t = t + 1$.

Perform a gradient step: $\mathbf{x}_g = \hat{\mathbf{x}}^{t-1} + \mu^t \mathbf{M}^*(\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^{t-1})$

Find a new support: $T^t = \hat{\mathcal{S}}_k(\mathbf{x}_g)$

Project to get a new estimate: $\hat{\mathbf{x}}^t = \mathbf{D}_{T^t}\mathbf{D}^\dagger_{T^t}\mathbf{x}_g$.

end while

Form the final solution $\hat{\mathbf{x}}_{\text{SSIHT}} = \hat{\mathbf{x}}^t$.

IV. Algorithms Guarantees

A uniform guarantee for the idealized version of SSIHT that has an access to the optimal projection and uses a constant step size $\mu^t = \mu$, is presented in [11]. The work in [11] deals with a general union of subspaces, $\mathcal{A}$, where in our case $\mathcal{A} = \{\mathbf{x} | \mathbf{x} = \mathbf{D}\boldsymbol{\alpha}, \|\boldsymbol{\alpha}\|_0 \leq k\}$. Using our notation Theorem 2 from [11] reads[3]:

Theorem 4.1 (Theorem 2 in [11]): Consider the problem $\mathcal{P}$ with $\hat{\mathcal{S}}_k = \mathcal{S}_k^*$ and apply SSIHT with a constant step size μ. If $1 + \delta_{2k} \leq \frac{1}{\mu} < 1.5(1 - \delta_{2k})$ then after a finite number of iterations t^*

$$\left\|\hat{\mathbf{x}}^{t^*} - \mathbf{x}_0\right\|_2 \leq c_3 \|\mathbf{e}\|_2, \quad (11)$$

where the constant c_3 is a function of δ_{2k} and μ.

In our work we extend the above in several ways: First, we refer to the case where an optimal projection is not known, and show that the same flavor guarantees apply for a near-optimal projection[4]. The price we seemingly have to pay is that $\sigma_{\mathbf{M}}$ enters the game. Second, we also consider the optimal step size and show that the same performance guarantees hold true in that case.

Theorem 4.2: Consider the problem $\mathcal{P}$ and apply SSIHT with a constant step size μ or an optimal changing step size. For any positive constant $\eta > 0$, let $b_1 := \frac{\eta}{1+\eta}$ and $b_2 := \frac{(C_k - 1)\sigma_{\mathbf{M}}^2 b_1^2}{C_k(1-\delta_{2k})}$. Suppose $\frac{b_2}{b_1^2} < 1$, $\frac{1}{\mu} \leq \sigma_{\mathbf{M}}^2$ and $1 + \delta_{2k} \leq \frac{1}{\mu} < \left(1 + \sqrt{1 - \frac{b_2}{b_1^2}}\right) b_1(1 - \delta_{2k})$. Then for $t \geq t^* \triangleq \frac{\log\left(\frac{\eta\|\mathbf{e}\|_2^2}{\|\mathbf{y}\|_2^2}\right)}{\log\left((1+\frac{1}{\eta})^2(\frac{1}{\mu(1-\delta_{2k})} - 1)C_k + (C_k - 1)(\mu\sigma_{\mathbf{M}}^2 - 1) + \frac{C_k}{\eta^2}\right)}$,[5]

$$\left\|\hat{\mathbf{x}}^t - \mathbf{x}_0\right\|_2^2 \leq \frac{(1+\eta)^2}{1 - \delta_{2k}} \|\mathbf{e}\|_2^2. \quad (12)$$

[3]Theorem 2 in [11] is more general and deals also with the case where $\hat{\mathcal{S}}_k$ is near-optimal up to an additive constant factor (in our definitions the factor is multiplicative). The error bound in the theorem has an additional constant factor that depends on the projection's near-optimality additive constant.

[4]Our work in fact improves the condition of the idealized case in [11] to be $\delta_{2k} \leq \frac{1}{3}$ instead of $\delta_{2k} \leq \frac{1}{5}$.

[5]For an optimal changing step-size the theorem conditions turn to be $\frac{b_2}{b_1^2} < 1$ and $1 + \delta_{2k} < (1 + \sqrt{1 - \frac{b_2}{b_1^2}}) b_1(1 - \delta_{2k})$ and we set $\mu = \frac{1}{1+\delta_{2k}}$ in t^*.

This theorem is a variant of Theorem 6.5 in [15] for AIHT and Theorem 2.1 in [20] for IHT. If, for example, $\sigma_{\mathbf{M}}^2 = 5$ and $C_k = 1.05$ then the conditions of Theorem 4.2 turn to be $\delta_{2k} \leq 0.289$ as mentioned before. For a better understanding of the nature of the theorem we refer the reader to the remarks after Theorems 6.2 and 6.5 in [15]. Briefly we comment on the selection of μ and η. For the step-size selection, note that an optimal changing step-size has the same theoretical guarantees as the optimal constant step-size $\mu = \frac{1}{1+\delta_{2k}}$. The advantage of the changing step-size method is that it does not need to compute (or estimate) the value of δ_{2k}. However, this comes at the cost of an additional complexity. Regarding the constant η, it gives a trade-off between satisfying the theorem conditions and the amplification of the noise. In particular, one may consider that the above theorem proves the convergence result for the noiseless case by taking η to infinity. This result is included in Lemma 4.4, which we present later, that guarantees in the case $\mathbf{e} = 0$ that $\mathbf{M}\hat{\mathbf{x}}^t$ converges geometrically to $\mathbf{M}\mathbf{x}_0$. Due to the uniqueness property that appears in [17], this implies that $\hat{\mathbf{x}}^t$ converges to $\mathbf{x}_0$.

We prove the theorem by presenting two key lemmas. The proofs rely on the ones in [15] that adopted ideas from [20] and [11]. Recall that the iterative algorithm tries to reduce the objective $\left\|\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^t\right\|_2^2$ over iterations t. Thus, the progress of the algorithm can be indirectly measured by how much the objective $\left\|\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^t\right\|_2^2$ is reduced at each iteration t. The two lemmas that we present capture this idea. The first lemma relates $\left\|\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^t\right\|_2^2$ to $\left\|\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^{t-1}\right\|_2^2$ and similar quantities at iteration $t-1$. We remark that the constraint $\frac{1}{\mu} \leq \sigma_{\mathbf{M}}^2$ in Theorem 4.2 may not be necessary and it is added only for having a simpler derivation of the results in this theorem. Furthermore, this is a very mild condition compared to $\frac{1}{\mu} < \left(1 + \sqrt{1 - \frac{b_2}{b_1^2}}\right) b_1(1 - \delta_{2k})$ and can only limit the range of values that can be used with the constant step size version of the algorithm.

Lemma 4.3: Consider the problem $\mathcal{P}$ and apply SSIHT with a constant step size μ satisfying $\frac{1}{\mu} \geq 1 + \delta_{2k}$ or an optimal step size[6]. Then, at the t-th iteration, the following holds:

$$\begin{aligned}\left\|\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^t\right\|_2^2 - \left\|\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^{t-1}\right\|_2^2 &\leq C_k \|\mathbf{y} - \mathbf{M}\mathbf{x}_0\|_2^2 \\ -C_k \left\|\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^{t-1}\right\|_2^2 &+ (C_k - 1)\mu\sigma_{\mathbf{M}}^2 \left\|\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^{t-1}\right\|_2^2 \\ +C_k \left(\frac{1}{\mu(1-\delta_{2k})} - 1\right) &\left\|\mathbf{M}(\hat{\mathbf{x}}^{t-1} - \mathbf{x}_0)\right\|_2^2. \end{aligned} \quad (13)$$

The proof of the above lemma is exactly the same as the proof of Lemma 6.6 in [15] with the change that here we use the $\mathbf{D}$-RIP instead of the $\boldsymbol{\Omega}$-RIP and the near-optimal projection scheme for synthesis instead of the one for analysis. The second lemma shows that once the objective $\left\|\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^{t-1}\right\|_2^2$ at iteration $t-1$ is small enough, then we are guaranteed to have small $\left\|\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^t\right\|_2^2$ as well. Given the presence of noise, this is quite natural; one cannot expect it to approach 0 but may expect it not to become worse. Moreover, the lemma also shows that if $\left\|\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^{t-1}\right\|_2^2$ is not small, then the objective in iteration t is necessarily reduced by a constant factor.

Lemma 4.4: Suppose that the same conditions of Theorem 4.2 holds true. If $\left\|\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^{t-1}\right\|_2^2 \leq \eta^2 \|\mathbf{e}\|_2^2$, then $\left\|\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^t\right\|_2^2 \leq \eta^2 \|\mathbf{e}\|_2^2$. Furthermore, if $\left\|\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^{t-1}\right\|_2^2 > \eta^2 \|\mathbf{e}\|_2^2$, then

$$\left\|\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^t\right\|_2^2 \leq c_4 \left\|\mathbf{y} - \mathbf{M}\mathbf{x}^{t-1}\right\|_2^2, \quad (14)$$

where $c_4 < 1$ and

$$c_4 := \left(1 + \frac{1}{\eta}\right)^2 \left(\frac{1}{\mu(1-\delta_{2k})} - 1\right) C_k + (C_k - 1)(\mu\sigma_{\mathbf{M}}^2 - 1) + \frac{C_k}{\eta^2}.$$

[6]For an optimal step size the bound is achieved with the value $\mu = \frac{1}{1+\delta_{2k}}$.

The Lemma's proof is similar to the one of Lemma 6.7 in [15]. The needed adaptations are similar to those done for Lemma 4.3. Having the two lemmas above, the proof of the theorem is straightforward.

Proof of Theorem 4.2: Since $\hat{\mathbf{x}}^0 = \mathbf{0}$, $\|\mathbf{y}\|_2^2 = \left\|\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^0\right\|_2^2$. Assuming that $\|\mathbf{y}\|_2 > \eta \|\mathbf{e}\|_2$ and applying Lemma 4.4 repeatedly, we obtain $\left\|\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^t\right\|_2^2 \leq \max(c_4^t \|\mathbf{y}\|_2^2, \eta^2 \|\mathbf{e}\|_2^2)$. Since $c_4^t \|\mathbf{y}\|_2^2 \leq \eta^2 \|\mathbf{e}\|_2^2$ for $t \geq t^*$, we have

$$\left\|\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^t\right\|_2^2 \leq \eta^2 \|\mathbf{e}\|_2^2 \tag{15}$$

for $t \geq t^*$. If $\left\|\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^0\right\|_2 = \|\mathbf{y}\|_2 \leq \eta \|\mathbf{e}\|_2$ then according to Lemma 4.4, (15) holds for every $t > 0$. Finally, we observe

$$\left\|\hat{\mathbf{x}}^t - \mathbf{x}_0\right\|_2^2 \leq \frac{1}{1-\delta_{2k}} \left\|\mathbf{M}(\hat{\mathbf{x}}^t - \mathbf{x}_0)\right\|_2^2 \tag{16}$$

and by the triangle inequality,

$$\left\|\mathbf{M}(\hat{\mathbf{x}}^t - \mathbf{x}_0)\right\|_2 \leq \left\|\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^t\right\|_2 + \|\mathbf{e}\|_2 . \tag{17}$$

By plugging (15) into (17) and then the resulting inequality into (16), the claim of the Theorem follows. □

V. Numerical Performance

We turn to check numerically whether SSIHT can recover signals in scenarios where IHT cannot. We perform a synthetic test similar to the one in [17] for signals that are sparse under a dictionary which is highly coherent and with linear dependencies between its columns. We generate a dictionary $\mathbf{D} = [\mathbf{D}_1, \mathbf{D}_2]$ where $\mathbf{D}_1, \mathbf{D}_2 \in \mathbb{R}^{d \times d}$, $d = 200$, $\mathbf{D}_1$ contains sparse columns with 2 non-zero entries which are 1 or -1 with probability 0.5 and $\mathbf{D_2}$ contains columns which are linear combinations of random 3 columns from $\mathbf{D}_1$ with random zero-mean white Gaussian weights. Each entry of the measurement matrix $\mathbf{M} \in \mathbb{R}^{m \times d}$ is distributed according to a normal Gaussian distribution, where m=$\lfloor \gamma d \rfloor$ and γ is the sampling rate – a value in the range $(0, 1]$. We set k to be $\lfloor \rho m \rfloor$ ($\rho \ll 1$) and measure the recovery rate of the representation $\boldsymbol{\alpha}$ and the signal $\mathbf{x}$ for various values of $\gamma \in \{0.1, 0.2, \ldots, 0.9\}$ and $\rho \in \{0.01, 0.02, \ldots, 0.05\}$. We compare SSIHT also to SSCoSaMP, where both use projection by thresholding. The adaptive changing step-size selection rule is used for IHT and SSIHT. Similar to what is done in [15], by uniqueness conditions it is better to apply the algorithms with sparsity $\tilde{k} = \max(k, m/2)$.

Figure 1 presents the recovery performance over 100 realizations per each parameter setting. As expected, IHT fails almost always in recovering the signal since it focuses on the representation, while SSIHT and SSCoSaMP succeed in several cases and their performance are similar. At a first glance, some would think that the SSIHT phase diagram implies that for a fixed k/m (e.g. 0.03) one may improve the recovery result if he uses less samples, i.e. smaller m/d. However, this observation misses the fact that for a fixed k/m, k is reduced together with m. Note that the recovery results of SSIHT and SSCoSaP can be improved by using other techniques for the projection, rather than thresholding, as done in [1] for SSCoSaMP.

VI. Conclusion

In this paper we have proposed a variant of the IHT algorithm – the Signal-Space IHT (SSIHT) – for recovering signals with sparse representations under highly coherent dictionaries. We have shown that IHT fails in recovering such signals, as it operates in the representation domain. SSIHT, on the other hand, targets the signal. A uniform recovery guarantee has been derived for the SSIHT, assuming the availability of a near optimal projection. Numerical simulations show that SSIHT succeeds in recovering signals for which IHT fails, even when the projection is not near-optimal.

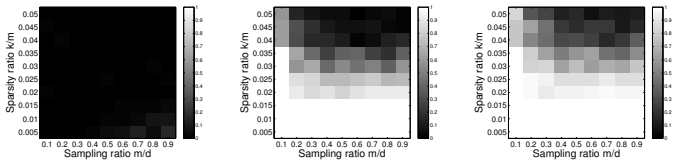


Fig. 1. IHT (left), SSIHT (middle) and SSCoSaMP (right) recovery rates for the synthetic experiment described in Section V. Color attribute: fraction of realizations in which a perfect recovery is achieved.

Acknowledgment

R. Giryes thanks the Azrieli Foundation for the Azrieli Fellowship. This research was supported by European Community's FP7- ERC program, grant agreement no. 320649.

References

[1] M. A. Davenport, D. Needell, and M. B. Wakin, "Signal space CoSaMP for sparse recovery with redundant dictionaries," *CoRR*, vol. abs/1208.0353, 2012.

[2] A. M. Bruckstein, D. L. Donoho, and M. Elad, "From sparse solutions of systems of equations to sparse modeling of signals and images," *SIAM Review*, vol. 51, no. 1, pp. 34–81, 2009.

[3] T. Blumensath and M. Davies, "Iterative hard thresholding for compressed sensing," *Appl. Comput. Harmon. Anal*, vol. 27, no. 3, pp. 265 – 274, 2009.

[4] E. J. Candès and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?" *IEEE Trans. Inf. Theory.*, vol. 52, no. 12, pp. 5406 –5425, Dec. 2006.

[5] S. Foucart, "Sparse recovery algorithms: sufficient conditions in terms of restricted isometry constants," in *Approximation Theory XIII*. Springer Proceedings in Mathematics, 2010, pp. 65–77.

[6] ——, "Hard thresholding pursuit: an algorithm for compressive sensing," *SIAM J. Numer. Anal.*, vol. 49, no. 6, pp. 2543–2563, 2011.

[7] R. Giryes and M. Elad, "RIP-based near-oracle performance guarantees for SP, CoSaMP, and IHT," *IEEE Trans. Signal Process.*, vol. 60, no. 3, pp. 1465–1468, March 2012.

[8] D. L. Donoho and M. Elad, "Optimal sparse representation in general (nonorthogoinal) dictionaries via l1 minimization," *Proceedings of the National Academy of Science*, vol. 100, pp. 2197–2202, Mar 2003.

[9] Y. Lu and M. Do, "A theory for sampling signals from a union of subspaces," *IEEE Trans. Signal Process.*, vol. 56, no. 6, pp. 2334 –2345, Jun. 2008.

[10] T. Blumensath and M. Davies, "Sampling theorems for signals from the union of finite-dimensional linear subspaces," *IEEE Trans. Inf. Theory.*, vol. 55, no. 4, pp. 1872 –1882, april 2009.

[11] T. Blumensath, "Sampling and reconstructing signals from a union of linear subspaces," *IEEE Trans. Inf. Theory.*, vol. 57, no. 7, pp. 4660–4671, 2011.

[12] E. J. Candès, Y. C. Eldar, D. Needell, and P. Randall, "Compressed sensing with coherent and redundant dictionaries," *Appl. Comput. Harmon. Anal*, vol. 31, no. 1, pp. 59 – 73, 2011.

[13] S. Nam, M. Davies, M. Elad, and R. Gribonval, "The cosparse analysis model and algorithms," *Appl. Comput. Harmon. Anal.*, vol. 34, no. 1, pp. 30 – 56, 2013.

[14] T. Peleg and M. Elad, "Performance guarantees of the thresholding algorithm for the cosparse analysis model," *IEEE Trans. on Information Theory*, vol. 59, no. 3, pp. 1832–1845, Mar. 2013.

[15] R. Giryes, S. Nam, M. Elad, R. Gribonval, and M. E. Davies, "Greedy-like algorithms for the cosparse analysis model," *to appear in the Special Issue in Linear Algebra and its Applications on Sparse Approximate Solution of Linear Systems*, 2013.

[16] R. Rubinstein, T. Peleg, and M. Elad, "Analysis K-SVD: A dictionary-learning algorithm for the analysis sparse model," *IEEE Trans. on Signal Processing*, vol. 61, no. 3, pp. 661–677, Feb. 2013.

[17] R. Giryes and M. Elad, "Can we allow linear dependencies in the dictionary in the synthesis framework?" in *ICASSP 2013.*

[18] R. Gribonval, M. E. Pfetsch, and A. M. Tillmann, "Projection onto the k-cosparse set is NP-hard," *submitted to IEEE Trans. Inf. Theory.*, 2013.

[19] A. Kyrillidis and V. Cevher, "Recipes on hard thresholding methods," in *CAMSAP, 2011*, Dec. 2011, pp. 353 –356.

[20] R. Garg and R. Khandekar, "Gradient descent with sparsification: an iterative algorithm for sparse recovery with restricted isometry property," in *ICML '09*. New York, NY, USA: ACM, 2009, pp. 337–344.

The Design of Non-redundant Directional Wavelet Filter Bank Using 1-D Neville Filters

Youngmi Hur, Fang Zheng
Department of Applied Mathematics and Statistics
Johns Hopkins University, Baltimore, Maryland 21218
Email: hur@jhu.edu, fzheng2@jhu.edu

***Abstract*—In this paper, we develop a method to construct non-redundant directional wavelet filter banks. Our method uses a special class of filters called Neville filters and can construct non-redundant wavelet filter banks in any dimension for any dilation matrix. The resulting filter banks have directional analysis highpass filters, thus can be used in extracting directional contents in multi-D signals such as images. Furthermore, one can custom-design the directions of highpass filters in the filter banks.**

I. Introduction

In the last couple of decades, wavelets have been a popular and useful tool in many applications such as signal and image processing. One of important remaining challenges in wavelets is to construct multi-D directional wavelet systems or wavelet filter banks.

There has been a lot of attempts to develop such wavelet systems or their variants for 2-D or 3-D signals, such as curvelets, contourlets, shearlets, etc. Despite many benefits of these existing systems, most of them are redundant with possibly huge redundancy factors, and they do not have a trivial generalization to higher dimensions. Although a recent study by the authors provides the construction of non-redundant wavelet filter banks with directional highpass filters for any dimension [1], it only deals with the dyadic dilation matrices. Other approaches based on anisotropic wavelet bases have also been proposed (see, for example, [2], [3], [4] and the references therein). However, these wavelets are designed in continuous domain and implementing them in discrete setting is not trivial.

In this paper, we develop a new method to construct non-redundant wavelet filter banks that can capture the directional information in multi-D signals. Our method is a general designing recipe in the sense that it can work in any dimension for any dilation matrix. In the design, one can even specify the number of directions and which directions to consider.

II. Preliminaries

In this section, we review some basic concepts and notations about wavelet filter bank construction. In particular, we review the concept of Neville filters and how to use Neville filters to build multi-D wavelet filter banks.

This work was supported in part by the National Science Foundation under Grant DMS-1115870.

A. Notation

In this paper, we use boldface to indicate vectors and matrices. A filter f is a a linear time-invariant operator characterized by its impulse response $\{f(\mathbf{k}) \in \mathbb{R} | \mathbf{k} \in \mathbb{Z}^d\}$. The z-transform of a filter is a Laurent polynomial

$$F(\mathbf{z}) = \textstyle\sum_{\mathbf{k}} f(\mathbf{k}) \mathbf{z}^{-\mathbf{k}}$$

where $\mathbf{z} = (z_1, z_2, \ldots, z_d)$ and $\mathbf{z}^{\mathbf{k}} := \prod_{i=1}^{d} z_i^{k_i}$. In this paper, we refer to both the z-transform $F(\mathbf{z})$ and the impulse response $f(\mathbf{k})$ as the filter, and sometimes we omit $\mathbf{z}$ and $\mathbf{k}$ in the parentheses for convenience. Define the *adjoint* of a filter as $[F(\mathbf{z})]^* := F(1/\mathbf{z})$. Throughout this paper, we assume all filters have finite impulse response.

A dilation matrix $\mathbf{D}$ is a $d \times d$ integer matrix with $|\det \mathbf{D}| := m > 1$. Given a dilation matrix $\mathbf{D}$, the set $\mathbb{Z}^d$ of integer grids can be split into m disjoint subsets

$$\mathbb{Z}^d = \textstyle\bigcup_{i=0}^{m-1} (\mathbf{D}\mathbb{Z}^d + \mathbf{t}_i), \quad \mathbf{t}_i \in \mathbb{Z}^d$$

where $\mathbf{t}_0 = \mathbf{0}$. We call $\{\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_{m-1}\}$ as a set of *(nonzero) distinct coset representatives* of the dilation matrix $\mathbf{D}$.

A filter bank (FB) consisting of an analysis bank and a synthesis bank is a set of filters. For a given dilation matrix $\mathbf{D}$, a filter in the analysis bank $\{A_i, i = 0, \ldots, l-1\}$ and a filter in the synthesis bank $\{S_i, i = 0, \ldots, l-1\}$ can be written as the sum of m *polyphase components*

$$A_i(\mathbf{z}) = \textstyle\sum_{j=0}^{m-1} \mathbf{z}^{\mathbf{t}_j} A_{i,j}(\mathbf{z}^{\mathbf{D}}), \quad a_{i,j}(\mathbf{k}) := a_i(\mathbf{D}\mathbf{k} - \mathbf{t}_j) \quad (1)$$

$$S_i(\mathbf{z}) = \textstyle\sum_{j=0}^{m-1} \mathbf{z}^{-\mathbf{t}_j} S_{i,j}(\mathbf{z}^{\mathbf{D}}), \quad s_{i,j}(\mathbf{k}) := s_i(\mathbf{D}\mathbf{k} + \mathbf{t}_j) \quad (2)$$

where $\mathbf{z}^{\mathbf{D}} := (\mathbf{z}^{D_1}, \mathbf{z}^{D_2}, \ldots, \mathbf{z}^{D_d})$, D_i is the ith column vector of $\mathbf{D}$. Then the pair of matrices

$$\begin{aligned} \mathbf{A}(\mathbf{z}) &:= [A_{i,j}(\mathbf{z})]_{i=0,\ldots,l-1; j=0,\ldots,m-1} \\ \mathbf{S}(\mathbf{z}) &:= [S_{j,i}(\mathbf{z})]_{j=0,\ldots,m-1; i=0,\ldots,l-1} \end{aligned}$$

is called the *polyphase matrix representation* [5] of the FB.

A FB satisfies the *perfect reconstruction* condition if the polyphase matrices satisfy $\mathbf{S}(\mathbf{z})\mathbf{A}(\mathbf{z}) = \mathbf{I}_m$, which can happen only when $l \geq m$. A FB is called *non-redundant* if $l = m$.

In this paper, we are only interested in non-redundant FBs satisfying the perfect reconstruction condition, and we assume there are exactly one lowpass filter A_0 in the analysis bank and one lowpass filter S_0 in the synthesis bank. The rest, $A_1, \ldots, A_{m-1}, S_1, \ldots, S_{m-1}$, are all highpass filters.

We use Π_N to denote the set of all polynomials of total degree less than N. We say a FB has $N \in \mathbb{N}$ *vanishing moments* [6] if, for any highpass filter f in the FB, $(f *' \pi)(\mathbb{Z}^d) = 0, \forall \pi \in \Pi_N$, or equivalently,

$$\textstyle\sum_{\mathbf{k}} f(-\mathbf{k})\mathbf{k}^{\mathbf{n}} = 0, \ \forall \mathbf{n} \in \mathbb{N}_0^d, \ |\mathbf{n}| < N$$

where $\mathbf{n} := (n_1, n_2, \ldots, n_d)$, $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$ and $|\mathbf{n}| := n_1 + n_2 + \ldots + n_d$. Here we used $(f *' \pi)(\cdot) := \sum_{\mathbf{k} \in \mathbb{Z}^d} f(\mathbf{k})\pi(\cdot - \mathbf{k})$.

B. Neville Filters and Their Use in Wavelet FB Construction

In [7], Kovačević and Sweldens introduce a class of filters called *Neville filters* (Definition 1) and their characterization (Result 1). When applied to a sampled polynomial, they result in the same polynomial but shifted by a shift parameter $\boldsymbol{\tau} \in \mathbb{R}^d$.

Definition 1. A filter f is a Neville filter of order N with shift $\boldsymbol{\tau}$ if $(f *' \pi)(\mathbb{Z}^d) = \pi(\mathbb{Z}^d + \boldsymbol{\tau})$, for any $\pi \in \Pi_N$. ■

Result 1 (Proposition 4 in [7]). A filter f is a Neville filter of order N with shift $\boldsymbol{\tau}$ if and only if f satisfies

$$\textstyle\sum_{\mathbf{k}} f(-\mathbf{k})\mathbf{k}^{\mathbf{n}} = \boldsymbol{\tau}^{\mathbf{n}}, \ \forall \mathbf{n} \in \mathbb{N}_0^d, \ |\mathbf{n}| < N. \tag{3}$$

■

In 1-D case, the construction of Neville filters of order N is straightforward. Once we fix the positions of N filter taps, we obtain a linear system with an $N \times N$ coefficient matrix from (3). Since the coefficient matrix in this case is a Vandermonde matrix, it is always solvable. In multi-D case, the solvability of the linear system not only depends on the number of filter taps but also on the geometric shape of the filter. Hence it is more challenging to construct a multi-D Neville filter with a prescribed order and shift. An approach based on an algorithm in [8] to solve this problem is proposed in [7], but it is highly non-trivial to control the shape of the filters using that approach.

Using the property of Neville filters, Kovačević and Sweldens propose a method for constructing wavelet FBs based on lifting scheme [9]. They use two lifting steps: predict (cf. R_i) and update (cf. U_i), as shown in (4) and (5) to build the wavelet FB with desirable vanishing moments:

$$\mathbf{A} = \begin{bmatrix} 1 & U_1 & \cdots & U_{m-1} \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & \cdots & 0 \\ -R_1 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -R_{m-1} & 0 & \cdots & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 1 - \sum_{i=1}^{m-1} U_i R_i & U_1 & \cdots & U_{m-1} \\ -R_1 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -R_{m-1} & 0 & \cdots & 1 \end{bmatrix} \tag{4}$$

$$\mathbf{S} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ R_1 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ R_{m-1} & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} 1 & -U_1 & \cdots & -U_{m-1} \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & -U_1 & \cdots & -U_{m-1} \\ R_1 & 1 - R_1 U_1 & \cdots & -R_1 U_{m-1} \\ \vdots & \vdots & \ddots & \vdots \\ R_{m-1} & -R_{m-1} U_1 & \cdots & 1 - R_{m-1} U_{m-1} \end{bmatrix}, \tag{5}$$

where R_i are called predict filters, U_i are called update filters, and $m = |\det \mathbf{D}|$. More precisely, the following is a variant of the result they prove in [7], written in terms of our terminology.

Result 2. Let $\{\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_{m-1}\}$ be a set of distinct coset representatives of the $d \times d$ dilation matrix $\mathbf{D}$. For $i = 1, \cdots, m-1$, let R_i be a d-D Neville filter of order N with shift $\boldsymbol{\tau}_i = \mathbf{D}^{-1}\mathbf{t}_i$, and U_i be the filter obtained by multiplying $1/m$ to the adjoint of a d-D Neville filter of order N with shifts $\boldsymbol{\tau}_i$. Then the analysis polyphase matrix constructed as (4) and the synthesis polyphase matrix constructed as (5) form a wavelet FB with N vanishing moments. ■

This construction works for any dilation matrix $\mathbf{D}$ in any dimension. It uses d-D Neville filters with prescribed orders and shifts to construct d-D wavelet FBs.

III. Directional wavelet FB design using 1-D Neville filters

In this section, we introduce a method to design directional wavelet FBs using 1-D Neville filters and the lifting based wavelet construction method reviewed in Section II-B. Let us first define an operator that maps 1-D filters to d-D filters.

Definition 2. Define the operator that maps a 1-D filter F to a d-D filter $\mathcal{M}_{\mathbf{t}}(F)$ along direction $\mathbf{t} \in \mathbb{Z}^d$ as

$$\mathcal{M}_{\mathbf{t}}(F)(\mathbf{z}) := F(\mathbf{z}^{\mathbf{t}}).$$

■

The following simple lemma, which says that the operator $\mathcal{M}_{\mathbf{t}}$ preserves the order of Neville filters is a key ingredient of our directional wavelet FB construction.

Lemma 1. *If F is a 1-D Neville filter of order N with shift $\tau \in \mathbb{R}$, then the d-D filter $\mathcal{M}_{\mathbf{t}}(F)$ is a Neville filter of order N with shift $\tau\mathbf{t}, \mathbf{t} \in \mathbb{Z}^d$.*

Proof: Let $G := \mathcal{M}_{\mathbf{t}}(F)$, and let g be the impulse response of G. Then, we have

$$g(\mathbf{k}) = \begin{cases} f(k), & \text{if } \mathbf{k} = k\mathbf{t} \text{ for some } k \in \mathbb{Z}, \\ 0, & \text{for all other } \mathbf{k} \in \mathbb{Z}^d. \end{cases}$$

where f is the impulse response of F. Therefore

$$\begin{aligned} \textstyle\sum_{\mathbf{k}} g(-\mathbf{k})\mathbf{k}^{\mathbf{n}} &= \textstyle\sum_k f(-k)(k\mathbf{t})^{\mathbf{n}} = \sum_k f(-k)k^{|\mathbf{n}|}\mathbf{t}^{\mathbf{n}} \\ &= \tau^{|\mathbf{n}|}\mathbf{t}^{\mathbf{n}} = (\tau\mathbf{t})^{\mathbf{n}}, \end{aligned}$$

for any $\mathbf{n} \in \mathbb{N}_0^d$, $|\mathbf{n}| < N$, where the second last equation holds because F is a 1-D Neville filter of order N with shift τ. Thus G is a d-D Neville filter of order N with shift $\tau\mathbf{t}$. ■

Example 1: Mapping 1-D Neville Filter to 2-D. $F(z) = 1/3z + 2/3$ is a 1-D Neville filter of order 2 with shift $\tau = 1/3$. Then mapping it to 2-D along direction $\mathbf{t} = (1, 1)$ results in $\mathcal{M}_{\mathbf{t}}(F)(\mathbf{z}) = 1/3 z_1 z_2 + 2/3$. It can be easily checked that $\mathcal{M}_{\mathbf{t}}(F)$ is a Neville filter of order 2 with shift $\tau\mathbf{t} = (1/3, 1/3)$. Figure 1 shows the impulse response of F and $\mathcal{M}_{\mathbf{t}}(F)$. ■

From Example 1, we see that the multi-D Neville filter constructed by the operator $\mathcal{M}_{\mathbf{t}}$ is directional along direction $\mathbf{t}$. We now discuss how to use these directional multi-D Neville filters to construct directional wavelet FB.

$$\begin{array}{cc} \frac{1}{3} & \underline{\frac{2}{3}} \end{array} \quad \longrightarrow \quad \begin{array}{cc} 0 & \underline{\frac{2}{3}} \\ \frac{1}{3} & 0 \end{array}$$

Fig. 1. Mapping 1-D Neville filter to 2-D. The impulse response of F and $\mathcal{M}_{\mathbf{t}}(F)$ in Example 1. Underlined position is the origin.

Let us first look at a simple case when the dilation matrix $\mathbf{D} = c\mathbf{I}_d$ where $c \in \mathbb{Z},\ c > 1$ and $\mathbf{I}_d$ is the identity matrix. In this case, $\mathbf{D}^{-1} = (1/c)\mathbf{I}_d$. The multi-D Neville filters used to construct predict and update filters in Result 2 need to have shift parameters $\boldsymbol{\tau}_i = \mathbf{D}^{-1}\mathbf{t}_i = (1/c)\mathbf{t}_i$. Therefore, it is possible to construct all these multi-D Neville filters by mapping a single 1-D Neville filter with shift $\tau = 1/c$ but with different directions $\mathbf{t}_i$. In this way, we can avoid constructing multi-D Neville filters directly, which is often difficult to do. Moreover, it can be shown that the highpass filters built on these multi-D Neville filters are also directional.

To generalize this idea to a general dilation matrix $\mathbf{D}$, let us consider the shift parameters $\boldsymbol{\tau}_i = \mathbf{D}^{-1}\mathbf{t}_i$ again. In this case, if we factor out $\tau = 1/m$ as the shift parameter for 1-D Neville filters, then $\boldsymbol{\tau}_i = \tau\tilde{\mathbf{t}}_i$, where $\tilde{\mathbf{t}}_i = m\mathbf{D}^{-1}\mathbf{t}_i \in \mathbb{Z}^d$, hence we can map a single 1-D Neville filter with shift $\tau = 1/m$ along different directions $\tilde{\mathbf{t}}_i$. For example, for dilation matrix

$$\mathbf{D} = \begin{bmatrix} 2 & -1 \\ 1 & 2 \end{bmatrix} \tag{6}$$

a set of distinct coset representatives of $\mathbf{D}$ are $\mathbf{t}_1 = (0,1), \mathbf{t}_2 = (1,1), \mathbf{t}_3 = (0,2), \mathbf{t}_4 = (1,2)$. The shift parameters of Neville filters needed to construct wavelet FB are $\boldsymbol{\tau}_1 = (1/5, 2/5), \boldsymbol{\tau}_2 = (3/5, 1/5), \boldsymbol{\tau}_3 = (2/5, 4/5), \boldsymbol{\tau}_4 = (4/5, 3/5)$. Therefore, we can construct all these multi-D Neville filters by mapping one 1-D Neville filter with shift $1/5$ along directions $\tilde{\mathbf{t}}_1 = (1,2), \tilde{\mathbf{t}}_2 = (3,1), \tilde{\mathbf{t}}_3 = (2,4), \tilde{\mathbf{t}}_4 = (4,3)$.

In fact, we can factor out any $\tau = 1/s$, where $s \in \mathbb{Z}$, as the shift parameter for 1-D Neville filters, as long as $\boldsymbol{\tau}_i = \tau\tilde{\mathbf{t}}_i$ and $\tilde{\mathbf{t}}_i = s\mathbf{D}^{-1}\mathbf{t}_i \in \mathbb{Z}^d$. In the simple case when $\mathbf{D} = c\mathbf{I}_d$, $s := c$ can be chosen, while in other cases such as (6), $s := m$ can be chosen. Therefore, we have the following theorem. For a general d-D dilation matrix $\mathbf{D}$ with $|\det \mathbf{D}| = m$, we can construct a directional wavelet FB with analysis highpass filters presenting at most $m-1$ different directions as follows.

Theorem 1. *Let $\{\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_{m-1}\}$ be a set of distinct coset representatives of $\mathbf{D}$. Let s be an integer such that $s\mathbf{D}^{-1}\mathbf{t}_i \in \mathbb{Z}^d$. For $i = 1, \cdots, m-1$, let P_i and Q_i be the 1-D Neville filters of order N with shift $1/s$. Set $\tilde{\mathbf{t}}_i = s\mathbf{D}^{-1}\mathbf{t}_i$. Let d-D filter $R_i := \mathcal{M}_{\tilde{\mathbf{t}}_i}(P_i)$ and $U_i := (1/m)[\mathcal{M}_{\tilde{\mathbf{t}}_i}(Q_i)]^*$. Then the analysis polyphase matrix given by (4) and the synthesis polyphase matrix given by (5) form a directional FB with N vanishing moments and the analysis highpass filters are placed along directions $\mathbf{t}_i$.* ■

Proof: Since P_i (resp. Q_i) is a 1-D Neville filter of order N with shift $1/s$, by Lemma 1, $R_i = \mathcal{M}_{\tilde{\mathbf{t}}_i}(P_i)$ (resp. $\mathcal{M}_{\tilde{\mathbf{t}}_i}(Q_i)$) is a d-D Neville filter of order N with shift $(1/s)\tilde{\mathbf{t}}_i = (1/s)s\mathbf{D}^{-1}\mathbf{t}_i = \mathbf{D}^{-1}\mathbf{t}_i$. Thus $U_i = (1/m)[\mathcal{M}_{\tilde{\mathbf{t}}_i}(Q_i)]^*$ is $1/m$ times the adjoint of Neville filter of order N with shift $\mathbf{D}^{-1}\mathbf{t}_i$. By Result 2, we see that (4) and (5) form a wavelet FB with N vanishing moments.

To prove the directionality of analysis highpass filters, consider the ith analysis highpass filter denoted by A_i. Since

$$R_i(\mathbf{z}) = \mathcal{M}_{\tilde{\mathbf{t}}_i}(P_i)(\mathbf{z}) = P(\mathbf{z}^{\tilde{\mathbf{t}}_i}) = P(\mathbf{z}^{s\mathbf{D}^{-1}\mathbf{t}_i}),$$

from (1) and (4), we see that $A_i(\mathbf{z})$ is equal to

$$-R_i(\mathbf{z}^{\mathbf{D}}) + \mathbf{z}^{\mathbf{t}_i} = -P_i(\mathbf{z}^{\mathbf{D}s\mathbf{D}^{-1}\mathbf{t}_i}) + \mathbf{z}^{\mathbf{t}_i} = -P_i(\mathbf{z}^{s\mathbf{t}_i}) + \mathbf{z}^{\mathbf{t}_i}.$$

If we replace $\mathbf{z}^{\mathbf{t}_i}$ with z in the last equation on the right hand side, we get a 1-D filter $-P_i(z^s) + z$. Thus A_i can be understood as the result of taking the 1-D filter $-P_i(z^s) + z$ and placing it in d-D space along direction $\mathbf{t}_i$. ■

Remark 1. In Theorem 1, a single 1-D Neville filter of order N and shift $1/m$ can be used for all of P_i and Q_i, or different 1-D Neville filters can be used. In fact P_i and Q_i can have different orders if we invoke more generalized version of Result 2 from [7]. In this case, if P_i's order is $\tilde{N}_i$ and Q_i's order is N_i, then the vanishing moments of the FB is given as $\min\{\tilde{N}_1, \ldots, \tilde{N}_{m-1}, N_1, \ldots, N_{m-1}\}$.

Remark 2. The analysis highpass filters A_i of the FB in Theorem 1 are placed along directions $\mathbf{t}_i \in \mathbb{Z}^d, i = 1, \ldots, m-1$ (not $\tilde{\mathbf{t}}_i = m\mathbf{D}^{-1}\mathbf{t}_i$). Therefore, by carefully choosing the distinct coset representatives of $\mathbf{D}$, one can custom-design the directions of the filters (cf. Example 2). There are *at most* $m-1$ different directions that can be presented by the analysis highpass filters.

In the next example, we illustrate how to use Theorem 1 to construct directional wavelet FB.

Example 2: 2-D Directional Wavelet FB with 2 Vanishing Moments. For dilation matrix $\mathbf{D} = 3\mathbf{I}_2$, since $|\det \mathbf{D}| = 9$, there are $9 - 1 = 8$ distinct coset representatives $\{\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_8\}$ that we can choose. We know that the directions of coset representatives are exactly the directions of resulting analysis highpass filters. Here we want to choose directions that divide the 2-D plane as equally as possible. Thus we choose $\mathbf{t}_1 = (1,0), \mathbf{t}_2 = (-1,0), \mathbf{t}_3 = (0,1), \mathbf{t}_4 = (0,-1), \mathbf{t}_5 = (2,1), \mathbf{t}_6 = (1,2), \mathbf{t}_7 = (-2,1), \mathbf{t}_8 = (-1,2)$. Then the resulting analysis highpass filters will present 6 different directions in the 2-D plane: approximately, 0° ($\mathbf{t}_1$, $\mathbf{t}_2$), 30° ($\mathbf{t}_5$), 60° ($\mathbf{t}_6$), 90° ($\mathbf{t}_3$, $\mathbf{t}_4$), 120° ($\mathbf{t}_8$) and 150° ($\mathbf{t}_7$) from the positive x-axis.

Next we pick a single 1-D Neville filter of order 2 with shift $1/3$ for all P_i and Q_i: $P_i(z) = Q_i(z) = 1/3z + 2/3$, for $i = 1, \ldots, 8$. Theorem 1 says that if we choose, for each i,

$$\begin{aligned} R_i(\mathbf{z}) &= P_i(\mathbf{z}^{\mathbf{t}_i}) = 1/3\mathbf{z}^{\mathbf{t}_i} + 2/3 \\ U_i(\mathbf{z}) &= (1/m)[Q_i(\mathbf{z}^{\mathbf{t}_i})]^* = (1/9)(1/3\mathbf{z}^{-\mathbf{t}_i} + 2/3) \end{aligned}$$

then we get the wavelet FB with 2 vanishing moments, whose polyphase matrices are $\mathbf{A}$ and $\mathbf{S}$ in (4) and (5). Using formula (1) and (2), we can read off the corresponding filters. For example, the resulting synthesis lowpass filter S_0 is

$$S_0(\mathbf{z}) = 1 + \textstyle\sum_{i=1}^{8} \mathbf{z}^{-\mathbf{t}_i} R_i(\mathbf{z}^{\mathbf{D}})$$

$$\begin{array}{ccccccccc}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & \frac{2}{3} & \frac{1}{3} & \frac{2}{3} & 0 & 0 & 0 \\
0 & 0 & \frac{2}{3} & 0 & \frac{2}{3} & 0 & \frac{2}{3} & 0 & 0 \\
0 & 0 & \frac{1}{3} & \frac{2}{3} & \underline{1} & \frac{2}{3} & \frac{1}{3} & 0 & 0 \\
0 & 0 & 0 & 0 & \frac{2}{3} & 0 & 0 & 0 & 0 \\
\frac{1}{3} & 0 & 0 & 0 & \frac{1}{3} & 0 & 0 & 0 & \frac{1}{3} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & \frac{1}{3} & 0 & 0 & 0 & \frac{1}{3} & 0 & 0
\end{array}$$

(a) Synthesis lowpass filter S_0

$$\begin{array}{cccc} -\frac{1}{3} & 0 & 1 & \underline{-\frac{2}{3}} \end{array}$$

(b) $A_1 : \mathbf{t}_1 = (1, 0)$

$$\begin{array}{cccc} \underline{-\frac{2}{3}} & 1 & 0 & -\frac{1}{3} \end{array}$$

(c) $A_2 : \mathbf{t}_2 = (-1, 0)$

$$\begin{array}{c} -\frac{2}{3} \\ 1 \\ 0 \\ -\frac{1}{3} \end{array}$$

(d) $A_3 : \mathbf{t}_3 = (0, 1)$

$$\begin{array}{c} -\frac{1}{3} \\ 0 \\ 1 \\ \underline{-\frac{2}{3}} \end{array}$$

(e) $A_4 : \mathbf{t}_4 = (0, -1)$

$$\begin{array}{ccccccc} & & & 0 & \underline{-\frac{2}{3}} & & \\ & & & 0 & & & \\ & & & 0 & & & \\ 0 & 0 & 0 & 1 & 0 & 0 & \underline{-\frac{2}{3}} \\ & & & 0 & & & \\ & & & 0 & & & \\ -\frac{1}{3} & & & 0 & & & \end{array}$$

(f) $A_5 : \mathbf{t}_5 = (2, 1)$

$$\begin{array}{ccccccc} & & & 0 & \underline{-\frac{2}{3}} & & \\ & & & 0 & & & \\ & & & 0 & & & \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ & & & 0 & & & \\ & & & 0 & & & \\ & & & 0 & & & -\frac{1}{3} \end{array}$$

(g) $A_6 : \mathbf{t}_6 = (1, 2)$

$$\begin{array}{ccccccc} \underline{-\frac{2}{3}} & & & 0 & & & \\ & & & 0 & & & \\ 0 & 1 & 0 & 0 & & & \\ & & & 0 & & & \\ & & & 0 & & & \\ & & & 0 & & & \\ & & & 0 & & & -\frac{1}{3} \end{array}$$

(h) $A_7 : \mathbf{t}_7 = (-2, 1)$

$$\begin{array}{ccccccc} & & & 0 & & & \\ & & & 0 & & & \\ \underline{-\frac{2}{3}} & 1 & 0 & 0 & & & \\ & & & 0 & & & \\ & & & 0 & & & \\ & & & 0 & & & \\ & & & 0 & & & -\frac{1}{3} \end{array}$$

(i) $A_8 : \mathbf{t}_8 = (-1, 2)$

Fig. 2. 2-D directional wavelet FB with 2 vanishing moments in Example 2: (a) synthesis lowpass filter, (b)-(i) directional analysis highpass filters with each direction along the coset representatives: $\mathbf{t}_i,\ i = 1, \ldots, 8$.

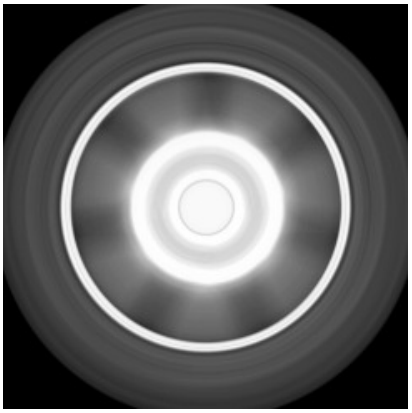
and the resulting analysis highpass filter associated with coset representative $\mathbf{t}_5 = (2, 1)$ is

$$A_5(\mathbf{z}) = -R_5(\mathbf{z}^{\mathbf{D}}) + \mathbf{z}^{\mathbf{t}_5} = -(1/3z_1^6 z_2^3 + 2/3) + z_1^2 z_2.$$

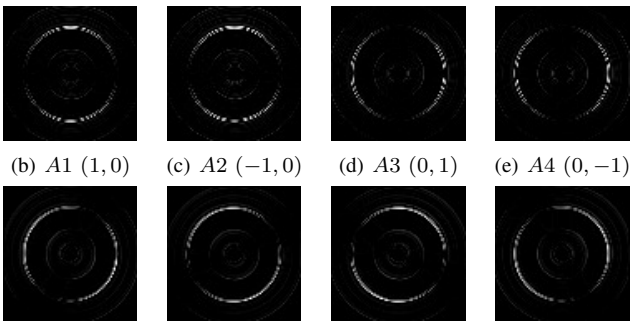
Figure 2 shows the synthesis lowpass filter S_0 and the analysis highpass filters $A_i,\ i = 1, \ldots, 8$. ■

IV. Experimental Result

We did an experiment using the 2-D directional wavelet FB constructed in Example 2. For an original image "circle" (Figure 3(a)), we did a 1-level-down decomposition using the analysis highpass filters obtained in Example 2 (as shown in Figure 2(b)-(i)). The images after passing through each highpass filter (wavelet coefficients) are shown in Figure 3(b)-(i). The result shows that different directional components of the circle are captured by different directional highpass filters. A highpass filter with direction $\mathbf{t}$ can mainly capture the directional content that is orthogonal to the direction $\mathbf{t}$.



(a) original



(b) $A1\ (1, 0)$ (c) $A2\ (-1, 0)$ (d) $A3\ (0, 1)$ (e) $A4\ (0, -1)$

(f) $A5\ (2, 1)$ (g) $A6\ (1, 2)$ (h) $A7\ (-2, 1)$ (i) $A8\ (-1, 2)$

Fig. 3. (a) The original image "circle", (b)-(i) the images after passing highpass filters $A1, \ldots, A8$.

V. Conclusion

In this paper, we developed a method to use 1-D Neville filters to build multi-D directional wavelet FBs. The resulting FB is a non-redundant FB which can capture the directional information in multi-D signals.

References

[1] Y. Hur and F. Zheng, "Coset Sum: an alternative to the tensor product in wavelet construction," *IEEE Trans. Inform. Theory*, 2013, to be published.

[2] R. A. DeVore, S. V. Konyagin, and V. N. Temlyakov, "Hyperbolic wavelet approximation," *Constructive Approximation*, vol. 14, no. 1, pp. 1–26, 1998.

[3] J.-P. Antoine, P. Vandergheynst, and R. Murenzi, "Two-dimensional directional wavelets in image processing," *International journal of imaging systems and technology*, vol. 7, no. 3, pp. 152–165, 1996.

[4] H. Triebel, "Wavelet bases in anisotropic function spaces," in *Proc. Function Spaces, Differential Operators and Nonlinear Analysis (FSDONA2004)*, Milovy, Czech Republic, 2005, pp. 370–387.

[5] G. Strang and T. Nguyen, *Wavelets and Filter Banks*. Wellesley: Wellesley-Cambridge Press, 1997.

[6] D.-R. Chen, B. Han, and S. D. Riemenschneider, "Construction of multivariate biorthogonal wavelets with arbitrary vanishing moments," *Adv. Comput. Math.*, vol. 13, no. 2, pp. 131–165, 2000.

[7] J. Kovačević and W. Sweldens, "Wavelet families of increasing order in arbitrary dimensions," *Image Processing, IEEE Transactions on*, vol. 9, no. 3, pp. 480–496, 2000.

[8] C. de Boor and A. Ron, "On multivariate polynomial interpolation," *Constructive Approximation*, vol. 6, no. 3, pp. 287–302, 1990.

[9] W. Sweldens, "The lifting scheme: A custom-design construction of biorthogonal wavelets," *Appl. Comput. Harmon. Anal.*, vol. 3(2), pp. 186–200, 1996.

Sparse Approximation of Ion-Mobility Spectrometry Profiles by Minutely Shifted Discrete B-splines

Masaru Kamada
Department of Computer and Information Sciences
Ibaraki University
Hitachi, Ibaraki 316–8511, Japan
m.kamada@cis.ibaraki.ac.jp

Masakazu Ohno
Graduate School of Science and Engineering
Ibaraki University
Hitachi, Ibaraki 316–8511, Japan

***Abstract*—Employing discrete B-splines instead of the Gaussian distribution, we construct an algorithm for the analysis of ion-mobility spectrometry profiles. The algorithm is suitable for hardware implementation because the discrete B-splines are supported by a simple digital filter to compute their weighted sum and their correlations with a given signal. Minutely shifted discrete B-splines are deployed of which weighted sum is to approximate a given profile with non-negative weights. Closely neighboring discrete B-splines are almost linearly dependent so that they may cause numerical instability in the approximation process. But numerical experiments deny this anxiety at least for the final results. Varying the width of discrete B-splines, we obtain a number of different approximations. Out of sufficiently precise approximations, we choose the sparse one in the sense that it comprises few discrete B-splines with large weights.**

I. Introduction

Ion-mobility spectrometry [1] is a method of discriminating chemical molecules in the atmosphere. Its capability of identifying tiny amounts of various chemicals has made it possible to analyze odor and flavor and to detect poisons, drugs and explosives. The analysis is mainly composed of physical and computational processes.

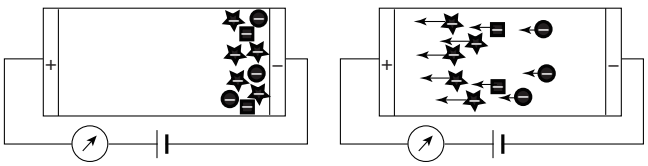
The physical process proceeds in this way: (i) Chemical molecules are ionized and injected near the cathode as shown in Fig. 1(a). (ii) The ions move toward the anode with the acceleration proportional to their charge-mass-ratio as illustrated in Fig. 1(b). Light ions reach the anode earlier than the heavier ones on the average. The ions bump and bounce against air and other molecules during their travel so that even ions of the same kind arrive at the anode in different traveling times. (iii) The ions give their charges to the anode which constitute the electric current called *profile* like the curve in Fig. 2(a). The profile is modeled as a weighted sum of several distributions as schematized in Fig. 2(b). Each distribution is traditionally supposed to be Gaussian because any random displacements of ions by their collision with other molecules amount to a Gaussian distribution if they happen infinitely many times.

The computational process identifies each different distribution in a given profile. Its weight and average tell, respectively, how much and what kind of ions are present. The standard algorithm employs the steepest descent method to search for locally optimal values of unknown parameters such as average, variance and weight of an unknown number of Gaussian distributions. This search has to be conducted sequentially so that it consumes much time even on the latest fast CPUs.

While a tiny chip from Owlstone Nanotech [2] and a system solution from ATONARP [3] have already made it possible to complete the physical process in a few milliseconds, the computational algorithm is still sequentially searching for local optima at much computational cost. In this paper, we shall approach a new algorithm which matches up to the compact and fast physical system. This approach is characterized by the following four features:

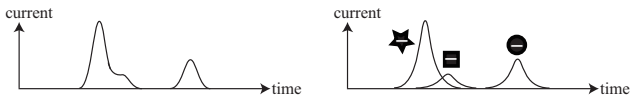
(i) Instead of the Gaussian distribution, we use the B-spline [4] of order m that is defined as the m-fold convolution integral of a uniform distribution and represents the distribution of ion position after m collisions if one causes a uniformly random displacement. The B-spline is a good substitute since it tends to the Gaussian at the limit $m \to \infty$. We can even say that the Gaussian was not the perfect choice because it has infinitely long tails that never exist in reality. We had better take a large m but do not have to make it infinity.

(ii) For the sake of simpler computation, the B-splines are further replaced by their discrete version[1] defined as the m-fold discrete convolution of the uniform discrete distribution over n sampling points. The discrete B-splines can be generated by only additions and subtractions [6]. There is also



(a) Initial state (b) Analyzing process

Fig. 1. Schematics of the ion-mobility spectrometry



(a) Profile (ion current) (b) Composite model

Fig. 2. Ion-mobility spectrometry profile

[1]The discrete B-splines tend to the original B-splines when $n \to \infty$ [5].

a fast digital filter to compute their correlations with a given signal [7].

(iii) We dare to deploy the discrete B-splines shifted by a minute interval as analyzing components of which weighted sum is to approximate a given profile even though we risk numerical instability in the approximation process due to the almost linear dependency among the overcrowding components. Otherwise, the algorithm would fall back to the slow sequential search for an unknown number of arrival times. The weights are constrained to be non-negative since ion counts cannot be negative numbers.

(iv) The other unknown parameter n, which represents how widely the arrival times distribute, is sought exhaustively. Since the above approximation process is rather simple and suitable for hardware implementation, we can try approximations with various values of n in parallel to find its best value. Among the values of n that result in good approximations with sufficiently small errors, we shall choose the one giving a sparse approximation in the sense that the approximation comprises few discrete B-splines with large weights.

The algorithm is a sort of sparse approximation method that arose from this particular application field. It works empirically fine. A proper formulation within the general theory of sparse approximation is yet to be established.

II. Summary of discrete B-splines

The B-spline of order m is defined as the m-fold convolution integral of a rectangle function [4]. It tends to the Gaussian distribution at the limit $m \to \infty$ by the central limit theorem. The discrete B-spline to be used in this paper is defined recursively by

$$b_m[k] = (b_{m-1} * b_1)[k] = \sum_{l=-\infty}^{\infty} b_{m-1}[k-l]b_1[l] \qquad (1)$$

as the m-fold discrete convolution of a sampled rectangle

$$b_1[k] = \begin{cases} 1, & k = 0, 1, 2, \cdots, n-1 \\ 0, & \text{otherwise.} \end{cases} \qquad (2)$$

It tends to the original B-spline at the limit $n \to \infty$ [5]. The z-transform of $b_m[\cdot]$ is

$$B_m(z) = \sum_{k=-\infty}^{\infty} b_m[k]z^{-k} = \left(\frac{1-z^{-n}}{1-z^{-1}}\right)^m. \qquad (3)$$

The inner product or correlation of two discrete B-splines $b_m[\cdot - r]$ and $b_m[\cdot - l]$ can be rearranged in the form of discrete convolution

$$\begin{aligned} \langle b_m[\cdot - r], b_m[\cdot - l]\rangle &= \sum_{k=-\infty}^{\infty} b_m[k-r]b_m[k-l] \\ &= \sum_{k=-\infty}^{\infty} b_m[k]b_m[-(l-r-k)] \\ &= \sum_{k=-\infty}^{\infty} b_m[k]\tilde{b}_m[l-r-k] \\ &= (b_m * \tilde{b}_m)[l-r], \end{aligned}$$

where we have set $\tilde{b}_m[\cdot] = b_m[-\cdot]$. By its z-transform

$$\begin{aligned} & B_m(z)B_m(z^{-1})z^{l-r} \\ = & \left(\frac{1-z^{-n}}{1-z^{-1}}\right)^m \left(\frac{1-z^n}{1-z}\right)^m z^{l-r} \\ = & \left(\frac{1-z^{-n}}{1-z^{-1}}\right)^m \left(\frac{1-z^{-n}}{1-z^{-1}}\right)^m z^{(n-1)m+l-r} \\ = & \left(\frac{1-z^{-n}}{1-z^{-1}}\right)^{2m} z^{m(n-1)+l-r}, \end{aligned}$$

we know that

$$\langle b_m[\cdot - r], b_m[\cdot - l]\rangle = b_{2m}[l - r + m(n-1)]. \qquad (4)$$

Given weighting coefficients $c[\cdot]$ of which z-transform is

$$C(z) = \sum_{l=-\infty}^{\infty} c[l]z^{-l},$$

we can express the weighted sum of discrete B-splines $b_m[\cdot - l]$ in the form of a discrete convolution

$$q[k] = \sum_{l=-\infty}^{\infty} c[l]b_m[k-l] = (c * b_m)[k] \qquad (5)$$

of which the representation by the transfer functions is

$$C(z)B_m(z) = C(z)\left(\frac{1-z^{-n}}{1-z^{-1}}\right)^m. \qquad (6)$$

So we can generate $q[\cdot]$ as the output of a digital filter having the transfer function $\left(\frac{1-z^{-n}}{1-z^{-1}}\right)^m$ for the input $c[\cdot]$. This digital filter can be implemented in two steps: the m-th order accumulation $\left(\frac{1}{1-z^{-1}}\right)^m$ and the m-th order difference $(1-z^{-n})^m$. Although the accumulation may overflow in the first step, it has been known that the final output $q[\cdot]$ stays correct as long as we use the integer arithmetic in the 2's complement representation of which bit-length is long enough to accommodate the theoretical range of $q[\cdot]$ [6]. Since the amplitude of the final output $q[\cdot]$ is bounded by

$$\begin{aligned} \sup_k |q[k]| &\le \sup_l |c[l]| \sum_{k=-\infty}^{\infty} |b_m[k]| \\ &= \sup_l |c[l]| \sum_{k=-\infty}^{\infty} b_m[k] \\ &= \sup_l |c[l]|\, B(1) \\ &= \sup_l |c[l]|\, n^m, \end{aligned} \qquad (7)$$

it cannot be magnified more than n^m times the amplitude of the input $c[\cdot]$. So it suffices for correct computation to add guard bits of the length $m\lceil \log_2 n \rceil$.

For a given profile $p[\cdot]$, let

$$P(z) = \sum_{k=-\infty}^{\infty} p[k]z^{-k}.$$

Then its inner products or correlations with $b_m[\cdot - l]$ can be represented in the form of discrete convolution

$$\begin{aligned}\langle p[\cdot], b_m[\cdot - l]\rangle &= \sum_{k=-\infty}^{\infty} p[k] b_m[k - l] \quad (8)\\ &= \sum_{k=-\infty}^{\infty} p[k] b_m[-(l - k)]\\ &= (p * \tilde{b}_m)[l] \quad (9)\end{aligned}$$

of which the representation by the transfer functions is

$$\begin{aligned}P(z)B_m(z^{-1})z^l &= P(z)\left(\frac{1 - z^n}{1 - z}\right)^m z^l\\ &= P(z)\left(\frac{1 - z^{-n}}{1 - z^{-1}}\right)^m z^{(n-1)m+l}. \quad (10)\end{aligned}$$

So we can compute the inner products by inputting $p[\cdot]$ to a digital filter having the transfer function $\left(\frac{1-z^{-n}}{1-z^{-1}}\right)^m$ and sampling its output at $(n-1)m + l$. This transfer function is the same as the one for generating weighted sums and can also be implemented efficiently in the two steps.

III. Algorithm

We have to approximate a given profile $p[\cdot]$ by a weighted sum of the discrete B-splines $b_m[\cdot - l]$ deployed by the most minute interval 1 under the constraint that the weights should be non-negative.

A. Digital filter to compute inner products $\langle p[\cdot], b_m[\cdot - l]\rangle$

The inner products $\langle p[\cdot], b_m[\cdot - l]\rangle$ of a given profile $p[\cdot]$ and the discrete B-splines $b_m[\cdot - l]$ should usually be evaluated by the standard multiply-and-add architecture according to their definition (8). But, by virtue of (9) and (10), we can do the same only by using the digital filter depicted in Fig. 3.

The first half of the filter in Fig. 3 represents the m-fold accumulation free from the parameter n so that this part has to be operated just once for a given profile. We can evaluate the inner product for different n only by operating the second half. Its computational cost is almost only m subtractions per an inner product on the average.

The mutual inner products $\langle b_m[\cdot - r], b_m[\cdot - l]\rangle$ among the discrete B-splines can be precomputed by (4) and stored in a data table.

B. Non-negative least-square approximation

From the inner products, we are to determine the weighting coefficients $c[\cdot]$ so that the weighted sum

$$q[k] = \sum_{l=0}^{L-1} c[l] b_m[k - l] \quad (11)$$

of the discrete B-splines $b_m[\cdot - l]$ approximate the profile $p[\cdot]$ best in the sense

$$\begin{aligned}E &= \langle p[\cdot] - q[\cdot], p[\cdot] - q[\cdot]\rangle\\ &= \sum_k (p[k] - q[k])^2 \longrightarrow \text{min.} \quad (12)\end{aligned}$$

Such coefficients can be determined by solving the normal linear equations

$$\sum_{l=0}^{L-1} c[l] \langle b_m[\cdot - r], b_m[\cdot - l]\rangle = \langle p[\cdot], b_m[\cdot - l]\rangle, \quad r = 0, 1, 2, ..., L - 1. \quad (13)$$

The resulting coefficients $c[\cdot]$ may be negative whereas they should be constrained to be non-negative.

The least-square approximation under this constraint can be solved by overwriting the negative coefficients by zero, discarding the coefficients and corresponding discrete B-splines from the linear equations, and solving the linear equations repeatedly until all the coefficients get non-negative [8].

Although all the four arithmetic operations in the floating point representation are required to solve the linear equations, this process is not so slow since the discrete B-splines are locally supported to make the equations banded and because the number of involved discrete B-splines decreases during the iterations.

The only and major concern is numerical instability in solving the linear equations. The minutely shifted discrete B-splines are so crowded that they are almost linearly dependent. The numerically obtained initial approximation result is quite imprecise despite the mathematical fact that the initial approximation must theoretically be an exact interpolation having no errors at all. It has been empirically observed that neighboring discrete B-splines are likely to have coefficients of opposite signs. Since the discrete B-splines with negative coefficients should be discarded, the discrete B-splines get sparser in the next iteration. The non-negativity constraint happened to bring in such a nice side effect. In that way, all the numerical experiments up to now with test data taken from real profiles finished successfully at the end.

C. Evaluation of mean square error

It follows from (5) that the approximate profile $q[\cdot]$ can be computed from $c[\cdot]$ by the digital filter depicted in Fig. 4. The mean square error is evaluated from the output $q[\cdot]$ and the original profile $p[\cdot]$ by

$$E_1 = \sqrt{\frac{\sum_k (p[k] - q[k])^2}{\sum_k (p[k])^2}}. \quad (14)$$

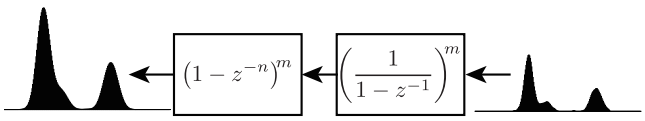


Fig. 3. Digital filter to compute $\langle p[\cdot], b_m[\cdot - l]\rangle$ at $(n-1)m+l$ from $p[k]$

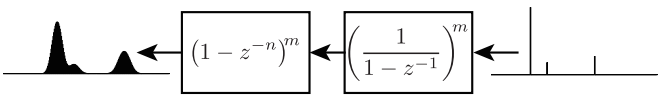


Fig. 4. Digital filter to compute $q[k]$ from $c[k]$

D. Evaluation of sparsity

Among a number of approximations obtained by the above processes *A*, *B* and *C* for various values of n, we choose the best one that has sufficiently small errors and is composed of few discrete B-splines with large weights.

In the case that the mean square error E_2 is very small, the absolute difference $\sum_k |p[k] - q[k]|$ is also small and the sums $\sum_k p[k]$ and $\sum_k q[k]$ of the original and approximate profiles are close to each other because both $p[\cdot]$ and $q[\cdot]$ are non-negative. In this case, the modified coefficients $n^m c[l]$ for the discrete B-splines $\frac{1}{n^m}\ b_m[\cdot - l]$ normalized by its sum $\sum_k b_m[k-l] = n^m$ satisfy

$$q[k] = \sum_{l=0}^{L-1} (n^m\ c[l]) \left(\frac{1}{n^m}\ b_m[k-l] \right) \tag{15}$$

and

$$\sum_{l=0}^{L-1} n^m c[l] = \sum_k q[k] \approx \sum_k p[k] = \text{constant}. \tag{16}$$

In this situation, sparsity in the sense that the approximation $q[\cdot]$ should comprise few large portions is translated into that the coefficients $n^m c[l]$ should comprise few and large ones because $\sum_l n^m c[l]$ is constant and $n^m c[l]$ is non-negative. An index to evaluate this sparsity is

$$E_2 = \sqrt{\sum_{l=0}^{K-1} (n^m c[l])^2}. \tag{17}$$

We take the sparsest approximation giving the largest E_2 out of the good approximations having the mean-square error E_1 smaller than a threshold among various approximations for different values of n.

IV. Numerical Result

The algorithm was applied to test data taken from real profiles. Figure 5 shows approximations of a profile within a short window of 128 sampling points. The order of the discrete B-splines is fixed as $m = 4$. The original profile $p[\cdot]$ is plotted in black, its approximations $q[\cdot]$ for various n is in red. The green curves represent the discrete B-splines weighted by their coefficients to compose the approximation.

The cases for $n \leq 12$ cleared the precision bar conditioned by $E_1 < 0.1$. The case $n = 12$ gave the largest E_2 to be selected as the sparsest among the precise approximations.

Figure 6 shows a whole profile. The best n gets larger as the time passes so that it was sought within each short window. The best n is 12 for the two left hills and 14 for the right one.

V. Conclusions

The discrete B-splines were employed to construct an algorithm for the analysis of ion-mobility spectrometry profiles. This application field requested modification of the standard B-spline approximations in two aspects: the deployment of B-splines by a minute shift interval and the non-negativity constraint on coefficients. The former put us in danger of numerical instability and the latter pulled us out of it.

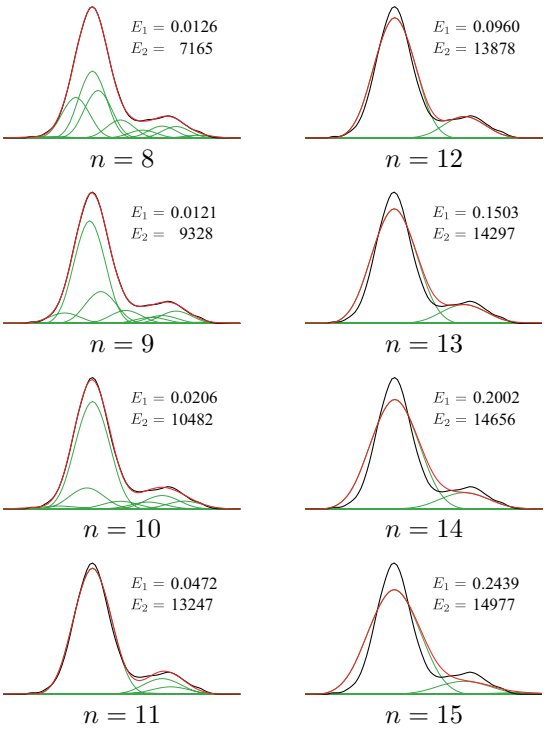


Fig. 5. Approximations of a partial sample profile

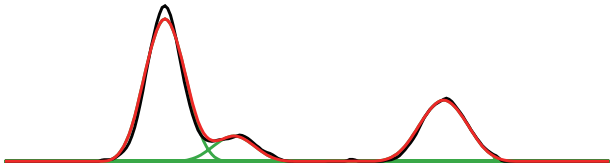


Fig. 6. Approximation of a whole sample profile

A next step is to come up with a single index to balance the approximation error and the sparsity. We may probably have to reformulate the problem within the general theory of sparse approximation. Before a large scale test of the algorithm against various profile data, we should see by simulations whether an artificial profile built up of discrete B-splines is identified in the noise-free and noisy cases.

References

[1] R. W. Purves, R. Guevremont, S. Day, C. W. Pipich and M. S. Matyjaszczyk, Mass spectrometric characterization of a high-field asymmetric waveform ion mobility spectrometer, *Rev. Sci. Instrum.*, **69**, 4094-4105, 1998.

[2] Owlstone Nanotech, Inc. `http://www.owlstonenanotech.com/technology/`

[3] ATONARP, Inc. `http://www.atonarp.com/en/`

[4] M. Unser, Splines: A perfect fit for signal and image processing, *IEEE Signal Process. Magazine*, **16**(6), 22–38,1999.

[5] K. Ichige and M. Kamada, An approximation for discrete B-splines in time domain, *IEEE Signal Process. Lett.*, **4**, 82–84, 1997.

[6] T. Saramäki, Y. Neuvo and S. K. Mitra, Design of computationally efficient interpolated FIR filters, *IEEE Trans. Circuits & Syst.*, **35**(1), 70– 88, 1988.

[7] K. Ichige, M. Kamada and R. Ishii, A simple scheme of decomposing and reconstructing continuous-time signals by B-splines, *IEICE Trans. Fundamentals*, **E81-A**, 2391–2399, 1998.

[8] C. L. Lawson and R. J. Hanson, *Solving Least Squares Problems*, SIAM, 1987.

Tracking Dynamic Sparse Signals with Kalman Filters: Framework and Improved Inference

Evripidis Karseras , Kin Leung and Wei Dai
Department of Electrical and Electronic Engineering, Imperial College, London, UK
{e.karseras11, kin.leung, wei.dai1}@imperial.ac.uk

***Abstract*—The standard Kalman filter performs optimally for conventional signals but tends to fail when it comes to recovering dynamic sparse signals. In this paper a method to solve this problem is proposed. The basic idea is to model the system dynamics with a hierarchical Bayesian network which successfully captures the inherent sparsity of the data, in contrast to the traditional state-space model. This probabilistic model provides all the necessary statistical information needed to perform sparsity-aware predictions and updates in the Kalman filter steps. A set of theorems show that a properly scaled version of the associated cost function can lead to less greedy optimisation algorithms, unlike the ones previously proposed. It is demonstrated empirically that the proposed method outperforms the traditional Kalman filter for dynamic sparse signals and also how the redesigned inference algorithm, termed here Bayesian Subspace Pursuit (BSP) greatly improves the inference procedure.**

I. Introduction

The Kalman filter has been the workhorse approach in the area of linear dynamic system modelling in both practical and theoretic scenarios. The escalating trend towards sparse signal representation has rendered this estimator to be useless when it comes to tracking *dynamic sparse signals*. It is easy to verify that the estimation process behind the Kalman filter is not fit for sparse signals. Intuitively, the Gaussian *prior* distribution placed over the system's observations does not place any sparsity constraints over the space of all possible solutions.

The Kalman filter was externally modified in the bibliography to admit sparse solutions. The idea in [1] and [2] is to enforce sparsity by thresholds. Work in [3] adopts a probabilistic model but signal amplitudes and support are estimated separately. Finally, the techniques presented in [4] use prior sparsity knowledge into the tracking process. All these approaches typically require a number of parameters to be pre-determined. It also remains unclear how these methods perform towards model and parameter mismatch.

For a single time instance of the sparse reconstruction problem, the Relevance Vector Machine (RVM) introduced in [10] was used with great success in Compressed Sensing applications [5] and basis selection [6]. The hierarchical Bayesian network behind the RVM achieves highly sparse models for the observations not only providing estimates for sparse signals but on their full posterior distributions as well. This is of great importance since it provides all the necessary statistical information to use in the prediction step of the tracking process. Additionally, the inference procedure used in this framework allows for automatic determination of the active components hence the need for a pre-determined level of sparsity is eliminated. This is an appealing attribute for an on-line tracking algorithm.

In this work the aforementioned Bayesian network is employed to extend the state-space model adopted in the traditional Kalman filter. This way the problem of modelling sparsity is tackled efficiently. The resulting statistical information from the inference procedure is then incorporated in the Kalman filter steps thus producing sparsity-aware state estimates.

A set of theorems dictate that a proper scaling of the cost function associated with the inference procedure can lead to more efficient inference algorithms. The techniques initially proposed are greedy methods at heart. By scaling the cost function with the noise variance, and by using knowledge gained from well known compressed sensing algorithms, it is possible to redesign these methods to admit better qualities. The gains are two fold. Firstly, the improved inference mechanism bears far better qualities than the one previously proposed. Secondly, the proposed method outperforms the traditional Kalman filter in terms of reconstruction error when it comes to dynamic sparse signals.

In Section II we present the basic idea for amalgamating the Bayesian network of the RVM in the Kalman filter, termed here Hierarchical Bayesian Kalman filter (HB-Kalman). In Section III we present as set of theorems and explain the motivation to improve upon previous techniques. Additionally we provide the steps for a revised inference algorithm based on the Subspace Pursuit (SP) reconstruction algorithm in [8], termed here Bayesian Subspace Pursuit (BSP). In Section IV we demonstrate the performance of the proposed methods in some synthetic scenarios.

II. Hierarchical Bayesian Kalman Filter

The system model is described by the following equations:

$$\boldsymbol{x}_t = \boldsymbol{F}_t \boldsymbol{x}_{t-1} + \boldsymbol{z}_t, \tag{1}$$

$$\boldsymbol{y}_t = \boldsymbol{\Phi}_t \boldsymbol{x}_t + \boldsymbol{n}_t. \tag{2}$$

where vectors $\boldsymbol{x}_t, \boldsymbol{y}_t$ denote the system's state and observation respectively. The state innovation and observation noise processes are modelled by $\boldsymbol{z}_t$ and $\boldsymbol{n}_t$ respectively.

The authors would like to acknowledge the European Commission for funding SmartEN ITN (Grant No. 238726) under the Marie Curie ITN FP7 programme.

We assume that signal $\boldsymbol{x}_t \in \mathbb{R}^n$ is sparse in some domain. which is considered to remain the same at all time instances (e.g the frames of a video are sparse in the wavelet domain). This allows to set the state transition matrix $\boldsymbol{F}_t$ equal to the unitary matrix $\boldsymbol{I}$. Equation (1) becomes:

$$\boldsymbol{x}_t = \boldsymbol{x}_{t-1} + \boldsymbol{z}_t$$

As in the standard Kalman filter we adopt the Gaussian assumption so that: $p(\boldsymbol{z}_t) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{Z}_t)$, $p(\boldsymbol{n}_t) = \mathcal{N}\left(\boldsymbol{0}, \sigma^2\boldsymbol{I}\right)$; and $p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_{t-1}, \boldsymbol{Z}_t)$ and $p\left(\boldsymbol{y}_t|\boldsymbol{x}_t\right) = \mathcal{N}\left(\boldsymbol{\Phi}\boldsymbol{x}_t, \sigma^2\boldsymbol{I}\right)$. At each time instance, the Kalman filter involves the prediction step where the parameters of $p(\boldsymbol{x}_t|\boldsymbol{y}_{t-1})$ are calculated, while the update step evaluates those of $p(\boldsymbol{x}_t|\boldsymbol{y}_t)$. The advantages of the standard Kalman filter include the ability to track the full statistics, and that the mean squared error solution coincides with the maximum posterior solution which has a closed form. The major issue when applying the filter to dynamic sparse signals, is that the solution is typically not sparse. This drawback is due to the fact that in the standard approach, the covariance matrix $\boldsymbol{Z}_t$ is priorly given. Variants of the Kalman filter such as the non-linear Kalman filter also suffer because of the special nature of the of the non-linearities associated with sparse reconstruction.

To alleviate this problem, the key idea behind Sparse Bayesian Learning (SBL) [10] is employed. As opposed to the traditional Kalman filter where the covariance matrix $\boldsymbol{Z}_t$ of $\boldsymbol{z}_t$ is given, here it is assumed that the state innovation process is given by:

$$\boldsymbol{z}_t \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{A}_t^{-1}\right),$$

where $\boldsymbol{A} = \operatorname{diag}(\boldsymbol{\alpha}) = \operatorname{diag}\left([\alpha_1, \cdots, \alpha_n]_t\right)$, and the hyper-parameters α_i are unknown and have to be learned from $\boldsymbol{y}_t$. To see how this promotes a sparse solution, let us drop the subscript t for simplicity. Then it holds that:

$$p(\boldsymbol{x}|\boldsymbol{\alpha}) = \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{A}^{-1}\right) = \prod_{i=1}^{n} \mathcal{N}\left(0, \alpha_i^{-1}\right).$$

By driving $\alpha_i = +\infty$ it means that $p(x_i|\alpha_i) = \mathcal{N}(0, 0)$; hence it is certain that $x_i = 0$. What remains is to find the maximum likelihood solution of $\boldsymbol{\alpha}$ for the given observation vector $\boldsymbol{y}$. The explicit form of the likelihood function $p\left(\boldsymbol{y}|\boldsymbol{\alpha}, \sigma^2\right)$ was derived in [10] and a set of fast algorithms to estimate $\boldsymbol{\alpha}$ and consequently $\boldsymbol{z}$ and $\boldsymbol{x}$ are proposed in [9].

Finally the principles behind the Kalman filter and SBL are put together. Similar to the standard Kalman filter, two steps, prediction and update, need to be performed at each time instance. In the prediction step, one has to evaluate:

$$\boldsymbol{\mu}_{t|t-1} = \boldsymbol{\mu}_{t-1|t-1}, \quad \boldsymbol{\Sigma}_{t|t-1} = \boldsymbol{\Sigma}_{t-1|t-1} + \boldsymbol{A}_t^{-1},$$
$$\boldsymbol{y}_{t|t-1} = \boldsymbol{\Phi}_t\boldsymbol{\mu}_{t|t-1}, \quad \boldsymbol{y}_{e,t} = \boldsymbol{y}_t - \boldsymbol{y}_{t|t-1}.$$

where the notation $t|t-1$ means prediction at time instance t for measurements up to time instance $t-1$. In the update step, one computes:

$$\boldsymbol{K}_t = \boldsymbol{\Sigma}_{t|t-1}\boldsymbol{\Phi}_t^T(\sigma^2\boldsymbol{I} + \boldsymbol{\Phi}_t\boldsymbol{\Sigma}_{t|t-1}\boldsymbol{\Phi}_t^T)^{-1},$$
$$\boldsymbol{\mu}_{t|t} = \boldsymbol{\mu}_{t|t-1} + \boldsymbol{K}_t\boldsymbol{y}_{e,t}, \quad \boldsymbol{\Sigma}_{t|t} = (\boldsymbol{I} - \boldsymbol{K}_t\boldsymbol{\Phi}_t)\boldsymbol{\Sigma}_{t|t-1}.$$

Differently from the standard Kalman filter, one has to perform the additional step of learning the hyper-parameters $\boldsymbol{\alpha}_t$. From Equation (2) we get $\boldsymbol{y}_{e,t} = \boldsymbol{\Phi}_t\boldsymbol{z}_t + \boldsymbol{n}_t$ where a sparse $\boldsymbol{z}_t$ is preferred to produce a sparse $\boldsymbol{x}_t$. Following the analysis in [10] and [9], maximising the likelihood $p(\boldsymbol{y}_t|\boldsymbol{\alpha}_t)$ is equivalent to minimising the following cost function:

$$\mathcal{L}(\boldsymbol{\alpha}_t) = \log|\boldsymbol{\Sigma}_\alpha| + \boldsymbol{y}_{e,t}^T\boldsymbol{\Sigma}_{\boldsymbol{\alpha}}^{-1}\boldsymbol{y}_{e,t}, \tag{3}$$

where $\boldsymbol{\Sigma}_{\boldsymbol{\alpha}} = \sigma^2\boldsymbol{I} + \boldsymbol{\Phi}_t\boldsymbol{A}_t^{-1}\boldsymbol{\Phi}_t^T$. The algorithms described in [9] can be applied to estimate $\boldsymbol{\alpha}_t$. Note that the cost function $\mathcal{L}(\boldsymbol{\alpha})$ is not convex. The obtained estimate $\boldsymbol{\alpha}_t$ is generally sub-optimal and details on the estimation of the globally optimal $\boldsymbol{\alpha}_t$ are given in the next section.

III. Bayesian Subspace Pursuit

Here we discuss the performance guarantees for a single time instance of the inference procedure. For convenience, subscript t is dropped and focus is turned to Equation (2) where $\boldsymbol{x}|\boldsymbol{\alpha} \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{A}^{-1}\right)$. This was analysed in [6] for the purpose of Basis Selection. It had also been proven in [6] that a maximally sparse solution of $\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{x}$ attains the global minimum of the cost function. However, the analysis did not specify the conditions to avoid local minima. By contrast, we provide a more refined analysis. Due to space constraints, only the main results are presented.

We follow [6] by driving the noise variance $\sigma^2 \to 0$. The following Theorem specifies the behaviour of the cost function $\mathcal{L}(\boldsymbol{\alpha})$.

Theorem 1. *For any given $\boldsymbol{\alpha}$, define the set $\mathcal{I} \triangleq \{1 \le i \le n:\ 0 < \alpha_i < \infty\}$. Then it holds that:*

$$\lim_{\sigma^2\to 0} \sigma^2\mathcal{L}(\boldsymbol{\alpha}) = \left\|\boldsymbol{y} - \boldsymbol{\Phi}_{\mathcal{I}}\boldsymbol{\Phi}_{\mathcal{I}}^{\dagger}\boldsymbol{y}\right\|_2^2, \tag{4}$$

where $\boldsymbol{\Phi}_{\mathcal{I}}$ is a sub-matrix of $\boldsymbol{\Phi}$ formed by the columns indexed by $\mathcal{I}$, and $\boldsymbol{\Phi}_{\mathcal{I}}^{\dagger}$ denotes the pseudo-inverse of $\boldsymbol{\Phi}_{\mathcal{I}}$.

Furthermore, if $|\mathcal{I}| < m$ and $\boldsymbol{y} \in \operatorname{span}(\boldsymbol{\Phi}_{\mathcal{I}})$, then $\mathcal{L}(\boldsymbol{\alpha}) \to -\infty$ and $\sigma^2\mathcal{L}(\boldsymbol{\alpha}) \to 0$ as $\sigma^2 \to 0$.

Two observations can be obtained: (a) the scenarios analysed in [6] can be seen as special cases of Theorem 1 where $\mathcal{L}(\boldsymbol{\alpha}) \to -\infty$; and (b) a proper scaling of the cost function gives the squared ℓ_2-norm of the reconstruction error. Reconstruction is then equivalent to recovering a support set that minimises the reconstruction distortion. This principle is the same as the one behind many greedy algorithms such as the OMP [7] and SP [8]. Theorem 1 suggests such connections.

According to Theorem 1 the key quantities concerning the algorithms described in [9] must be scaled by the noise variance. The original formulae can be found in [9] while the revised ones are given below:

$$\sigma^{-2}\boldsymbol{\Sigma}_x = \left(\sigma^2\boldsymbol{A}_{\mathcal{I}} + \boldsymbol{\Phi}_{\mathcal{I}}^T\boldsymbol{\Phi}_{\mathcal{I}}\right)^{-1}, \quad \boldsymbol{\mu}_x = \sigma^{-2}\boldsymbol{\Sigma}_x\boldsymbol{\Phi}_{\mathcal{I}}^T\boldsymbol{y},$$
$$\sigma^2\boldsymbol{C}_{-i}^{-1} = \boldsymbol{I} - \boldsymbol{\Phi}_{\mathcal{I}-i}\left(\sigma^2\boldsymbol{A}_{\mathcal{I}-i} + \boldsymbol{\Phi}_{\mathcal{I}-i}^T\boldsymbol{\Phi}_{\mathcal{I}-i}\right)^{-1}\boldsymbol{\Phi}_{\mathcal{I}-i}^T,$$
$$\bar{s}_i = \sigma^2 s_i = \boldsymbol{\phi}_i^T\left(\sigma^2\boldsymbol{C}_{-i}^{-1}\right)\boldsymbol{\phi}_i, \quad \bar{q}_i = \sigma^2 q_i = \boldsymbol{\phi}_i^T\left(\sigma^2\boldsymbol{C}_{-i}^{-1}\right)\boldsymbol{y}.$$

Subscript $\mathcal{I}$ denotes the set of indices for which $0 < \alpha_i < +\infty$. The notation $\mathcal{I} - i$ means removal of index i from $\mathcal{I}$.

Subsequently formula [9, Equation (20)] for the optimal α_i given all other α_j, $j \neq i$, becomes:

$$\alpha_i = \frac{\bar{s}_i^2}{\bar{q}_i^2 - \sigma^2 \bar{s}_i}$$

Finally the scaled cost function becomes:

$$\bar{\mathcal{L}} = \sigma^2 \mathcal{L} = \sigma^2 \log \left|\sigma^2 \boldsymbol{I} + \boldsymbol{\Phi}_{\mathcal{I}} \boldsymbol{A}_{\mathcal{I}}^{-1} \boldsymbol{\Phi}_{\mathcal{I}}^T\right| + \boldsymbol{y}^T \left(\boldsymbol{I} - \boldsymbol{\Phi}_{\mathcal{I}} (\sigma^2 \boldsymbol{A}_{\mathcal{I}} + \boldsymbol{\Phi}_{\mathcal{I}}^T \boldsymbol{\Phi}_{\mathcal{I}})^{-1} \boldsymbol{\Phi}_{\mathcal{I}}^T\right) \boldsymbol{y}.$$

Let us clarify the importance of using the scaled quantities. Assume that $\sigma^2 = 0$. It is then easy to show that the original formulae in [9] result in poor performance. Scaling the cost function (and consequently these quantities) is necessary when we want to account for a given noise variance. This may seem irrelevant but in many tracking applications the noise floor is assumed to be estimated in some way or provided by the manufacturer for specific devices. The initial work in [10] provides the formula to infer the noise level from the observations. The scaled versions of the aforementioned quantities can still be applied if desired.

We now have a better understanding of the inference procedure but it still remains unclear what the selection criterion for the basis functions should be. In [9] selection is based on the value of α_i which maximises the difference $\Delta\mathcal{L}$ in the likelihood function, while algorithms such as the OMP and SP make decisions on different grounds. The following Theorem sheds some light on this matter.

Theorem 2. *Assume the noiseless setting $\boldsymbol{y} = \boldsymbol{\Phi x}$ where $\boldsymbol{\Phi} \in \mathbb{R}^{m \times n}$ and $\boldsymbol{\phi}_i^T \boldsymbol{\phi}_i = 1$ for all $1 \leq i \leq n$. Furthermore assume that $t = \max \left|\boldsymbol{\phi}_i^T \boldsymbol{\phi}_j\right|$ for $1 \leq i \neq j \leq n$. Then an algorithm similar to the one in [9] based on one of the following criteria recovers all s-sparse signals exactly given the sufficient condition $t < 0.375/s$; (a) the maximum $\sigma^2 \Delta\mathcal{L}$, (b) the maximum x_i or (c) the minimum α_i.*

Theorem 2 is the starting point for redesigning the inference algorithm. Based on the scaled quantities we can rederive the algorithm in [9] termed Fast Marginal Likelihood Maximisation (FMLM). It is possible to have variants with OMP-like performance guarantees based on different criteria as Theorem 2 suggests. Actually the inference algorithm then greatly resembles the OMP; where the basis functions are recovered sequentially with decreasing order of correlation with the residual signal. For brevity we only present the version based on maximising $\boldsymbol{x}_i$ hence the algorithm is termed FMLM-$\boldsymbol{x}_i$. The steps are given in Algorithm 1.

Theorem 3. *Assume that the same conditions hold as in Theorem 2. An algorithm similar to the one in [9] based on the less greedy criterion of maximum $\theta_i = \bar{q}_i$, recovers all s-sparse signals exactly given the sufficient condition $t < 0.5/s$. The algorithm presented in Algorithm 2 recovers all s-sparse signals exactly if matrix $\boldsymbol{\Phi}$ satisfies the RIP with parameter $\delta_{3s} < 0.205$.*

Theorem 3 suggests further improvements to the performance guarantees, to match those of the OMP by altering the optimisation criterion. Also, results from [8] motivate us to extend the FMLM procedure to a less greedy optimisation procedure by borrowing ideas from the SP algorithm. The SP selects a subset of basis functions at each time instance based also on correlation maximisation, but adds a backtracking step so as to retain only the sparse components with the largest magnitudes. The redesigned algorithm termed here Bayesian Subspace Pursuit is described in Algorithm 2.

Algorithm 1 FMLM-$\boldsymbol{x}_i$

Input: $\boldsymbol{\Phi}, \boldsymbol{y}, \sigma^2$
Initialise:
- $\hat{T} = \{\text{index } i \in [1, n] \text{ for maximum } |\boldsymbol{\phi}_i^T \boldsymbol{y}|\}$.

Iteration:
- Calculate values of α_i and $[\boldsymbol{\mu}_x]_i$ for $i \in [1, n] \setminus \hat{T}$.
- $T' = \hat{T} \cup \{\text{index } i \text{ corresponding to the maximum value of } [\boldsymbol{\mu}_{\boldsymbol{x}}]_i \text{ for } i \notin \hat{T}\}$.
- Calculate values α_i for $i \in T'$.
- $\tilde{T} = \{i \in T' : 0 < \alpha_i < +\infty\}$.
- If $|\bar{\mathcal{L}}_{\tilde{T}} - \bar{\mathcal{L}}_{\hat{T}}| = 0$ then compute $\sigma^{-2}\boldsymbol{\Sigma}_x$, $\boldsymbol{\mu}_x$ for $\tilde{T}$ and quit. Set $\hat{T} = \tilde{T}$ and continue otherwise.

Output:
- Estimated support set $\tilde{T}$ and sparse signal $\tilde{\boldsymbol{x}}$ with $|\tilde{T}|$ non-zero components, $\tilde{\boldsymbol{x}}_{\tilde{T}} = \boldsymbol{\mu}_x$.
- Estimated covariance matrix $\sigma^{-2}\boldsymbol{\Sigma}_x$.

Algorithm 2 Bayesian Subspace Pursuit

Input: $\boldsymbol{\Phi}, \boldsymbol{y}, \sigma^2$
Initialise:
- $\hat{T} = \{\text{index } i \in [1, n] \text{ for minimum } \alpha_i = \frac{1}{|\boldsymbol{\phi}_i^T \boldsymbol{y}|}\}$.

Iteration:
- Store $\alpha_{max} = \arg\max_{i \in \hat{T}} |\alpha_i|$.
- Calculate values α_i and $\theta_i = \bar{q}_i^2 - \bar{s}_i$ for $i \in [1, n]$.
- Calculate values $t_{\theta_i > 0} = |\{i \in [1, n] : \theta_i > 0\}|$ and $t_{\alpha_i \leq a_{max}} = |\{i \in [1, n] : |\alpha_i| \leq a_{max}\}|$.
- If $t_{\theta_i > 0} = 0$ then $s = t_{\alpha_i \leq a_{max}} + 1$ else $s = t_{\theta_i > 0} + t_{\alpha_i \leq a_{max}}$.
- $T' = \hat{T} \cup \{\text{indices corresponding to } s \text{ smallest values of } \alpha_i \text{ for } i \in [1, n]\}$.
- Compute $\sigma^{-2}\boldsymbol{\Sigma}_x$ and $\boldsymbol{\mu}_x$ for T'.
- $\tilde{T} = \{\text{indices corresponding to } s \text{ largest non-zero values of } |\boldsymbol{\mu}_x| \text{ for which } 0 < \alpha_i < +\infty\}$.
- If $|\bar{\mathcal{L}}_{\tilde{T}} - \bar{\mathcal{L}}_{\hat{T}}| = 0$ then quit. Otherwise set $\hat{T} = \tilde{T}$ and continue.

Output:
- Estimated support set $\tilde{T}$ and sparse signal $\tilde{\boldsymbol{x}}$ with $|\tilde{T}|$ non-zero components, $\tilde{\boldsymbol{x}}_{\tilde{T}} = \boldsymbol{\mu}_x$.
- Estimated covariance matrix $\sigma^{-2}\boldsymbol{\Sigma}_x$ for $\tilde{T}$.

IV. Empirical Results

A. Single Time Instance

We concentrate on the performance of the algorithms for a single time instance and for $\sigma^2 = 0$. The algorithms under comparison are the FMLM algorithm as originally presented

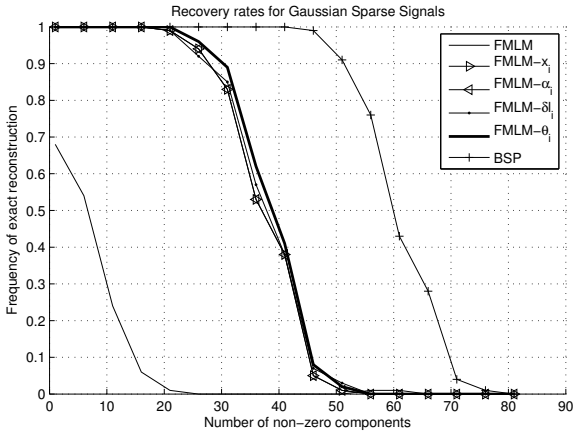


Figure 1. Exact reconstruction rates for $m = 128, n = 256$

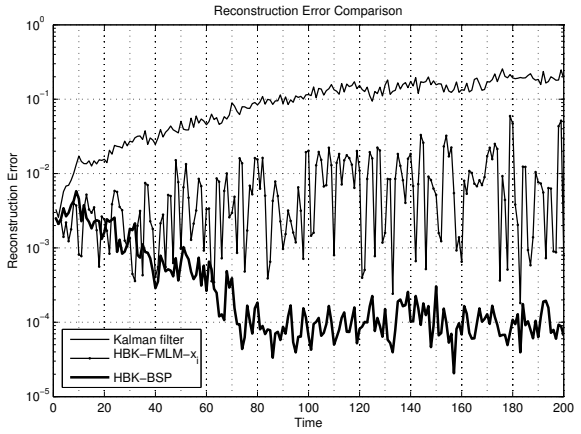


Figure 2. Reconstruction error comparison for $m = 128$.

in [9], the variants based on the scaled quantities; FMLM-$\boldsymbol{x}_i$, FMLM-$\boldsymbol{\alpha}_i$, FMLM-$\boldsymbol{\delta} l_i$, FMLM-$\boldsymbol{\theta}_i$ and the BSP. The experiment is as follows:

1) Generate $\boldsymbol{\Phi} \in \mathbb{R}^{128\times 256}$ with i.i.d entries from $\mathcal{N}\left(0, \frac{1}{m}\right)$.
2) Generate T uniformly at random so that $|T| = K$.
3) Choose values for $\boldsymbol{x}_T$ from $\mathcal{N}(0, 1)$.
4) Compute $\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{x}$ and apply a reconstruction algorithms. Compare estimate $\hat{\boldsymbol{x}}$ to $\boldsymbol{x}$.
5) Repeat experiment for increasing values of K and for 100 realisations.

The results from this procedure are depicted in Figure 1. The first critical observation is that the original FMLM performs poorly when $\sigma^2 = 0$ due to the improperly scaled cost function. The three scaled variants of FMLM based on the criteria mentioned in Theorem 2 perform - within computational accuracy - in the same manner. We observe the increase in performance for FMLM-$\boldsymbol{\theta}_i$, a consequence of altering the selection criterion to $\theta_i = \bar{q}_i$. Even though changing the criterion gives theoretically better performance as Theorem 3 suggests, empirically this gain is not great. By redesigning the inference algorithm based on ideas from the SP we are able to achieve far better performance, as the curve for the BSP algorithm shows.

B. Dynamic Sparse Signal

We now compare the proposed method, HB-Kalman filter against the original Kalman filter. Signal $\boldsymbol{x}_t \in \mathbb{R}^n$ is assumed to be sparse in its natural basis with support set $\mathcal{S}$ chosen uniformly at random from $[1, n]$ where $n = 256$. The magnitudes of the non-zero entries of $\boldsymbol{x}_t$ evolve according to Equation II with $Z_i = \sigma_z^2 I$ with $\sigma_z^2 = 0.1$. The simulation time for this experiment was $T = 200$ time instances. At two randomly chosen time instances: $T = 50$ and $T = 150$, a change in the support of $\boldsymbol{x}_t$ is introduced. A non-zero component is added to the support of $\boldsymbol{x}_{50}$ and a non-zero component is removed from the support of $\boldsymbol{x}_{150}$. Apart from these two time instances the support of $\boldsymbol{x}_t$ remains unchanged. At $T = 1$ the support is initialised with $K = 30$ non-zero components. Observation variance is set to $\sigma_n^2 = 0.01$ for the entire simulation time. We compare the following techniques; the classic Kalman filter, the HBK with FMLM-$\boldsymbol{x}_i$ as the optimisation procedure and the HBK with BSP.

In this scenario noisy measurements $\boldsymbol{y}_t$ are taken by choosing the design matrix $\boldsymbol{\Phi}_t \in \mathbb{R}^{128\times 256}$ as described in subsection IV-A and is re-sampled at each time instance. The number of observations m remains constant at each time instance. In Figure 2 we primarily notice how the HBK outperforms the original Kalman filter, direct consequence of the sparse dynamic model. The HBK-BSP captures the evolution in the support set with greater success due to the improved optimisation algorithm.

REFERENCES

[1] N. Vaswani, "Kalman filtered compressed sensing," in *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*, oct. 2008, pp. 893 –896.

[2] A. Carmi, P. Gurfil, and D. Kanevsky, "Methods for sparse signal recovery using kalman filtering with embedded pseudo-measurement norms and quasi-norms," *Signal Processing, IEEE Transactions on*, vol. 58, no. 4, pp. 2405 –2409, april 2010.

[3] J. Ziniel, L.C. Potter, and P. Schniter, "Tracking and smoothing of time-varying sparse signals via approximate belief propagation," in *Signals, Systems and Computers (ASILOMAR), 2010 Conference Record of the Forty Fourth Asilomar Conference on*. IEEE, 2010, pp. 808–812.

[4] A. Charles, M.S. Asif, J. Romberg, and C. Rozell, "Sparsity penalties in dynamical system estimation," in *Information Sciences and Systems (CISS), 2011 45th Annual Conference on*. IEEE, 2011, pp. 1–6.

[5] Shihao Ji, Ya Xue, and L. Carin, "Bayesian compressive sensing," *Signal Processing, IEEE Transactions on*, vol. 56, no. 6, pp. 2346 –2356, june 2008.

[6] D.P. Wipf and B.D. Rao, "Sparse bayesian learning for basis selection," *Signal Processing, IEEE Transactions on*, vol. 52, no. 8, pp. 2153 – 2164, aug. 2004.

[7] J.A. Tropp, "Greed is good: algorithmic results for sparse approximation," *IEEE Trans. Inform. Theory*, vol. 50, no. 10, pp. 2231 – 2242, oct. 2004.

[8] Wei Dai and Olgica Milenkovic, "Subspace pursuit for compressive sensing signal reconstruction," *IEEE Trans. Inform. Theory*, vol. 55, pp. 2230–2249, 2009.

[9] M.E. Tipping, A.C. Faul, et al., "Fast marginal likelihood maximisation for sparse Bayesian models," in *International workshop on artificial intelligence and statistics*. Jan, 2003, vol. 1.

[10] Michael E. Tipping, "The relevance vector machine," 2000.

The Variation Detracting Property of some Shannon Sampling Series and their Derivatives

Andi Kivinukk
Department of Mathematics,
Tallinn University, Narva Road 25
Tallinn, 10120, Estonia
Email: andik@tlu.ee

Tarmo Metsmägi
Department of Mathematics,
Tallinn University, Narva Road 25
Tallinn, 10120, Estonia
Email: tmetsmag@tlu.ee

***Abstract*—In this paper are considered some generalized Shannon sampling operators which preserve the total variation of functions and their derivatives. For that purpose will be used the averaged kernel functions of certain even bandlimited kernel functions.**

I. Introduction

In this paper we investigate the generalized Shannon sampling operators, which preserve the total variation of functions and their derivatives.

Dealing with the class of functions of bounded variation $BV[0,1]$, the Bernstein polynomials have the following the total variation preserving property (due to G.G.Lorentz, 1937)

$$V_{[0,1]}[B_n f] \leq V_{[0,1]}[f],$$

where $f \in BV[0,1]$. In [3] this was called as the *variation detracting property,* sometimes called also as the variational diminishing property. Such kind of the total variation preserving property is known for many positive operators [1].

There have been also interests in the variation detracting property for the derivative of the Bernstein operator (e.g. [5]) expressed by the inequality

$$V_{[0,1]}[(B_n f)'] \leq V_{[0,1]}[f'].$$

The generalized Shannon sampling operators [4] for the uniformly continuous and bounded functions $f \in C(\mathbb{R})$ are given by ($t \in \mathbb{R}$; $W > 0$)

$$(S_W f)(t) := \sum_{k=-\infty}^{\infty} f(\frac{k}{W}) s(Wt-k). \tag{1}$$

The variation detracting property for the generalized Shannon sampling operators could be in form:

let $f \in BV(\mathbb{R})$ implies $S_W f \in BV(\mathbb{R})$ and

$$V_{\mathbb{R}}[S_W f] \leq C_0(S_W)\ V_{\mathbb{R}}[f]$$

is valid, where the constant $C_0 = C_0(S_W)$ depends on the norm of the operator $S_W : C(\mathbb{R}) \to C(\mathbb{R})$. The variation detracting property for the derivatives of the generalized Shannon sampling operators could be in form:

let $f' \in BV(\mathbb{R})$ implies $(S_W f)' \in BV(\mathbb{R})$ and

$$V_{\mathbb{R}}[(S_W f)'] \leq C_1(S_W)\ V_{\mathbb{R}}[f']$$

for some constant $C_1 = C_1(S_W)$ depending on the norm of the operator S_W.

For any $f \in C(\mathbb{R})$ the operators S_W are well-defined, if the kernel function s satisfies the condition

$$\sup_{u\in\mathbb{R}} \sum_{k=-\infty}^{\infty} |s(u-k)| < \infty, \tag{2}$$

hence $s \in L^1(\mathbb{R})$. Moreover [4], $\{S_W\}_{W>0}$ defines a family of bounded linear operators from $C(\mathbb{R})$ into itself, having its operator norm

$$\|S_W\| = \sup_{u\in\mathbb{R}} \sum_{k=-\infty}^{\infty} |s(u-k)| \quad (W>0). \tag{3}$$

In our approach the kernel functions of sampling operators defined above will be some even band-limited functions s, i.e. $s \in L^1(\mathbb{R})$, and these are given as the Fourier transform of an even window function $\lambda \in C_{[-1,1]}$, $\lambda(0) = 1$, $\lambda(u) = 0$ ($|u| \geqslant 1$) by the equality

$$s(t) := s(\lambda; t) := \int_0^1 \lambda(u) \cos(\pi t u)\ du. \tag{4}$$

These type of window (also called as the apodization) functions have been widely used in applications (e.g., [2], [10] and literature cited there), in Signal Analysis in particular, very long time.

The leading idea to consider the variation detracting property of (1) is to construct some related kernels to the kernel (4). For the kernels defined by (4) holds the following proposition.

Proposition 1. *Define the related kernels to the kernel (4) as follows:*

$$s_m(t) := \int_0^1 \frac{\lambda(u)}{\operatorname{sinc}(mu)} \cos(\pi t u) du \tag{5}$$

for $0 < m \leq 1$*, and*

$$s_{m,n}(t) := \int_0^1 \frac{\lambda(u)}{\operatorname{sinc}(mu)\operatorname{sinc}(nu)} \cos(\pi t u) du \tag{6}$$

for $0 < m, n \leq 1$. *Then*

$$s(t) = \frac{1}{2m}\int_{-m}^{m} s_m(t+x)dx \quad (7)$$

$$= \frac{1}{4mn}\int_{-m}^{m} dx \int_{-n}^{n} s_{m,n}(t+x+y)dy, \quad (8)$$

$$s_m(t) = \frac{1}{2n}\int_{-n}^{n} s_{m,n}(t+y)dy. \quad (9)$$

Remark 1. The kernels (5) and (6) are well-defined, if in case $m = 1$ or $n = 1$ we assume the existence of the continuous left derivative with value $\lambda\,'(1-) = 0$. If $m = n = 1$, we assume the existence of the continuous left second derivative with value $\lambda\,''(1-) = 0$.

Since (8) and (9), we have the following

Corollary 1. *If* $s_{m,n} \in L^1(\mathbb{R})$, *then* $s_m \in L^1(\mathbb{R})$ *and*

$$\|s\|_1 \leq \|s_m\|_1 \leq \|s_{m,n}\|_1.$$

Remark 2. If $s \in L^1(\mathbb{R})$, then by (4) $s \in B^1_\pi \subset L^1(\mathbb{R})$, where B^1_π is a Bernstein class [6]. Under conditions of Remark 1, and by Proposition 1, equations (5), (6), assuming $s_{m,n} \in L^1(\mathbb{R})$, the kernels $s_m, s_{m,n} \in B^1_\pi$ as well.

Some examples of kernels defined by (4), which appear in our applications, are given as follows:

1) $\lambda(u) = 1$ defines the sinc function (the exceptional case, because $\operatorname{sinc}(\cdot) \notin L^1(\mathbb{R})$)

$$\operatorname{sinc}(t) := \frac{\sin \pi t}{\pi t};$$

2) $\lambda(u) = \operatorname{sinc} u$ defines the Lanczos' kernel (with corresponding operator L_W), which by (4) appears to be the averaged sinc-function

$$s_L(t) = \frac{1}{2}\int_{-1}^{1} \operatorname{sinc}(t+v)\, dv; \quad (10)$$

3) $\lambda(u) = \cos\frac{\pi u}{2}$ defines the Rogosinski-type kernel (with corresponding operator R_W) in the form

$$r_0(t) = \frac{1}{2\pi} \cdot \frac{\cos \pi t}{\frac{1}{4} - t^2}.$$

II. Total variation on $\mathbb{R}$

Let us consider functions of bounded variation in the following meaning.

Definition. We say that $f \in BV(\mathbb{R})$, the space of all functions of bounded variation on $\mathbb{R}$, iff there exists a.e. derivative $f' \in L^1(\mathbb{R})$; and we define the total variation of $f \in BV(\mathbb{R})$ as

$$V_{\mathbb{R}}[f] := \|f'\|_1 = \int_{-\infty}^{\infty} |f'(v)|\, dv.$$

Next we give some statements and properties of $BV(\mathbb{R})$.

Proposition 2. *For* $f \in BV(\mathbb{R})$ *and any monotone increasing sequence* $\{t_k\}_{k=-\infty}^{\infty} \subset \mathbb{R}$ *with* $\lim_{k\to\pm\infty} t_k = \pm\infty$ *one has:*

1)

$$\sum_{k=-\infty}^{\infty} |f(t_k+v) - f(t_{k-1}+v)| \leqslant V_{\mathbb{R}}[f] \quad (v \in \mathbb{R});$$

2) f *is bounded and there exist* $\lim_{t\to\pm\infty} f(t)$;

3) f *is locally integrable, and*

$$\frac{1}{b-a}\sum_{k=-\infty}^{\infty}\int_a^b |f(t_k+v) - f(t_{k-1}+v)|dv \leqslant V_{\mathbb{R}}[f].$$

Proposition 3. *Let* $f' \in BV(\mathbb{R})$, *and let* $a \in \mathbb{R}$, $W > 0$, $t_k = \frac{k}{W}$, $k \in \mathbb{Z}$. *Then*

$$W\sum_{k=-\infty}^{\infty}\Big|f(t_k+a) - 2f(t_{k-1}+a) + f(t_{k-2}+a)\Big| \leq V_{\mathbb{R}}[f'].$$

III. The variation detracting property

The variation detracting property of certain sampling operators will be considered for $BV(\mathbb{R})$, the space of all functions of bounded variation on $\mathbb{R}$. This property is important in practice, since often signals are discontinuous but still with bounded variation.

Theorem 1. *Assume* $s_m \in L^1(\mathbb{R})$. *If there exists a number* $b \in \mathbb{R}$ *such that* $\pm m - b \in \mathbb{Z}$, *then for* $f \in BV(\mathbb{R})$ *we have* $S_W f \in BV(\mathbb{R})$ *and*

$$V_{\mathbb{R}}[S_W f] \leq \|s_m\|_1\, V_{\mathbb{R}}[f].$$

The proof of Theorem 1 was essentially presented in [7]. By Corollary 1 and Theorem 1 we obtain under assumptions of Theorem 1

Corollary 2.

$$V_{\mathbb{R}}[S_W f] \leq \|s_{m,n}\|_1\, V_{\mathbb{R}}[f].$$

For the proof of Theorem 2 we need the following technical lemma.

Lemma. *For any sequence* $(a_0, a_1, a_2, ...)$ *and* $M, N \in \mathbb{N}$ *we have*

$$\sum_{i=1}^{M}\sum_{j=1}^{N}\big(a_{i+j} - 2a_{i+j-1} + a_{i+j-2}\big) = a_0 - a_M - a_N + a_{M+N}.$$

The variation detracting property for derivatives will be read as follows.

Theorem 2. *Assume the kernel* $s_{m,n} \in L^1(\mathbb{R})$ *for* $m \neq 0, n \neq 0$, *such that there exists* $b \in \mathbb{R}$ *with* $\pm m \pm n - b \in \mathbb{Z}$. *If* f *is bounded and* $f' \in BV(\mathbb{R})$, *then* $(S_W f)' \in BV(\mathbb{R})$ *and*

$$V_{\mathbb{R}}[(S_W f)'] \leq \|s_{m,n}\|_1\, V_{\mathbb{R}}[f'].$$

Proof. By (7) we get (see also Remarks 1 and 2)

$$s'(t) = \frac{1}{2m}\big(s_m(t+m) - s_m(t-m)\big)$$

and hence by (9) we obtain

$$\begin{aligned} & s''(t) = \frac{1}{2m}\big(s'_m(t+m) - s'_m(t-m)\big) \\ = \;& \frac{1}{4mn}\int_{-n}^{n}\big(s'_{m,n}(t+m+y) - s'_{m,n}(t-m+y)\big)dy \\ = \;& \frac{1}{4mn}\big(s_{m,n}(t+m+n) - s_{m,n}(t+m-n) \\ - \;& s_{m,n}(t-m+n) + s_{m,n}(t-m-n)\big). \end{aligned}$$

So by (1) we have

$$\begin{aligned} (S_W f)''(t) &= W^2\sum_k f\big(\frac{k}{W}\big)s''(Wt-k) \\ &= \frac{W^2}{4mn}\sum_k f\big(\frac{k}{W}\big)\big(s_{m,n}(Wt-k+m+n) \\ &- s_{m,n}(Wt-k+m-n) \\ &- s_{m,n}(Wt-k-m+n) \\ &+ s_{m,n}(Wt-k-m-n)\big). \end{aligned}$$

Since $s_{m,n} \in L^1(\mathbb{R})$, then by (6) $s_{m,n} \in B^1_\pi \subset L^1(\mathbb{R})$ and for $s_{m,n}$ the condition (2) is satisfied by Nikolskii's inequality. Therefore, since f is bounded, the series here is absolutely and uniformly convergent. Under assumptions of Theorem 2 on a number $b \in \mathbb{R}$ we get

$$\begin{aligned} & (S_W f)''(t) \\ = \;& \frac{W^2}{4mn}\sum_k \Big(f\big(\frac{k+m+n-b}{W}\big) - f\big(\frac{k+m-n-b}{W}\big) \\ - \;& f\big(\frac{k-m+n-b}{W}\big) + f\big(\frac{k-m-n-b}{W}\big)\Big) \\ \times \;& s_{m,n}(Wt-k+b). \end{aligned} \tag{11}$$

The application of Lemma with $M = 2m, N = 2n$,

$$a_l = f\big(\frac{k+l-m-n-b}{W}\big),$$

$(k \in \mathbb{Z}, W > 0, l = 0, 1, ...)$ gives

$$\begin{aligned} & f\big(\frac{k-m-n-b}{W}\big) - f\big(\frac{k+m-n-b}{W}\big) \\ - \;& f\big(\frac{k-m+n-b}{W}\big) + f\big(\frac{k+m+n-b}{W}\big) \\ = \;& \sum_{i=1}^{2m}\sum_{j=1}^{2n}\Big(f\big(\frac{k+i+j-m-n-b}{W}\big) \\ - \;& 2f\big(\frac{k+i+j-1-m-n-b}{W}\big) \\ + \;& f\big(\frac{k+i+j-2-m-n-b}{W}\big)\Big). \end{aligned}$$

Now by (11) we have

$$\begin{aligned} (S_W f)''(t) = \frac{W^2}{4mn}\sum_k\sum_{i=1}^{2m}\sum_{j=1}^{2n}\Big(f\big(\frac{k+i+j-m-n-b}{W}\big) \\ - 2f\big(\frac{k+i+j-1-m-n-b}{W}\big) \\ + f\big(\frac{k+i+j-2-m-n-b}{W}\big)\Big)s_{m,n}(Wt-k+b). \end{aligned}$$

Estimating and integrating over $\mathbb{R}$ yields

$$\begin{aligned} \|(S_W f)''\|_1 \le \|s_{m,n}\|_1\frac{W}{4mn}\sum_{i=1}^{2m}\sum_{j=1}^{2n}\sum_k\Big|f\big(\frac{k}{W}+a_{ij}\big) \\ - 2f\big(\frac{k-1}{W}+a_{ij}\big) + f\big(\frac{k-2}{W}+a_{ij}\big)\Big|, \end{aligned}$$

where $a_{ij} = \frac{i+j-m-n-b}{W}$. The application of Proposition 3 yields

$$\begin{aligned} \|(S_W f)''\|_1 &\le \|s_{m,n}\|_1\frac{1}{4mn}\sum_{i=1}^{2m}\sum_{j=1}^{2n}V_{\mathbb{R}}[f'] \\ &= \|s_{m,n}\|\;V_{\mathbb{R}}[f']. \end{aligned}$$

This concludes the proof.

In analogous way we can consider the variation detracting property for the second derivative. We have the following

Theorem 3. *Assume the kernel $s_{m,n,r} \in L^1(\mathbb{R})$ for $0 < m, n, r < 1$ such that there exists $b \in \mathbb{R}$ with $\pm m \pm n \pm r - b \in \mathbb{Z}$. Here*

$$s_{m,n,r}(t) := \int_0^1 \frac{\lambda(u)}{\operatorname{sinc}(mu)\operatorname{sinc}(nu)\operatorname{sinc}(ru)}\cos(\pi t u)du.$$

If f is bounded and $f'' \in BV(\mathbb{R})$, then $(S_W f)'' \in BV(\mathbb{R})$ and

$$V_{\mathbb{R}}[(S_W f)''] \le \|s_{m,n,r}\|_1\;V_{\mathbb{R}}[f''].$$

IV. Applications

1) If we take $\lambda_L(u) = \operatorname{sinc} u$, the Lanczos' window function, then we get

$$\frac{\lambda_L(u)}{\operatorname{sinc}\frac{u}{2}} = \cos\frac{\pi u}{2} \equiv \lambda_R(u),$$

which is the Rogosinski window. Applying Theorem 1 with $m = b = \frac{1}{2}$, by Corollary 2 we obtain

$$V_{\mathbb{R}}[L_W f] \le \|r_0\|_1\;V_{\mathbb{R}}[f].$$

If we take $m = n = \frac{1}{2}$ in (6), we get

$$\frac{\lambda_L(u)}{\operatorname{sinc}^2\frac{u}{2}} = \frac{\pi u}{2}\cot\frac{\pi u}{2} \equiv \lambda_{F_1}(u),$$

which defines the Favard-type kernel s_{F_1} ([9], Section 4.1.3; [11], Chapter 3, Section 1). Applying Theorem 2 with $m = n = \frac{1}{2}$, $b = 0$, we obtain

$$V_{\mathbb{R}}[(L_W f)'] \le \|s_{F_1}\|_1\;V_{\mathbb{R}}[f'].$$

If we take $m = n = r = \frac{1}{2}$ in Th. 3, then we get

$$\frac{\lambda_L(u)}{\operatorname{sinc}^3 \frac{u}{2}} = \Big(\frac{\pi u}{2}\Big)^2 \frac{\cot \frac{\pi u}{2}}{\sin \frac{\pi u}{2}} \equiv \lambda_{F_2}(u),$$

which defines the Favard-type kernel s_{F_2} (see [11], Ch. 3, Sect. 1). Applying Theorem 3 with $m = n = r = b = \frac{1}{2}$, we obtain

$$V_{\mathbb{R}}[(L_W f)''] \le \|s_{F_2}\|_1 \; V_{\mathbb{R}}[f''].$$

Remark 3. In [8] we proved that $\|r_0\|_1 = \|L_W\|$, thus the variation detracting property takes a very natural shape,

$$V_{\mathbb{R}}[L_W f] \le \|L_W\| \; V_{\mathbb{R}}[f].$$

2) If we take $\lambda_R(u) = \cos \frac{\pi u}{2}$, the Rogosinski window function, we get

$$\frac{\lambda_R(u)}{\operatorname{sinc} \frac{u}{2}} = \lambda_{F_1}(u).$$

Applying Theorem 1 with $m = b = \frac{1}{2}$, we get

$$V_{\mathbb{R}}[R_W] \le \|s_{F_1}\|_1 \; V_{\mathbb{R}}[f].$$

If we take $m = n = \frac{1}{2}$ in (6), we again have the Favard-type window

$$\frac{\lambda_R(u)}{\operatorname{sinc}^2 \frac{u}{2}} = \Big(\frac{\pi u}{2}\Big)^2 \cdot \frac{\cot \frac{\pi u}{2}}{\sin \frac{\pi u}{2}} = \lambda_{F_2}(u).$$

Applying Theorem 2 with $m = n = \frac{1}{2},\ b = 0$, we have

$$V_{\mathbb{R}}[(R_W f)'] \le \|s_{F_2}\|_1 \; V_{\mathbb{R}}[f'].$$

3) Let $\lambda_H(u) = \cos^2 \frac{\pi u}{2}$ be the Hann window function, then

$$\frac{\lambda_H(u)}{\operatorname{sinc} u} = \frac{\pi u}{2} \cot \frac{\pi u}{2} = \lambda_{F_1}(u).$$

Applying Theorem 1 with $m = b = 1$ we get

$$V_{\mathbb{R}}[H_W f] \le \|s_{F_1}\|_1 \; V_{\mathbb{R}}[f].$$

If we take $m = 1,\ n = \frac{1}{2}$ in (6), we have

$$\frac{\lambda_H(u)}{\operatorname{sinc} u \cdot \operatorname{sinc} \frac{u}{2}} = \frac{\cos \frac{\pi u}{2}}{\operatorname{sinc}^2 \frac{u}{2}} = \lambda_{F_2}(u),$$

which defines the Favard-type kernel. Applying Theorem 2 with $m = 1,\ n = \frac{1}{2},\ b = \frac{1}{2}$, we obtain

$$V_{\mathbb{R}}[(H_W f)'] \le \|s_{F_2}\|_1 \; V_{\mathbb{R}}[f'].$$

V. Conclusion

We investigated the variation detracting property of the generalized Shannon sampling operators,

$$(S_W f)(t) := \sum_{k=-\infty}^{\infty} f(\frac{k}{W}) s(Wt - k),$$

which preserve the total variation of functions and their derivatives, i.e.

$$V_{\mathbb{R}}[(S_W f)^{(k)}] \le \|s_{m,n,r}\|_1 \; V_{\mathbb{R}}[f^{(k)}],$$

where $k = 0, 1, 2$ and $s_{m,n,r}$ are certain related kernels to the original kernel

$$s(t) := s(\lambda; t) := \int_0^1 \lambda(u) \cos(\pi t u).$$

As applications we considered some, in literature known examples of kernels, which realize the variation detracting property of the generalized Shannon sampling operators.

Acknowledgment

This research was partially supported by the Estonian Sci. Foundation, grant 8627 and by Estonian Center of Excellence "Mesosystems - Theory and Applications", AU/8211.

The authors are grateful to their colleague G. Tamberg (Tallinn University of Technology) for some ideas concerning the kernels s_m.

References

[1] J. A. Adell and J. de la Cal, Bernstein-type operators diminish the φ-variation, *Constr. Approx.*, 12: 489-507, 1996.

[2] H. H. Albrecht, A family of cosine-sum windows for high resolution measurments. In *IEEE International Conference on Acoustics, Speech and Signal Processing, Salt Lake City, May 2001,* pages 3081-3084. Salt Lake City, 2001.

[3] C. Bardaro, P. L. Butzer, R. L. Stens and G. Vinti, Convergence in variation and rates of approximation for Bernstein-type polynomials and singular convolution integrals, *Analysis (Munich)*, 23: 299–346, 2003.

[4] P. L. Butzer, W. Splettstösser and R. L. Stens, The sampling theorems and linear prediction in signal analysis. *Jahresber.Deutsch. Math-Verein,* 90:1-70, 1988.

[5] T. N. T. Goodman, Variation diminishing properties of Bernstein polynomials on triangles, *J. Approx. Theory*, 50: 111–126, 1987.

[6] J. R. Higgins, *Sampling Theory in Fourier and Signal Analysis*, Clarendon Press, Oxford, 1996.

[7] A. Kivinukk, On some Shannon sampling series with the variation detracting property, *Proc. of the 9th Intern. Conf. on Sampling Theory and Applications* , Singapore, May 2-6, 2011, A. Khong, F. Oggier (Eds.), Nanyang Technological University, Singapore 2011, 1 – 4

[8] A. Kivinukk and G. Tamberg, On sampling series based on some combinations of sinc fubctions, *Proc. of the Estonian Academy of Sciences. Physics Mathematics*, 51: 203-220, 2002.

[9] N. P. Korneichuk, *Exact Constants in Approximation Theory.* Cambridge Univ. Press, 1991.

[10] H. D. Meikle, *A New Twist to Fourier Transforms.* Dover, Berlin, 2004.

[11] V. V. Zhuk, *Approximation of Periodic Functions. (Russ.)* Leningrad Univ. Press, Leningrad, 1982.

Jointly filtering and regularizing seismic data using space-varying FIR filters

Apostolos Kontakis†, Xander H. Campman†, Geert J. T. Leus*, Zijian Tang† and Mike Danilouchkine†
*Circuits and Systems group, Faculty of EEMCS, Delft University of Technology
Address: Mekelweg 4, 2628 CD, Delft, The Netherlands
Email: G.J.T.Leus@tudelft.nl
†Novel Geophysical Measurements (PTI/EN), Shell Global Solutions International B.V.
Address: Kessler Park 1, 2288 GS Rijswijk, The Netherlands
Email: A.Kontakis@student.tudelft.nl, {Xander.Campman, Zijian.Tang, Mike.Danilouchkine}@shell.com

***Abstract*—Array forming in seismic data acquisition can be likened to FIR filtering. Misplacement of the receivers used to record seismic waves can lead to degraded performance with respect to the filtering characteristics of the array. We propose two methods for generating linear space-varying filters that take receiver misplacements into account and demonstrate their performance on synthetic data.**

I. Introduction

Variations in the sampling interval when sampling a signal are often difficult or impossible to prevent. This might make processing the sampled signal problematic, if the related tools can only handle uniformly sampled signals. The motivation for our work comes from the field of seismic data acquisition. A very common practice in this field, is to sum together the (possibly weighted) output signal of multiple seismic receivers. This process is known as array forming and is used to improve the signal-to-noise ratio and to reduce the amount of recorded data.

Being a weighted summation of the output of a finite number of receivers, array forming can be likened to FIR filtering. More specifically, we can view array forming as the application of a linear space-invariant (LSI) filter, since usually the same set of weights is used for the array elements of all arrays in a field.

A problem arises, however, when the receivers are misplaced due to e.g., terrain difficulties. Usually the array weights are designed for a specific geometrical layout of the array elements. When this geometrical layout is violated, it can prove detrimental for the filtering capabilities of the array, as shown in [1]. Fortunately, advances in acquisition hardware have enabled us a) to record the output of each individual receiver and b) to know with high (but limited) accuracy the actual location of each receiver. This makes more sophisticated techniques viable for array forming/filtering that can compensate for irregularities in sampling.

A number of solutions have been suggested for the problem of array forming/filtering nonuniformly sampled data, which can be roughly divided in three categories. Methods of the first category interpolate an FIR filter that is designed to filter uniformly sampled data. We shall call this FIR filter the prototype filter. An example is given in [2], where the prototype filter is interpolated to the actual locations of the receivers. These interpolated filter coefficients (or equivalently, the array weights) are then reweighted based on the sampling density at the area around each receiver. We shall refer to this method as geometry-compensating filtering (GCF). The second category involves methods that approximate the outputs of the prototype filter applied to the uniformly sampled data. An example is given in [3], which uses the projections onto convex sets (POCS) framework applied to nonuniformly sampled data. The third category involves reconstruction of the data at the regular sampling locations. The prototype filter can then be applied to the reconstructed data. The methods given in [4], [5] are examples of data reconstruction.

The goal of this work is to propose two methods that generate a linear space-varying (LSV) FIR filter suitable for filtering the nonuniformly sampled data. We will refer to these methods as **Method A** and **Method B**. The LSV FIR filter designed by any of these two methods generates the filter output at equi-spaced intervals. In this respect, both methods combine filtering and regularization in one operator. The difference is that

- **Method A** approximates the prototype filter in the spatial domain, which has been already designed,
- **Method B** approximates the *ideal* response of the prototype filter in the wavenumber domain[1]. In other words, **Method B** also skips the intermediate step of designing the prototype filter.

Compared to existing works, **Method A** has a similar flavor as the approach of [2] since they both interpolate a prototype filter to the actual locations of the receivers. However, the interpolation in [2] is driven by the geometry of the receiver arrays while in this paper, the interpolation is based on the band-limited assumption on the received signal, which is also utilized in [4], [5]. In reality, such a band-limited assumption is often valid, and the corresponding interpolation process can yield a better approximation to the prototype filter response in the wavenumber domain as will be demonstrated in the paper.

[1]Wavenumber domain is also known as spatial frequency domain.

II. Proposed Algorithms

For simplicity, we only consider filtering along one spatial dimension x. The continuous data is represented by $d(x)$; the data samples gathered at the N locations of the receivers $x_j,\ j = 0, \ldots, N-1$ are denoted by $d(x_j)$. The nominal locations of the receivers are defined on the grid $\bar{x}_n = n\Delta x$ while the actual locations of the receivers are assumed to lie on the denser grid $\bar{\bar{x}}_m = m\delta x = m(\Delta x/M)$ with M being a positive integer. This is not overly restrictive, since M can be large and the precise receiver locations are known with high, but limited, accuracy. We also define the indicator function $s(\bar{\bar{x}}_m)$ to take the value $s(\bar{\bar{x}}_m) = 1$ when a receiver is present at $\bar{\bar{x}}_m$ and the value zero otherwise. Suppose a prototype LSI FIR filter has already been defined on the nominal grid, whose ith tap is denoted as $h(\bar{x}_i)$. We assume that $h(\bar{x}_i) = 0$ if $\bar{x}_i < 0$ or $\bar{x}_i \geq L_f\Delta x$, where L_f is referred to as the spatial support of the FIR filter.

A. Method A

The filter to be designed in **Method A** is represented as a set of LSV FIR filters, whose filter taps $g_l(\bar{\bar{x}}_m)$ for $m = 0, 1, \ldots, NM-1$ are defined on the dense grid. The spatial support of $g_l(\bar{\bar{x}}_m)$ is the same as that of $h(\bar{x}_i)$, therefore $g_l(\bar{\bar{x}}_m) = 0$ if $\bar{\bar{x}}_m < 0$ or $\bar{\bar{x}}_m \geq L_f\Delta x$ (note that [4] imposes a different FIR constraint on the filter). We desire that the output of one such filter at output location $\check{x}_l$ should be identical as if the prototype filter were applied on uniformly sampled data. In other words,

$$\sum_{n=0}^{N-1} h(\check{x}_l - \bar{x}_n)d(\bar{x}_n) = \sum_{m=0}^{NM-1} g_l(\check{x}_l - \bar{\bar{x}}_m)s(\bar{\bar{x}}_m)d(\bar{\bar{x}}_m). \quad (1)$$

The output locations $\check{x}_l$ lie on the nominal grid, i.e., $\check{x}_l = l\Delta x = lM\delta x$. Utilizing the band-limited assumption on $d(\bar{\bar{x}}_m)$ it can be shown that

$$d(\bar{\bar{x}}_m) \approx \frac{1}{N}\sum_{p=-P}^{P}\left(\sum_{n=0}^{N-1} d(\bar{x}_n)e^{-j\frac{2\pi p}{N\Delta x}\bar{x}_n}\right)e^{j\frac{2\pi p}{N\Delta x}\bar{\bar{x}}_m}, \quad (2)$$

where $N = 2P+1$ (a similar expression can be derived when N is even). We can exchange the order of summation and arrive at

$$\begin{aligned} d(\bar{\bar{x}}_m) &\approx \sum_{n=0}^{N-1}\left(\frac{1}{N}\sum_{p=-P}^{P} e^{j\frac{2\pi p}{N\Delta x}(\bar{\bar{x}}_m - \bar{x}_n)}\right)d(\bar{x}_n), \\ &\approx \sum_{n=0}^{N-1}\underbrace{\frac{\sin(\frac{\pi}{\Delta x}(\bar{\bar{x}}_m - \bar{x}_n))}{N\sin(\frac{\pi}{N\Delta x}(\bar{\bar{x}}_m - \bar{x}_n))}}_{\mathrm{sincd}(N;\bar{x}_n,\bar{\bar{x}}_m)} d(\bar{x}_n). \end{aligned} \quad (3)$$

Substituting (3) in (1) yields

$$\begin{aligned} &\sum_{n=0}^{N-1} h(\check{x}_l - \bar{x}_n)d(\bar{x}_n) \approx \\ &\sum_{m=0}^{NM-1} g_l(\check{x}_l - \bar{\bar{x}}_m)s(\bar{\bar{x}}_m)\left(\sum_{o=0}^{N-1}\mathrm{sincd}(N;\bar{x}_o,\bar{\bar{x}}_m)d(\bar{x}_o)\right). \end{aligned} \quad (4)$$

We rewrite (4) in matrix-vector form as

$$\mathbf{h}_l^H\mathbf{d} \approx \mathbf{g}_l^H\mathbf{S}\mathbf{Q}\mathbf{d} \quad (5)$$

where

$\mathbf{h}_l$ is an $N\times 1$ vector with $h(\check{x}_l - \bar{x}_n)$ as its n-th element;
$\mathbf{g}_l$ is an $NM \times 1$ vector with $g_l(\check{x}_l - \bar{\bar{x}}_m)$ as its m-th element;
$\mathbf{S}$ is an $NM \times NM$ diagonal matrix with $s(\bar{\bar{x}}_m)$ as its m-th diagonal element;
$\mathbf{Q}$ is an $NM \times N$ matrix with $\mathrm{sincd}(N;\bar{x}_o,\bar{\bar{x}}_m)$ as its (m,o)th element and
$\mathbf{d}$ is an $N \times 1$ vector with $d(\bar{x}_n)$ as its n-th element.

A sufficient condition for (5) to hold is

$$\mathbf{h}_l^H \approx \mathbf{g}_l^H\mathbf{S}\mathbf{Q},$$

for which a suitable $\mathbf{g}_l$ can be found by solving the following least-squares problem

$$\min_{\mathbf{g}_l}\left\{||\mathbf{h}_l^H - \mathbf{g}_l^H\mathbf{S}\mathbf{Q}||_2^2\right\} \equiv \min_{\tilde{\mathbf{g}}_l}\left\{||\mathbf{h}_l^H - \tilde{\mathbf{g}}_l^H\tilde{\mathbf{Q}}||_2^2\right\}, \quad (6)$$

where $\tilde{\mathbf{g}}_l^H$ is formed by removing its elements corresponding to the zeros of $\mathbf{S}$. Similarly, $\tilde{\mathbf{Q}}$ is constructed after removing the rows of $\mathbf{Q}$ corresponding to the zero columns in $\mathbf{S}$. This eliminates $\mathbf{S}$ from (6). In order to limit the spatial support of $\mathbf{g}_l$, we remove the elements of $\tilde{\mathbf{g}}_l$ that correspond to elements m of $\mathbf{g}_l$ for which $m < lM$ or $m \geq (l+L_f)M$ holds, thus forming $\tilde{\tilde{\mathbf{g}}}_l$. The corresponding rows of $\tilde{\mathbf{Q}}$ are removed as well, to form $\tilde{\tilde{\mathbf{Q}}}$. The problem has a closed-form solution given by

$$\tilde{\tilde{\mathbf{g}}}_l^H = \mathbf{h}_l^H\tilde{\tilde{\mathbf{Q}}}^H(\tilde{\tilde{\mathbf{Q}}}\tilde{\tilde{\mathbf{Q}}}^H)^{-1}. \quad (7)$$

A different FIR filter $\mathbf{g}_l$ has to be calculated for each output location $\check{x}_l$. The solution can be seen as a composition of two operations: the first operation is $\mathbf{h}_l^H\tilde{\tilde{\mathbf{Q}}}^H = (\tilde{\tilde{\mathbf{Q}}}\mathbf{h}_l)^H$, which can be interpreted as an interpolation of the filter to the actual locations of the receivers. The second operation $(\tilde{\tilde{\mathbf{Q}}}\tilde{\tilde{\mathbf{Q}}}^H)^{-1}$ deconvolves the effects of nonuniform sampling.

B. Method B

In comparison to **Method A**, **Method B** does not rely on the knowledge of the prototype filter $\mathbf{h}_l$, which needs to be pre-designed. To this end, we will try to approximate the behavior of the prototype filter in the wavenumber domain. As a first step, let us use a circular convolution operator to describe the target LSV FIR filter as well as the prototype filter. Accordingly, (1) should be adapted to the form

$$\begin{aligned} &\sum_{n=0}^{N-1} h(\bar{x}_{(l-n)\bmod(N)})d(\bar{x}_n) = \\ &\sum_{m=0}^{NM-1} g_l'(\bar{\bar{x}}_{(lM-m)\bmod(NM)})s(\bar{\bar{x}}_m)d(\bar{\bar{x}}_m), \end{aligned} \quad (8)$$

where $g_l'(\bar{\bar{x}}_m)$ stands for the LSV FIR filter from **Method B**. In deriving the above, we have used the assumption that the output locations $\check{x}_l$ lie on the nominal grid, i.e., $\check{x}_l = l\Delta x = lM\delta x$. Just like **Method A**, we require that $g_l'(\bar{\bar{x}}_m)$ have a

spatial support in the interval $[0,\ L_f\Delta x)$. In other words, $g_l'(\bar{\bar{x}}_m) = 0$ if $\bar{\bar{x}}_m < 0$ or $\bar{\bar{x}}_m \geq L_f\Delta x$.

To get rid of the dependence on $h(\bar{x}_i)$ in (8), we resort to the wavenumber domain. Suppose the wavenumber response of the prototype filter is known as $h_w(k)$ with $-\pi \leq k \leq \pi$. If we use $\mathbf{H}_w$ to denote an $N \times N$ diagonal matrix whose nth diagonal element is given by the sample $h_w(k)$ at $k = -\pi + n\frac{2\pi}{N}$ for $n = 0, 1, \ldots, N-1$, then the wavenumber-domain counterpart of (8) can be expressed as

$$\mathbf{H}_w(\mathbf{Fd}) = \mathbf{FG'SQF}^H(\mathbf{Fd}), \tag{9}$$

where we have coined an $N \times NM$ matrix $\mathbf{G'}$ to represent the circular convolution of the filter with the (l,m)th element of $\mathbf{G'}$ given by $g_l'(\bar{\bar{x}}_{(lM-m)\bmod(NM)})$; $\mathbf{F}$ stands for the N-point DFT matrix, and $\mathbf{S}$ and $\mathbf{Q}$ are defined in (5). A sufficient condition for (9) to hold is

$$\mathbf{H}_w \approx \mathbf{FG'SQF}^H. \tag{10}$$

The right-hand side of (10) can be broken down to three parts:

- $\mathbf{G'}$ is the LSV filter. Its output is defined on the nominal grid and its input on the dense grid.
- $\mathbf{F}$ acts on the columns of $\mathbf{G'}$ to produce the DFT of the LSV filter $g_l(\check{x}_l - \bar{\bar{x}}_m)$ with respect to $\check{x}_l$.
- $\mathbf{SQF}^H$ acts of the rows of $\mathbf{G'}$ to produce the nonuniform DFT of $g_l(\check{x}_l - \bar{\bar{x}}_m)$ with respect to $\bar{\bar{x}}_m$.

The product of $\mathbf{FG'SQF}^H$ is a wavenumber connection matrix[2] [6], which can be viewed as the wavenumber response of the LSV FIR filter. Note that the off-diagonal elements in $\mathbf{FG'SQF}^H$ are in most cases non-zero, which means that the relation between the spectra of the data before and after filtering in the wavenumber domain is not simply an element-wise multiplication for nonuniform sampling.

From (10), we formulate the following least-squares problem

$$\min_{\mathbf{G'}}\{||\mathbf{W} \odot (\mathbf{H}_w - \mathbf{FG'SQF}^H)||_F^2 \tag{11}$$

where $\mathbf{W}$ is a weighting matrix that can apply an individual weight to each element of $\mathbf{H}_w - \mathbf{FG'SQF}^H$ with the symbol $\odot$ denoting the Hadamard (element-wise) product. Including such a weighting matrix is beneficial, for instance, to enable a better trade-off between different approximation errors in the passband, stopband as well as the "do-not-care" (transition) zones.

As in **Method A**, the matrix $\mathbf{S}$ can be eliminated by removing all the columns of $\mathbf{G'}$ and rows of $\mathbf{Q}$ with the same index as the elements of the diagonal of $\mathbf{S}$ that have the value zero. Let $\tilde{\mathbf{G}}'$ and $\tilde{\mathbf{Q}}$ respectively be the reduced form of $\mathbf{G'}$ and $\mathbf{Q}$ resulting from this column- and row-removal. Then (11) can be rewritten as

$$\min_{\mathbf{G'}}\{||\mathbf{W} \odot (\mathbf{H}_w - \mathbf{F}\tilde{\mathbf{G}}'\tilde{\mathbf{Q}}\mathbf{F}^H)||\}_F^2 \tag{12}$$

[2] In [6] its continuous counterpart is called "frequency connection function". We use the term "wavenumber connection matrix" for consistency with the rest of the terminology in this paper.

This problem can be restated in its vectorized form as

$$\min_{vec(\mathbf{G'})}\{||\ vec(\mathbf{W} \odot (\mathbf{H}_w - \mathbf{F}\tilde{\mathbf{G}}'\tilde{\mathbf{Q}}\mathbf{F}^H)||\}_2^2 \equiv$$
$$\min_{vec(\mathbf{G'})}\{||\ diag(vec(\mathbf{W}))\ vec(\mathbf{H}_w - \mathbf{F}\tilde{\mathbf{G}}'\tilde{\mathbf{Q}}\mathbf{F}^H)||\}_2^2, \tag{13}$$

Here $vec(\mathbf{A})$ returns a column-vector that stacks the columns of the matrix $\mathbf{A}$. The function $diag(\mathbf{v})$ returns a diagonal matrix with the vector $\mathbf{v}$ on its main diagonal.

Using the identity $vec(\mathbf{ABC}) = (\mathbf{C}^T \otimes \mathbf{A})\ vec(\mathbf{B})$, where $\otimes$ denotes the Kronecker product, (13) can be rewritten as

$$\min_{vec(\mathbf{G'})}\{||\ diag(vec(\mathbf{W}))\ vec(\mathbf{H}_w) -$$
$$diag(vec(\mathbf{W}))(\mathbf{F}^*\tilde{\mathbf{Q}}^T \otimes \mathbf{F})\ vec(\tilde{\mathbf{G}}))||\}_2^2$$

where * denotes the complex conjugate of a matrix. Let

$$\tilde{\mathbf{U}} = diag(vec(\mathbf{W}))(\mathbf{F}^*\tilde{\mathbf{Q}}^T \otimes \mathbf{F})$$
$$\tilde{\mathbf{g}}' = vec(\tilde{\mathbf{G}}').$$

The elements of $\tilde{\mathbf{g}}'$ and rows of $\tilde{\mathbf{U}}$ that should be removed due to the limited spatial support of the filter $g_l'(\bar{\bar{x}}_n)$ are given by the indexes of those elements of $\tilde{\mathbf{g}}'$ that correspond to the zero elements of $g_l'(\bar{\bar{x}}_m)$. If we call $\tilde{\tilde{\mathbf{g}}}'$ and $\tilde{\tilde{\mathbf{U}}}$ the results after the corresponding row and element removal, the solution is given by

$$\tilde{\tilde{\mathbf{g}}}' = (\tilde{\tilde{\mathbf{U}}}^H\tilde{\tilde{\mathbf{U}}})^{-1}\tilde{\tilde{\mathbf{U}}}^H\ diag(vec(\mathbf{W}))\ vec(\mathbf{H}_w) \tag{14}$$

$\mathbf{G'}$ can be constructed from the elements of $\tilde{\tilde{\mathbf{g}}}'$ and can be applied to the nonuniformly sampled data.

III. Results

The performance of **Method A** and **Method B** was examined using synthetically created seismic data. A portion of the spectral content of the synthesized data can be found at the lower part of Fig. 1. The reflections from the Earth's subsurface appear mostly on the lower part of the wavenumber spectrum in the range $-0.02m^{-1} \leq k' \leq 0.02m^{-1}$. This region appears highlighted in all figures. All the plots have been smoothed by a 5-tap moving average filter. The peaks found at $k' = \pm 0.055m^{-1}$ are due to waves propagating along the Earth's surface. Their presence is often undesired and should be removed before further processing of the seismic data. Notice that when the data are not uniformly sampled, the wavenumber content appears smeared. This is a well-known side-effect introduced by nonuniform sampling [7].

We generated 100 different realizations of the receiver locations and filtered the nonuniformly sampled data using **Methods A** and **B**. **Method A** approximates the prototype filter whose wavenumber response is given in the upper part of Figure 1. In **Method B**, the prototype filter is not given but is supposed to have a similar passband and stopband region in the wavenumber domain as in **Method A**. The LSV FIR filters resulting form **Method A** and **B** have the same spatial support as that of the prototype filter. The average spectrum of the 100

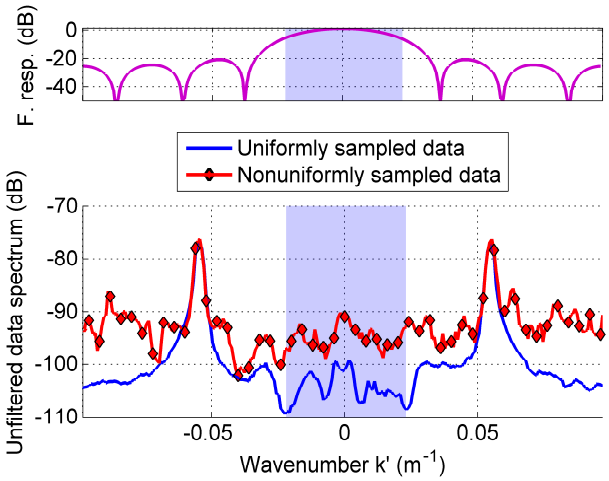


Fig. 1. Upper part: the wavenumber response of the prototype filter. Lower part: the wavenumber content for the data.

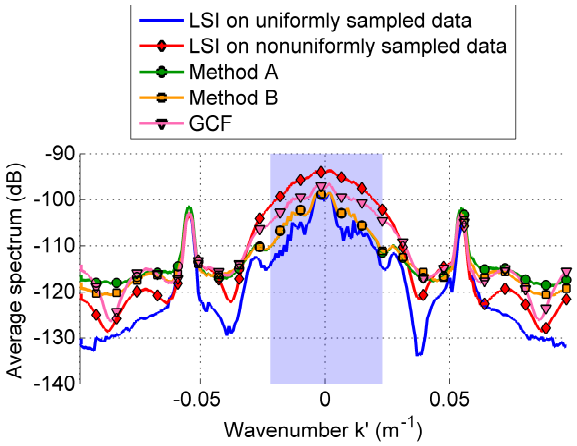


Fig. 2. Average spectrum of the filtered data (over 100 realizations).

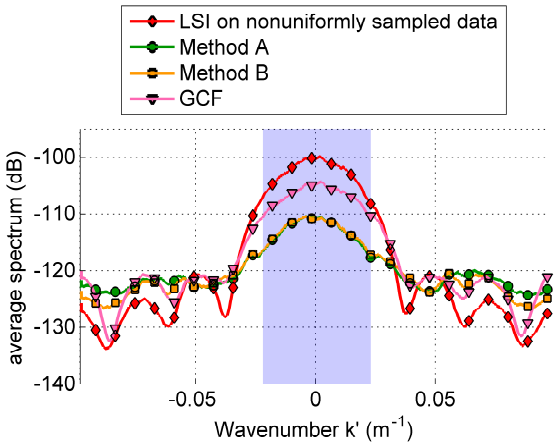


Fig. 3. Standard deviation from the average spectrum (over 100 realizations).

realizations can be seen in Fig. 2. In addition to **Methods A** and **B**, an adapted version of the GCF was implemented[3].

In addition, we also compared the proposed methods against 1) the ideal case where the prototype filter is applied on the uniformly sampled and 2) the case where the prototype filter is applied directly on nonuniform data. In the latter case, the spectral leakage induced by the irregularities is most obvious in the passband. In Fig. 2 we see that using the LSI filter on nonuniformly sampled data gives an output that differs almost 15dB at the edges of the passband. In contrast, **Methods A** and **B** give a filtered output that, on average, is closer to the ideal case in the low wavenumbers, exhibiting less than 5dB maximum difference from the uniformly sampled data case (Fig. 2) in the passband region. This is due to the fact that **Methods A** and **B** compensate for irregularities in sampling. The attenuation in the stopband, however, is less when using **Methods A** and **B**. This is because the FIR filters generated for each individual output will not, in general, have zeros at the same locations of their wavenumber responses. This leads to a more flat response in the stopband. GCF is somewhere in the middle, as it interpolates the filter to the locations of the receivers and compensates for sampling density, but does not deconvolve the effects of nonuniform sampling.

The standard deviation of the output spectrum can be seen in Fig. 3. **Methods A** and **B** exhibit in this case a significantly smaller standard deviation in the lower wavenumbers, for example, in Fig. 3, around 10dB lower than using the LSI filter. This means that **Methods A** and **B** may provide an output that is, generally, stable.

IV. Conclusion

We proposed two methods for generating LSV FIR filters suitable for filtering nonuniformly sampled data. The resulting filters yield a more accurate output in the passband than simply applying an LSI filter directly to nonuniformly sampled data. The output is also more stable, as it varies less for different realizations of the receiver locations.

[3]The method in [2] works on data having two spatial dimensions and is adapted here to the single spatial dimension case.

References

[1] P. Newman and J. Mahoney, "Patterns - with a pinch of salt," *Geophysical Prospecting*, vol. 21, no. 2, pp. 197–219, 1973.

[2] A. Özbek and R. Ferber, "Multi-dimensional filtering of seismic data sampled on an irregular grid," in *67th EAGE Conference & Exhibition*. EAGE, 2005.

[3] K. Kose and A. Cetin, "Low-pass filtering of irregularly sampled signals using a set theoretic framework [lecture notes]," *Signal Processing Magazine, IEEE*, vol. 28, no. 4, pp. 117–121, 2011.

[4] H. Johansson and P. Löwenborg, "Reconstruction of nonuniformly sampled bandlimited signals by means of time-varying discrete-time fir filters," *EURASIP Journal on Advances in Signal Processing*, vol. 2006, 2006.

[5] A. Duijndam, M. Schonewille, and C. Hindriks, "Reconstruction of band-limited signals, irregularly sampled along one spatial direction," *Geophysics*, vol. 64, no. 2, pp. 524–538, 1999.

[6] G. Margrave, "Theory of nonstationary linear filtering in the fourier domain," in *1997 SEG Annual Meeting*, 1997.

[7] G. Blacquière and L. Ongkiehong, "Single sensor recording: Anti-alias filtering, perturbations and dynamic range," in *2000 SEG Annual Meeting*, 2000.

Non-uniform Sampling Pattern Recognition Based on Atomic Decomposition

Tugdual Le Pelleter*, Taha Beyrouthy*, Robin Rolland†, Agnès Bonvilain*, Laurent Fesquet*†
*TIMA Laboratory
46, avenue Félix Viallet, 38031 GRENOBLE Cedex, France
Email: {tugdual.le-pelleter, taha.beyrouthy, agnes.bonvilain, laurent.fesquet}@imag.fr
†CIME Nanotech
3 parvis Louis Néel, BP 257, 38016 GRENOBLE Cedex, France
Email: robin.rolland-girod@grenoble-inp.fr

***Abstract*—Non-uniform sampling is an interesting scheme that can outperforms the uniform sampling with low activity signals. With such signals, it generates fewer samples, which means less data to process and lower power consumption. In addition, it is well-known that asynchronous logic is a low power technology. This paper deals with the coupling between a non-uniform sampling scheme and a pattern recognition algorithm implemented with an event-driven logic. This non-uniform analog-to-digital conversion and the specific processing have been implemented on an Altera FPGA platform. This paper reports the first results of this low-activity pattern recognition system and its ability to recognize specific patterns with very few samples. The objectives of this work target the future ultra-low power integrated systems.**

I. Introduction

The advances in microelectronics and wireless communication has facilitated the development of tiny sensor platforms, smart sensors that can be integrated in mobile devices. Currently, mobile communication devices have sophisticated internal hardware architectures and embed a wide range of internal sensors including three-axial accelerometers. Accelerometers are widely used in the context of health monitoring with the detection of motions, actions and activity.

The pattern recognition processing can be very challenging to be achieved on mobile devices because they have limited processing, memory and computing ressources. Figo et al. [1] evaluated preprocessing techniques for recognizing basic daily physical activities (jumping, running and walking) with accelerometric data performed on a mobile device. The battery autonomy that supplies the mobile device is an inherent limitation that depends on the computational costs, storage requirements and precision. Energy scavenging techniques such as thermoelectric effect, vibrations or body movements may also complement the battery power [2].

In activity monitoring, important amount of energy can be saved since activities occur sporadically. Efforts are made to avoid useless processing. For example, Jafari and Lotfian [3] propose a low-power architecture dedicated to the Dynamic Time Warping (DTW) for physical movement monitoring. The idea is to activate the processing unit that performs the pattern recognition only when necessary. The architecture consists of a granular decision making module (GDMM) that detects *a priori* relevant information in the signal.

Traditional activity recognition methods use uniform sampling scheme, however the non-uniform sampling defined by level crossing is a better candidate for signals with infrequent activity. In this approach, levels are disposed along the amplitude range of the signal and a sample is only captured when the input signal crosses one of the defined levels (cf. Fig.1.). Relation can be made with the field of symbolic dynamics where the data space is partitioned and associated to a symbol. Such method leads to data compression and is a way to save energy, because irrelevant samples are pruned off. The data partitioning is an active area of research. Some partitioning methods are the data mean, midpoint, median, equal-size intervals over the data range, or regions of the range with equal probability [4].

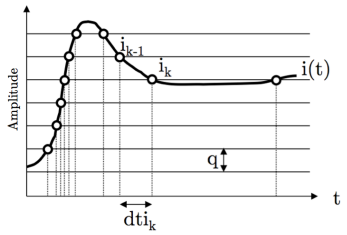


Fig. 1. Level crossing sampling with regularly spaced levels along the amplitude range of the signal

In this work, we propose a pattern recognition algorithm dedicated to mobile applications.

This paper is divided into three parts. We begin with discussion of the current state of pattern recognition methods for mobile devices. Secondly, we present our low-power pattern recognition method. In the third part, we present the hardware implementation. Finally, a discussion and a conclusion outline the method strength and weakness.

II. Related works

Benbasat and Paradiso [5] suggest to use the peaks structures in the accelerometer signal as a way to simplify the human gestures signature detection in inertial gesture recognition framework. The recognition is performed with simple

parameters : peaks magnitude, duration and number. This method, called *atomic gestures* decomposes complex human gesture into a set of peaks. For example, a straight-line motion generated by an arm movement will create a two-peaked trace. The inertial measures are obtained with two two-axis MEMS accelerometers and leading to a six degree-of-freedom inertial measurements. The concept of motion decomposition into atoms is used by Fan et al. [6] for the detection of translational and rotational motions using a tri-axial accelerometer. An interesting characteristic pointed out by the authors is that for translational motions the translation starts by a tangential acceleration followed by a tangential deceleration. As a consequence, the acceleration signature consists of two peaks with opposite signs. The specific pattern allows to perform the motion decomposition on the human gestures in order to generate the sequences of motion directions defining the translational motions (clockwise circular, counter clockwise circular and hop right). One of the advantage of the detection method based on atomic decomposition is its insensibility to time distortion. Indeed, the gestures vary between individuals and for the same individual.

A gesture can be performed at several speeds, leading to similar shaped patterns but with different duration. The pattern matching performed using the euclidean metric leads to poor results for highly time distorted patterns since it does not take into account the distortion due to its linear alignment. A more adapted method is the dynamic time warping (DTW) distance introduced by Sakoe and Chiba [7] in the context of speech recognition. It is widely used in speech recognition, for online signature verification and for fall detection in the elderly. This technique aims to find an optimal alignment between two given sequences. The non linear complexity in $O(N \cdot M)$ strongly restricts its implementation on mobile devices. Global constraints such as *Sakoe-Chiba band* [7] or adapted constraints [8] are used to linearize the computational cost. Jafari and Lotfian [3] implemented a low-power architecture of pattern recognition for mobile applications based on the DTW.

Many algorithms have been proposed for pattern recognition framework. The most popular method is the hidden Markov model (HMM). It can achieves high recognition rate in motion and gestures recognition. For example, Joselli and Clua [9] propose a detection method of patterns performed in the air for games processed on a mobile phone. The method is energy intensive so the implementation on mobile devices where battery cannot be easily recharged is restricted. In this work, we use a finite state machine (FSM) model used for its simplicity in hardware implementation. For more details about gesture recognition algorithms, we refer the reader to [10].

III. Tools & Methods

The proposed method is designed to perform pattern recognition in activity signals. To take into account their sporadic characteristic, a non classical sampling scheme called level crossing sampling (LC) is used. With the LC sampling, a sample is only captured when the continuous time input signal crosses one of the defined levels (cf. Fig. 1.). Contrary to the uniform sampling, the samples are not uniformly spaced along the time axis, because they depend on the signal variations. The time elapsed between samples i_k and i_{k-1} is defined by $dt_{i_k} = t_{i_n} - t_{i_{n-1}}$. For the hardware implementation of the LC sampling, a local timer of period T_c is dedicated to record dt_{i_k}. Contrary to the Nyquist sampling, the amplitude of the sample is known and the time elapsed between two samples is quantized. The Signal to Noise Ratio (SNR) depends on the timer period T_c and not on the number of quantization levels [11]. An empirical framework is proposed in this paper, to choose the minimum number of necessary levels with their respective amplitude to properly perform the pattern recognition.

A. Atomic decomposition

The manipulated data are time series. A time series $T = t_1, ..., t_m$ is an ordered set of m real-valued variables. The patterns of interest are considered as the local subsections of the time series, called subsequences. A subsequence C of length m is a sampling of length $n \leq m$ of contiguous positions from p, that is, $C = t_p, ..., t_{p+n-1}$ for $1 \leq p \leq m-n+1$.

In the proposed framework, the pattern recognition can be considered as a two-class classification problem, let {p,n} the two classes. A known process generates two different patterns to detect, that belongs to the positive class p. The subsequences generated by unknown processes are considered as noise and associated to the negative class n. The two different type of patterns of interest belonging to class p are depicted in Fig. 2. There are four types of variations to detect

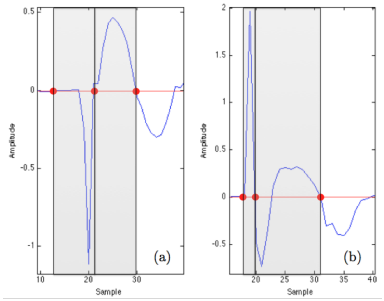


Fig. 2. The patterns consists of at least 2 peaks with opposite sign : (a) Negative peak P_1^- followed by positive peak P_2^+ (b) Positive peak P_1^+ followed by negative peak P_2^-

We suggest to detect these pair of opposite polarity peaks with two levels (positive and negative). Their respective amplitude is determined with the ROC curve.

B. Data labelization

The data labelization (cf. Fig. 3), consists of 3 steps : (1) The time series T is non-uniformly sampled with regularly spaced levels along the amplitude range. (2) The signal is then divided into two parts : positive T^+ and negative T^-. (3) Every peak bounded by contiguous zeros crossings, called epoch, are labeled. When the peaks belong to the positive (resp. negative) class, it is labeled 1 (resp. 0).

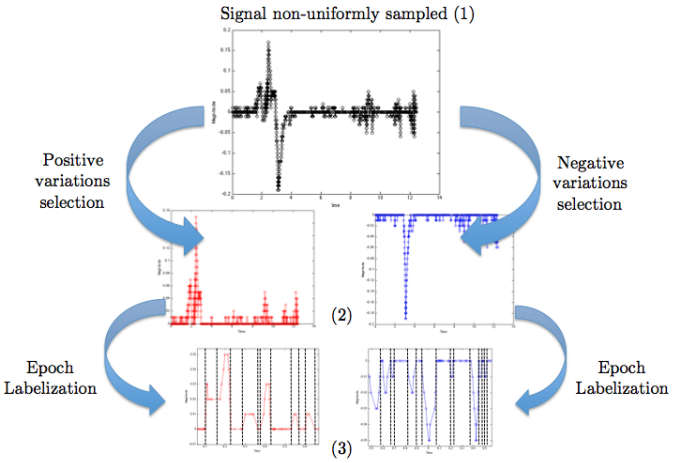


Fig. 3. Data labelization framework

C. ROC curve

The ROC curve is a technique mainly used for selecting a classifier in pattern recognition and is well-known in the medical decision community. Considering a two-class classification problem, each pattern of interest or instances are mapped to one element of the set $\{p, n\}$ of positive and negative class labels. There are four possible outcomes.

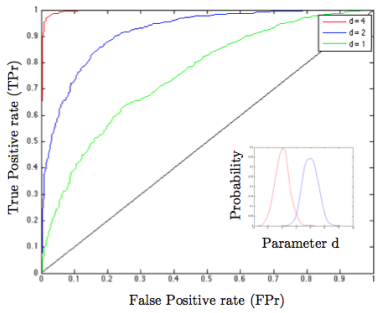


Fig. 4. The ROC graph plotted for two Gaussian data

If the instance is labeled positive and it is classified as positive, it is counted as a true positive tp. If it is classified as negative, it is counted as a false negative fp. If the pattern is negative and is classified as negative, it is counted as a true negative tn. If it is classified as positive, it is counted as a false positive fp. The number of positive (resp. negative) instances is noted P (resp. N). The true positive rate tpr (resp. false positive rate fpr) is defined by tp/P (resp. fp/N), where P (resp. N) is the number of positive (resp. negative) instances. The receiver operating characteristic (ROC) curve is a graph where the true positive rate tpr is plotted on the Y axis and the false positive rate fpr is plotted on the X axis. A ROC curve depicts relative tradeoffs between benefits (true positive) and costs (false positives). The Fig. 4. illustrates the ROC curve for two data sets with Gaussian distribution. Each set belongs to one class, positive p and negative n. As the distance d between the two distributions increases, the decision error decreases. When there is no overlapping between the two data sets, there is no error in classification, which is represented by the point (0,1) in the ROC curve. The ROC point (0,0) means that there is no false positive nor true positive and the point (1,1) represents a maximum rate of true positive and false negative. The classification performance can be calculated with the Area Under the Curve (AUC). We refer the reader to [12] for more information about the ROC curve.

In the time series T^+ (resp. T^-), two ROC curves are plotted to determine the optimal level for detecting peaks P_1^+ and P_2^+ (resp. P_1^+ and P_2^+). In the ROC space, the optimal level is considered as the one which generated the more distant point from the diagonal.

D. Algorithm

The pattern recognition algorithm is modeled by a finite state machine (FSM). The Fig. 5. shows the two types of signatures non-uniformly sampled with four levels. To validate a pattern, two peaks with opposite sign must occur in a pattern window δ. Two levels of minimum amplitude (positive and negative), defined as silence levels, are necessary to detect the pattern beginning.

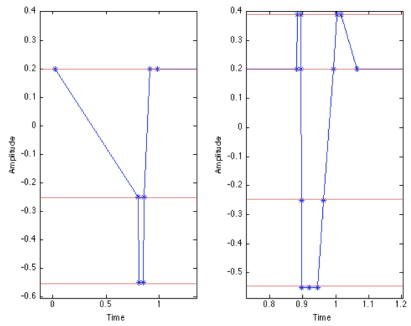


Fig. 5. The patterns depicted in Fig. 2 non-uniformly sampled with 4 levels

The algorithm is designed to detect a global pattern that contains a set of 4 patterns. Consequently, another window Δ is defined in which 4 patterns must be detected. This redundancy is used to decrease the false positive rate fpr without reducing the true positive rate tpr.

IV. Simulation & Results

The data manipulated in this paper consists of experiments with 5 different scenarii. They were sampled at 100 Hz, which is the minimum frequency for detecting the patterns of interest. Our procedure works directly with raw data and thus does not need to extract features and pre-processing. The pattern recognition algorithm was tested with MATLAB. The level crossing sampling scheme allows to drastically reduce the number of sample to process. The results in Tab. I. prove that the non-uniform sampling is adapted to sporadic signals. Indeed, the highest ratio between the number of samples for non-uniform sampling and uniform sampling is less than 1 %. The performance results are summarized in Tab.II. It globally

shows good performances with a very low number of false positive detections. The method proves that well positioning the levels along the amplitude axis allows to prune off non-relevant samples that leads to useless processing.

Record	N_{US} (100 Hz)	N_{NUS}	N_{NUS}/N_{US} [%]
1	8 344 216	39 240	0.47
2	8 640 070	17 000	0.20
3	8 661 893	33 198	0.38
4	8 314 179	12 697	0.15
5	8 060 294	63 394	0.79

TABLE I
RATION BETWEEN THE NUMBER OF SAMPLE FOR NON-UNIFORM SAMPLING N_{NUS} AND UNIFORM SAMPLING N_{US} AT 100 HZ

Record	**False detection**	**True detection**	**Detection rate** [%]
1	0	31/37	84
2	0	26/36	72
3	0	12/30	40
4	0	25/25	100
5	1	33/39	85

TABLE II
THE PATTERN RECOGNITION ALGORITHM DETECTION RATE

V. HARDWARE IMPLEMENTATION

The non-uniform sampling scheme based on level crossing can be implemented with the Asynchronous Analog-to-Digital Converter (A-ADC). The A-ADC, depicted in Fig. 6., consists of four parts [13]. The difference quantifier compares the continuous input signal $i(t)$ and the reference V_{ref}. If the continuous input signal $i(t)$ increases, Up = 1 and Down = 0.

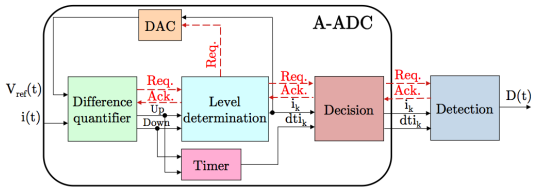


Fig. 6. Block diagram of the A-ADC with the detection block connected to its outputs

In the other case, when $i(t)$ decreases, Up = 0 and Down = 1. The signals Up and Down feed the level determination block that allows to select the appropriate reference signal recorded in memory. Then it is converted to continuous value with a DAC to make the comparison possible with $i(t)$. The timer aims to determine the time elapsed since the previous sample. When the A-ADC is fed with a new sample k, its time dt_{i_k} and amplitude i_k is determined by the decision block. The A-ADC being asynchronous, it establishes a communication with its neighbors blocks in order to exchange data with ack and req signals. The A-ADC and the detection algorithm were implemented on an Altera DE1 Board with Cyclone II EP2C20F484. The input signal was generated with a GBF Tektronix AFG3021 that generates an arbitrary signal from a MATLAB file. $V_{ref}(t)$, obtained at the DAC output, is the input signal $i(t)$ approximated with four level determined with the proposed method (cf. Fig. 7). $wr(t)$ is the write-read signal command of the digital-to-analog converter. It can be considered as the A-ADC activity signal.

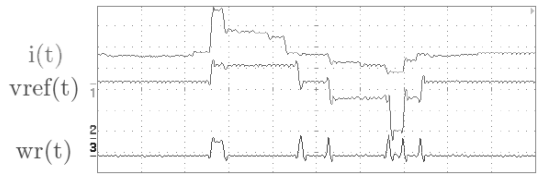


Fig. 7. The approximation of the input signal $i(t)$ with four levels

VI. CONCLUSION

In this paper, we presented a threshold-based partitioning scheme to perform pattern recognition with a level crossing sampling scheme method without pre-processing. The non-uniform sampling is well-adapted to sporadic signals and allows to only capture samples with relevant information. The results show that less than 1 % of data necessary to perform the pattern recognition and to preserve a good detection rate of about 76 %, similar to that of the uniform sampling scheme. Moreover, the proposed method allows to determine the most adequate levels, thanks to the ROC curve.

REFERENCES

[1] D. Figo, P. Diniz, D. Ferreira, and J. Cardoso, "Preprocessing techniques for context recognition from accelerometer data," *Personal and Ubiquitous Computing*, vol. 14, pp. 645–662, 2010.
[2] J. Paradiso and T. Starner, "Energy scavenging for mobile and wireless electronics," *IEEE Pervasive Computing*, vol. 4, pp. 18–27, 2005.
[3] R. Jafari and R. Lotfian, "A low power wake-up circuitry based on dynamic time warping for body sensor networks," in *International Conference on Body Sensor Networks (BSN)*, 2011.
[4] C. S. Daw, C. E. A. Finnery, and E. R. Tracy, "A review of symbolic analysis of experimental data," *Review of Scientific Instruments*, vol. 74, pp. 915–932, 2003.
[5] A. Benbasat and J. Paradiso, "An inertial measurement framework for gesture recognition and applications," *Gesture and Sign Language in Human-Computer Interaction*, vol. 2298, pp. 9–20, 2002.
[6] G. Fan, Fitriani, and W.-B. Goh, "Generic motion gesture detection scheme using only a triaxial accelerometer," in *IEEE 15th International Symposium on Consumer Electronics (ISCE)*, 2011.
[7] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 26, pp. 43–49, 1978.
[8] D. Yu, X. Yu, Q. Hu, J. Liu, and A. Wu, "Dynamic time warping constraint learning for large margin nearest neighbor classification," *Information Sciences*, vol. 181, pp. 2787–2796, 2011.
[9] M. Joselli and E. Clua, "grmobile: a framework for touch and accelerometer gesture recognition for mobile games," in *IEEE VIII Brazilian Symposium on Games and Digital Entertainment*, 2009.
[10] S. Mitra and T. Acharya, "Gesture recognition: A survey," *IEEE Transactions on systems, man, and cybernetics*, vol. 37, pp. 311–324, 2007.
[11] E. Allier, L. Fesquet, G. Sicard, and M. Renaudin, "Low power asynchronous a/d conversion," in *Proceedings of the 12th international workshop on power and timing, modeling, optimization and simulation (PATMOS)*, 2002.
[12] T. Fawcett, "An introduction to roc analysis," *Pattern Recognition Letters 27*, vol. 27, pp. 861–874, 2006.
[13] E. Allier, G. Sicard, L. Fesquet, and M. Renaudin, "Asynchronous level crossing analog to digital converters," *Measurement*, vol. 37, pp. 296–309, 2005.

Particle Filter Acceleration Using Multiscale Sampling Methods

Yaniv Shmueli
School of Computer Science
Tel Aviv University
yaniv.shmueli@cs.tau.ac.il

Gil Shabat
School of Electrical Engineering
Tel Aviv University
gil@eng.tau.ac.il

Amit Bermanis
School of Mathematics
Tel Aviv University
amitberm@post.tau.ac.il

Amir Averbuch
School of Computer Science
Tel Aviv University
amir@math.tau.ac.il

***Abstract*—We present a multiscale based method that accelerates the computation of particle filters. Particle filter is a powerful method that tracks the state of a target based on non-linear observations. Unlike the conventional way that calculates weights over all particles in each cycle of the algorithm, we sample a small subset from the source particles using matrix decomposition methods. Then, we apply a function extension algorithm that uses the particle subset to recover the density function for all the rest of the particles. As often happens, the computational effort is substantial especially when tracking multiple objects takes place. The proposed algorithm reduces significantly the computational load. We demonstrate our method on both simulated and on real data such as tracking in videos sequences.**

***Index Terms*—particle filter, multiscale methods, nonlinear tracking**

I. Introduction

Particle filter (PF) is a powerful method for target state tracking based on non-linear observations obtained by a Monte-Carlo approach [1]. The advantages of PF over different tracking methods such as Kalman filter are in its ability to use non-linear models and the ability to use non-Gaussian distributions. On the other hand, the disadvantage of the PF is its use of Monte Carlo as the performance of the PF strongly depends on the number of particles used. A large number of particles will simulate the required distributions more accurately leading to better results but also increase significantly the computational load. In many cases, weight computation of each particle can be computationally expensive. A common example for this case is object tracking in videos where the weight of each particle is determined by the Bhattacharyya coefficient [2] or by Earth-Moving-Distance (EMD) [3], which requires to evaluate histograms over a large number of bins, especially when color is involved. When the number of particles is either moderate or large (typically a few thousands) the computational load becomes a serious bottleneck to achieve real-time processing.

In this work, we propose a new method to evaluate particles weights using multiscale function extension (MSE) algorithm [4]. The MSE approach consists of two steps: subsampling and extension. In the subsampling step, the particles are sampled to achieve maximal coverage using a small fraction of the actual number of particles with respect to their density. This is done by a special type of matrix decomposition called Interpolative-Decomposition(ID). Then, the weights are computed for this small set of particles. In the next step (extension), the weights are extracted for the rest of the particles using the MSE method. The method uses coarse-to-fine hierarchy of the multiscale decomposition of a Gaussian kernel that represents the similarity between the particles. This generates a sequence of subsamples, which we refer to as adaptive grids, and a sequence of approximations to a given empirical function on the data, as well as their extensions to any missing multidimensional data point. Since in many cases the computational load of the weights is heavy, this approach can reduce the computational load significantly and accelerate the PF, allowing us to use more particles. Increasing the number of particles is needed since many of today tasks are geared to track objects "buried" in huge data streams such as video, communication and telemetric data.

Particle filters were studied in many works and used in several applications domains such as computer vision, robotics, target tracking and finance. While PF can be robust to both the input observations distribution and the behavior of the additive noise, its implementation is computationally intensive. Making it working in real-time (computationally efficient) has become a major challenge when objects tracking is done in high dimensional state space, or when dealing with multiple target tracking. Comprehensive tutorials and surveys on the different variations and recent progress in PF methods are given in [1], [5].

II. Particle Filter Algorithm

In general, PF is a model estimation technique based on simulation that uses Monte Carlo methods for solving a recursive Bayesian filtering problem [1]. It is used for estimating the state x_n at time n from a noisy observations $y_1, ..., y_n$. A dynamic state space equations are used for modeling and prediction. The basic idea behind PF is to use a sufficiently large number of "particles". Each particle is an independent random variable which represents a possible state of the target. For example, a state can be a location and velocity. In this case, each particle represents a possible location and velocity of the target from a proposed distribution. The system model is applied to the particles in order to perform prediction to the next state. Then, each particle is assigned a weight, which represents its reliability or the probability that it represents the real state of the target. The actual location (the output

of the PF) is usually determined as the maximal likelihood of the particle's distribution. The algorithm robustness and accuracy are determined by the number of computed particles. A large number of particles is more likely to cover a wider state subspace in the proximity of the target, as well as a better approximation of the state distribution function. However, the cost of such improved tracking produces higher computational load since each particle needs to be both advanced and weighted while this is repeated in each cycle.

III. Multiscale Function Extension

Given a set of N particles $\mathcal{P}_N = \{p_1, p_2, \ldots, p_N\}$ and their mutual distances, we wish to estimate the value of their weight function through a small subset $\mathcal{P}_n$ of n particles ($n < N$ is a predefined number), for which we compute the weights directly. Formally, our goal is to interpolate the weight function $W : \mathcal{P}_n \to \mathbb{R}$ to $\mathcal{P}_N$, given a distance function $d : \mathcal{P}_N \times \mathcal{P}_N \to R$. For that purpose we use the multiscale function extension (MSE) [4], which is a multiscale, numerically stable interpolation method.

Each scale of the MSE is divided into two phases - a subsampling phase and an extension phase. The first phase is done by a special decomposition, known as interpolative decomposition (ID), of an affinities matrix associated with $\mathcal{P}_n$. The second phase extends the function from $\mathcal{P}_n$ to $\mathcal{P}_N$, using the output of the first (sampling) phase. The essentials of the MSE are describe in sections III-A and III-B. For further reading we refer the reader to [4].

We use the following notation: s denotes the scale parameter, $s = 0, 1, \ldots$, $\epsilon_s = 2^{-s}\epsilon_0$ for some positive number ϵ_0, and $g^{(s)}(r) = \exp\{-r^2/\epsilon_s\}$. For a fixed scale s we define the function $g_j^{(s)} : \mathcal{P}_N \to \mathbb{R}$, $g_j^{(s)}(p) = g^{(s)}(d(p_j, p))$ to be the Gaussian of width ϵ_s, centered at p_j. $A^{(s)}$ is the $n \times n$ affinities matrix associated with $\mathcal{P}_n$, whose (i,j)-th entry is $g^{(s)}(d(p_i, p_j))$. Note that the j-th column of $A^{(s)}$ is the restriction of $g_j^{(s)}$ to $\mathcal{P}_n$. $\mathcal{P}_n^c$ is the complementary set of $\mathcal{P}_n$ in $\mathcal{P}_N$. The spectral norm of a matrix A is denoted by $\|A\|$, and its j-th singular value (in decreasing order) is denoted by $\sigma_j(A)$.

A. Data subsampling through ID of Gaussian matrix

Let s be a fixed scale. Our goal is to approximate W by a superposition of the columns of the affinities matrix $A^{(s)}$, then to extend W to $p_* \in \mathcal{P}_n^c$, based on the affinities between p_* and the elements of $\mathcal{P}_n$. At first sight, we could solve the equation $A^{(s)}c = W$ and, using the radiality of $g^{(s)}$, to extend W to p_* by $\hat{W}(p_*) = \sum_{i=1}^{n} c_i g_j^{(s)}(p)$, which is exact on $\mathcal{P}_n$, i.e. $\hat{W}(p_j) = W(p_j), j = 1, 2, \ldots, n$. This method is known as Nyström extension [6], [7]. As proved in [4], the condition number of $A^{(s)}$ is big for small values of s, namely $A^{(s)}$ is numerically singular. On the other hand, too big s would be resulted in a short distance interpolation. Moreover, even if we would choose such s for which $A^{(s)}$ is numerically non-singular and the interpolation is not for too short distance, interpolation by a superposition of translated Gaussian of fixed width, would not necessarily fit the nature of W.

In order to overcome the numerical singularity of $A^{(s)}$, we apply an interpolative decomposition (ID). The deterministic version of the ID algorithm can be found in [4], whose complexity is $\mathcal{O}(mn^2)$, and a randomized version can be found in [8]. The latter is based on random-projections and its complexity is $\mathcal{O}(k^2 n \log n)$. Since each column of $A^{(s)}$ corresponds to a single data point in $\mathcal{P}_n$, selection of columns subset from $A^{(s)}$ is equivalent for subsampling of $\mathcal{P}_n$ data points.

B. Multiscale function extension

Let $A^{(s)} = B^{(s)}P^{(s)}$ be the ID of $A^{(s)}$, where $B^{(s)}$ is an $n \times k$ matrix, whose columns constitute a subset of the columns of $A^{(s)}$, and let $\mathcal{P}^{(s)} = \{p_{s_1}, \ldots, p_{s_k}\}$ the associated sampled dataset. The extension of W to $\mathcal{P}_n^c$ is done by orthogonally projecting W on the columns space of $B^{(s)}$, and extending the projected function to $\mathcal{P}_n^c$ in a similar manner to Nyström extension method, using the radiality of $g^{(s)}$. The single scale extension (SSE) algorithm can be found in [4] (Algorithm 3), whose complexity is $\mathcal{O}(nk^2)$.

Obviously, $w^{(s)}$ is not necessarily equal to w, namely the output of the SSE algorithm is not an interpolant of w. In this case we apply the SSE algorithm once again to the residual $w - w^{(s)}$ with narrower Gaussian, that ensures a bigger numerical rank of the next-scale affinities matrix $A^{(s+1)}$ and, as a consequence, a wider subspace to project the residual on. Such approach is described in Algorithm 4 in [4]. We shell call this algorithm the multiscale extension (MSE) Algorithm, whose complexity is $\mathcal{O}(n^3)$.

IV. Multiscale Particle Filter (MSPF)

In order to accelerate the PF computation, while executing it with a large number of particles, we will apply an intelligent sampling of the particles, followed by an extension method to compute the weights of the rest. This will allow us to compute a relatively small number of particle weights in each cycle. Such approach can be effective if the particle weight calculation is computationally expensive.

A. Particle Subsampling

In each cycle of the PF algorithm, we first resample a new set of N particles from the set $\mathcal{P} = \{x_t^{(n)}, w_t^{(n)}\}, n = 1, .., N$ using their weights as the distribution function. Once we apply the dynamic model on each particle and advance it, we need to compute their new weights. To do that, we first select a small subset of the particles. We wish to find a good set of particle candidates that will capture the geometry of the weight function $W : \mathcal{P}_n \to \mathbb{R}$. To find such candidates, we define a distance metric between the particles. In our experiments we used the euclidean distance between each two particles viewed as vectors, but other metrics can be used as well. We select the particle candidates using the ID Algorithm described in Section III-A. We construct an affinity matrix $A^{(s)}$ containing the affinities between the particles, using a Gaussian kernel

that is based on the given distance metric $d(p_i,_p j)$ between the particles.

$$[A^{(s)}]_{ij} = \exp\left(\frac{-d(p_i,p_j)^2}{\epsilon_s}\right), i,j = 1,..,N. \quad \text{(IV.1)}$$

We calculate the affinities for all the particles in $\mathcal{P}$ so $A^{(s)}$ is an $N \times N$ matrix defined by Equation IV.1. The number of candidates we wish to receive is at most k. The value k is usually selected according to the computation budget we have to calculate the weight function in each cycle. The output of the ID algorithm will be a set $\mathcal{P}_k$ of k particles selected from $\mathcal{P}$. We compute the weights of the k particles we selected, as we do in the original PF algorithm.

B. Weight Calculation using Function Extension

Now that we have a set of particles $\mathcal{P}_k = \{x_t^{(n)}, w_t^{(n)}\}, n = 1,..,k$ with their calculated weight values, we can continue and compute the weights of the rest of the particles. Using the MSE algorithm, we compute the weight value of each of the other $N-k$ particles, by using the set $\mathcal{P}_k$, and the first k columns of the affinity matrix $A_k^{(s)}$. These columns contains the affinities between each pair of particles in $\mathcal{P}_k$ and the affinities between particles in $\mathcal{P}_k$ and all other particles. The output of the MSE algorithm is the weights of the $N-k$ particles that were not selected in the previous step. The extension method allow us to avoid a direct computation of the weight for the rest of the $N-k$ particle. This is especially effective when we can not compute the weight for all particles if the observation has some missing data, or if the computation is too intense. Once we calculated all the weights we select the particle with the maximum likelihood (weight) as the prediction result and continue to the next cycle.

Algorithm IV.1: Multiscale Particle filter
Input: Initial state x_0 and current observations $y_1, ..., y_T$
Output: Estimated observations $x_1, ..., x_T$

1: Initialize weights $w_0^{(n)} = \frac{1}{N}$, and $x_0^{(n)} \sim p(x_0)$, $n = 1, ..., N$.
2: **for** time steps t=1,...,T **do**
3: Resample N new particles by their distribution determined by weights: $w_t^{(n)}$
4: Prediction: Apply the dynamic model on each particle to estimate next state using x_{t-1} and $y_1, ..., y_t$

$$\tilde{x}_t^{(n)} \sim q(x_t^{(n)} | x_{t-1}^{(n)}, y_1, ..., y_t)$$

5: Selection : Select a subset of size k out of the new particles $\tilde{x}_t^{(n)}$, by computing the affinity matrix $A^{(s)}$ and using the ID Algorithm.
6: Calculate weights of the k selected particles using

$$w_t^{(n)} \propto \frac{p(y_t|x_t^{(n)})p(x_t^{(n)}|x_{t-1}^{(n)})}{q(x_t|x_{t-1}^{(n)}, y_1, ..., y_t)}, n = 1, ..., N$$

7: Weight extension: Calculate the weights of the $N-k$ particles using the Multiscale Function Extension Algorithm (See Algorithm 4 in [4]).
8: Normalize weights:

$$\tilde{w}_t^{(n)} = \frac{w_t^{(n)}}{\sum_{i=1}^N w_t^{(i)}}$$

9: Set x_t to be the particle $\tilde{x}_t^{(i)}$ with maximum weight $\tilde{w}_t^{(i)} \geq \tilde{w}_t^{(n)}, n = 1, .., N$
10: **end for**

V. Experimental Results

To test the performance of Algorithm IV.1, we preformed several experiments of tracking objects in both synthetic and real videos, and comparing the results to other tracking methods.

A. Multiple Target Tracking

he MSPF Algorithm IV.1 was tested on a video sequence that contains multiple objects. In this case, the tracking was achieved by the application of two different PF types algorithms, where each had its own set of particles and a separate set of observations. Each particle describes a state of a single target. Another approach to track multiple objects is to create a "super-state" particle, which describes the state of all the objects inside the video sequence. In this case, the number of fields inside the particle vector was $n \times k$ where n is the number of targets and k is the number of parameters required to describe a single target. In the latter scenario, the MSE algorithm outperformed standard interpolation methods since it handled better data points in high dimensions. The advantage of using the "super-state" particle is by enabling to advance a particle state by dynamic model equations that took into account the state of all the objects within a particle including dependencies between objects.

In order to test the tracking performance using the "super-state" particle, we tracked two tennis players in a video sequence. The players are represented by a single particle with $6 \times 2 = 12$ coordinates, 6 for each player (location in x and y, velocity in x and y, width and hight). In each algorithm cycle, the prediction step advanced the particles by the application of the model equations separately to each coordinate. The weight calculation was done in each region separately and then multiplied the Bhattacharyya coefficient to obtain a single weight. Then, the extension step was applied as before using the weighted Euclidean metric for each particle that has 12 coordinates. By using Algorithm IV.1, we were able to track both targets successfully with the lowest computational cost in comparison to other extension methods that are based on standard interpolation such as B-splines, cubics and nearest neighbor. We used 1500 particles to track both players. In each step of the algorithm, we calculated the weights for 150 selected particles and interpolated the weights for the other 1350 particles by using the MSE method. The complete videos of the basketball and tennis games tracking can be seen in our

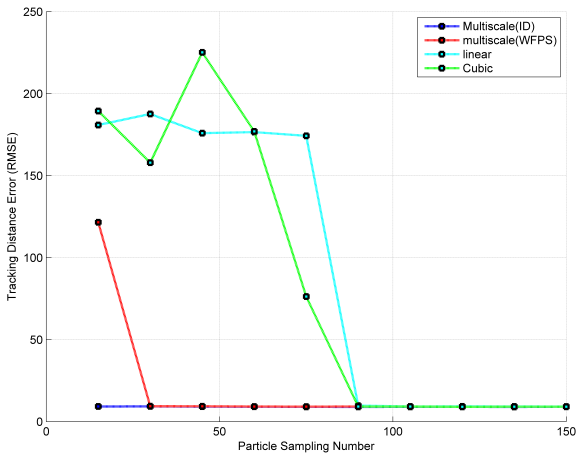


Fig. V.1. Comparing the RMSE between different methods - multiscale with ID sampling , Multiscale with WFPS sampling, linear approximation, cubic approximation.

website[1].

B. Comparison with Other Approximation Methods

In order to compare the proposed method with different approximation methods, we applied the PF algorithm and in each run we used a different approximation method to calculate the particle's weights. In this comparison, we used a synthetic movie. We generated a video sequence by moving a colored disc over a still image. The disc moved along a non-linear parametric function. This allows us to know the ground truth of the target at any frame. We applied the PF algorithm to the synthetic video sequence several times, each with different interpolation method. We compared the total Root Mean Square Error (RMSE) for each approximation method measured on the distance between the estimation and the real location of the target. The MSPF Algorithm IV.1 had the lowest error rate even when we use a sampling factor between 2%-5% from the total number of particles. When such subsampling factor was used, all the other tested methods fail (error grew).

Overall, the MSPF Algorithm IV.1 achieved the lowest computational time while maintaining a low error rate.

C. Comparison with the EMD Measurement

Recently, the Earth Moving Distance (EMD) [3] was used for particles weight computation since this particle weight fits deformable objects [9]. We tested Algorithm IV.1 with the EMD metric to demonstrate how well the extension scheme fits it. Several runs were conducted on the "Lemming" sequence from the PROST database. Weights were calculated for 10% from the total number of particles while the rest of the particles were estimated using the MSE Algorithm.

Table V.1 shows the time differences between the naive version of the PF algorithm that uses the EMD metric and our implementation that uses the MSE Algorithm. For the latter, 10% of the particles were sampled, and the MSE was applied to the other 90% of the particles.

[1] http://www.cs.tau.ac.il/~yanivshm/mspf

TABLE V.1
EMD ACCELERATION TIME COMPARISONS IN [SEC]. SAMPLING WAS 10% FROM THE TOTAL NUMBER OF PARTICLES.

# of Particles	Time [no MSE]	Time [with MSE]	Acceleration Factor
2000	63	10.6	5.9
4000	125	32	3.9
6000	187	75.4	2.5
8000	260	151	1.7
10000	294	266	1.1

VI. CONCLUSION

In this work, we presented several contributions. We reduced the PF computational time by applying a novel multiscale extension (MSE) method to reduce the particle weight calculation. This allowed us to use more particles within a given computational budget thus improving the performance of the PF. We tested our modified PF algorithm on real video sequences to track a single and multiple targets, and compared it with other extension methods.

ACKNOWLEDGMENT

This research was partially supported by the Israel Science Foundation (Grant No. 1041/10) and by the Israeli Ministry of Science & Technology (Grant No. 3-909).

REFERENCES

[1] M. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking," *Signal Processing, IEEE Transactions on*, vol. 50, no. 2, pp. 174–188, 2002.

[2] A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by their probability distributions," *Bull. Calcutta Math. Soc*, vol. 35, no. 99-109, p. 4, 1943.

[3] Y. Rubner, C. Tomasi, and L. Guibas, "The earth mover's distance as a metric for image retrieval," *International Journal of Computer Vision*, vol. 40, no. 2, pp. 99–121, 2000.

[4] A. Bermanis, A. Averbuch, and R. Coifman, "Multiscale data sampling and function extension," *Applied and Computational Harmonic Analysis*, vol. http://dx.doi.org/10.1016/j.acha.2012.03.002, 2012.

[5] A. Doucet and A. Johansen, "A tutorial on particle filtering and smoothing: Fifteen years later," *Handbook of Nonlinear Filtering*, pp. 656–704, 2009.

[6] C. Baker, *The numerical treatment of integral equations*. Clarendon press Oxford, 1977, vol. 13.

[7] B. Flannery, W. Press, S. Teukolsky, and W. Vetterling, "Numerical recipes in c," *Press Syndicate of the University of Cambridge, New York*, 1992.

[8] P. Martinsson, V. Rokhlin, and M. Tygert, "A randomized algorithm for the decomposition of matrices," *Appl. Comput. Harmon. Anal.*, 2010.

[9] S. Avidan, D. Levi, A. Bar-Hillel, and S. Oron, "Locally orderless tracking," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 1940–1947.

Analysis of Multistage Sampling Rate Conversion for Potential Optimal Factorization

Zhengmao Ye, Habib Mohamadian
Southern University, Baton Rouge, Louisiana, USA
Email: zhengmao_ye@subr.edu, habib_mohamadian@subr.edu

Abstract — **Digital multistage sampling rate conversion has many engineering applications in fields of signal and image processing, which is to adapt the sampling rates to the flows of diverse audio and video signals. The FIR (Finite Impulse Response) polyphase sampling rate converter is one of typical schemes that are suitable for interpolation or decimation by an integer factor. It also guarantees the stability performance with the stable gain margin and phase margin. The big challenge occurs upon implementation when a very high order filter is needed with large values of L (positive integer factor of interpolator) and/or M (positive integer factor of decimator). Narrowband linear phase filter specifications are hard to achieve, however. It leads to extra storage space, additional computation expense and detrimental finite word length effects. The multistage sampling rate converter has been introduced to factorize the L and M ratio into a product of ratios of integers or prime numbers. The optimal number of stages and optimal converting factors are both critical terms to minimize the computation time and storage requirements. Filter structure analysis is conducted in this study to search for the potential factors that could have a remarkable impact to optimize the sampling rate conversion.**

Keywords- Polyphase FIR Filter, Interpolation, Decimation, Sampling Rate Conversion, Multistage, Multirate, Optimization

I. INTRODUCTION

The sampling process stems from obtaining discrete time signals from the continuous time signals at the regular time intervals. Sampling can be conducted for functions varying in space, time, or any other dimension. For discrete time signals, potential upsampling, downsampling and multirate multi-stage sampling rate conversion can be applied as well [1-2]. There are a wide variety of important real world applications on sampling. For instance, the emerging GPS (Global Positioning System) enabled cell phones offer new opportunities of data collection in the massive volumes at relatively cheaper cost than the dedicated probe vehicles. The traffic monitoring applications need to firstly determine whether the GPS-enabled cell phone is actually in an automobile and secondly, it needs to match the current GPS device location to a corresponding link on a GIS (Geographic Information System) map. A methodology is developed to determine relationships between the cell phone pinging sampling rate and accuracy of mode detection and map matching processes. It is found that the higher the number of pings per interval and the longer the data trace interval, the better the accuracy. The impact of the sampling frequency on the map matching algorithm is found to be a function of link length, current speed of a vehicle and period of the day [3]. In many cases, sampling rate conversion is required by digital systems dedicated to audio and speech processing in order to adapt the sample rate to different signal flows. For example, 8 kHz and 16 kHz for speech, 32 kHz for broadcasting, 44.1 kHz for CDs, and 48 kHz for studio work. The sampling rate conversion (SRC) is based on the objective criteria, such as complexity, integration cycle and performance characterization. The proposed SRC system has the capability of fully recovering characteristics and rounding noise behavior [4]. A linear phase Finite Impulse Response (FIR) filter of an arbitrary order is designed for the sampling-rate conversion by a rational factor of L/M (upsampling / downsampling) also. The coefficient symmetry of the linear-phase filter is exploited with a minimal number of delay elements. The number of multiplications per output sample is reduced approximately by a factor of two compared with the conventional polyphase implementation [5]. Similarly, a class of farrow-structure-based reconfigurable bandpass FIR filters for integer sampling rate conversion is introduced. Both Mth-band and general FIR filters can be realized and the filters work equally well for any integer factor and passband location. The proposed sampling rate converters provide the considerably higher efficiency and fewer filter coefficients [7]. Rational sampling rate conversion can also be performed in the domains of discrete Fourier transform and discrete cosine transform. Conversion error performance and computational complexity are based on the proposed fast transform algorithms. It can achieve substantial improvements on the conversion accuracy at the reduced computational cost, compared with the conventional lowpass filter [6]. To evaluate the performance of sampling receiver, a sub-Nyquist rate sampling receiver architecture is presented that exploits signal sparsity by employing compressive sensing techniques. The receiver works at sampling rates much lower than the Nyquist rate whose performance is quantified analytically. A new parallel path structure is used. The receiver performance is quantified analytically. It is shown that an instantaneous receiver signal bandwidth of 1.5 GHz and a Signal to Interference plus Noise Ratio (SINR) of 44 dB are achievable [8]. Upsampling and downsampling can further be applied to digital image processing to enhance the image quality [9-10].

From the most recent research outcomes, the problem of optimal multistage sampling rate converters has never been solved in terms of both the optimal number of stages and the optimal converting factor. For concern of the potential optimal solution, a case study has been made based on an example of the cascades of FIR polyphase filter design when converting the sampling rate for a stream of signals from audio DVD data (96 kHz) to audio CD data (44.1 kHz) with some interesting new results.

II. Single Stage Sample Rate Conversion

In order to convert a stream of signals from 96 kHz to 44.1 kHz, the FIR lowpass filter structure is selected to perform the single stage sampling rate conversion. By nature, in contrast to the Infinite Impulse Response (IIR) filter, FIR filters are always stable and easy to implement with all poles located at the origin. On the other hand, a higher order is necessary. For a matter of simplicity, upsampling and downsampling are processed in the single stage, where anti-aliasing filtering is applied before downsampling (decimation) and anti-imaging filtering is applied after upsampling (interpolation). The combined cutoff frequency is selected as the minima of anti-aliasing filtering and anti-imaging filtering. The frequency response of the single stage sampling rate conversion is shown in Fig. 1. Due to the large values of L (147) and M (320), a narrowband lowpass FIR filter has been produced. It is tough to implement while a very high order filter is necessary. The finite word length effect is thus generated. Also extra storage space and long simulation time are needed. The multistage sampling rate converter is an alternative to the single stage sampling rate converter. The multistage structure serves as a tradeoff to the harmful finite word length effects. The conversion ratio can be translated into a product of ratios, where smaller factorized values of the interpolation factor (L) and decimation factor (M) will be achieved. From literatures there is no systematic approach in determining an optimal number of stages and an optimal structure to factorize a set of L/M ratio so as to minimize the computation time and storage space. The trial and error method is mostly applied so far.

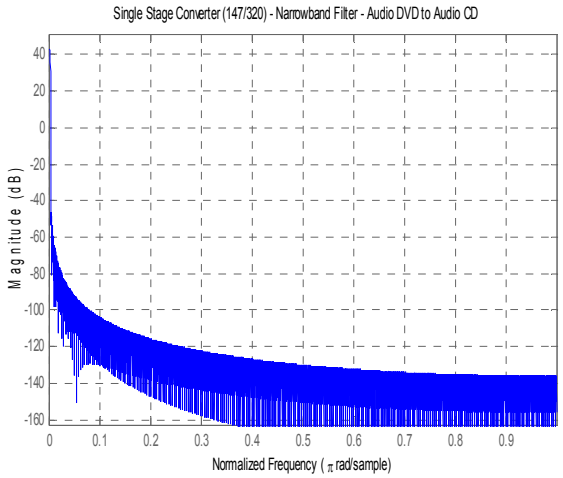


Fig. 1 Single Stage Sampling Rate Conversion - Narrowband Filter

III. MultiStage Sample Rate Conversion

The multistage sampling rate converter has been designed to substitute the single stage sample rate conversion. Overall it is a downsampling process (decimation) and three stage cascade structure has been applied. The block diagram is shown in Fig. 2, where three composite anti-aliasing and anti-imaging filters H1(f), H2(f) and H3(f) are applied to each stage, respectively.

Fig. 2 Multiple Stage Sampling Rate Conversion

Since the three stage factorization approaches for the required L/M ratio of 147/320 are not unique, three typical cases are chosen whose filter specifications are shown in Table 1. Here, the stage conversion rates of three cases have been selected as:

A. [7, 7, 3] / [8, 5, 8]
B. [7, 3, 7] / [8, 5, 8]
C. [7, 3, 7] / [8, 4, 10]

where a combination of two decimators and one interpolator is applied in Case A, while combinations of three adjustable decimators are applied in Case B and Case C. At each stage of all 3 cases, the ideal gains of the low frequency passbands, sampling rates and cutoff frequencies are also provided.

Table 1. Specifications for Multiple Stage Converter Design

Sample Rate Conversion 96 to 44.1K	L/M	H(f) Passband	fs (kHz)	F (kHz)	Order
Single Stage	147/320				7680
A - Stage 1	**7/8**	**7**	**96**	**6**	**192**
A - Stage 2	7/5	7	84	6	168
A - Stage 3	3/8	3	117.6	7.35	192
B - Stage 1	**7/8**	**7**	**96**	**6**	**192**
B - Stage 2	3/5	3	84	8.4	120
B - Stage 3	7/8	7	50.4	3.15	192
C - Stage 1	**7/8**	**7**	**96**	**6**	**192**
C - Stage 2	3/4	3	84	10.5	96
C - Stage 3	7/10	7	63	3.15	240

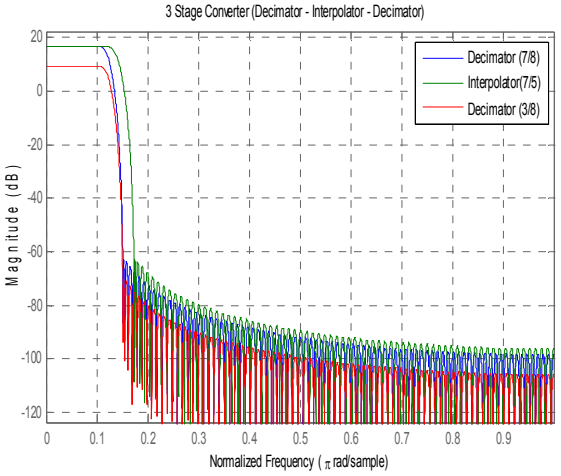


Fig. 3 Three Stage Sampling Rate Conversion – Case A

In addition, small integer factors are the solutions with the low fractional rate conversions in each case. Thus the FIR polyphase sample rate converters are applied to decimator and interpolator design. The reason is that the polyphase FIR filters are well suitable for interpolation or decimation by a small integer factor. In this way, detrimental finite word length effects can be avoided. The regular lowpass filters can be formulated in this way rather than the single stage narrowband lowpass filter. It can be shown in Table 1 that the orders of

three stage sampling rate converters are much lower than that of the single sampling rate converters (i.e., 7680). Using the multistage approach, an audio DVD to audio CD converter can be realized. The frequency responses of the multirate three stage sampling rate converters are plotted in Figs. 3-5. At the same time, among three different cases, case A has a relatively higher order design than Case B and Case C.

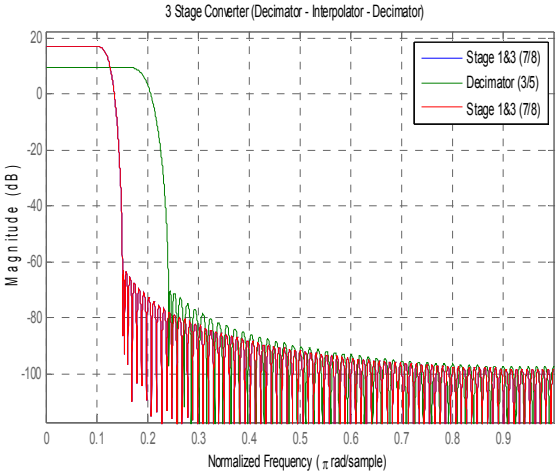


Fig. 4 Three Stage Sampling Rate Conversion – Case B

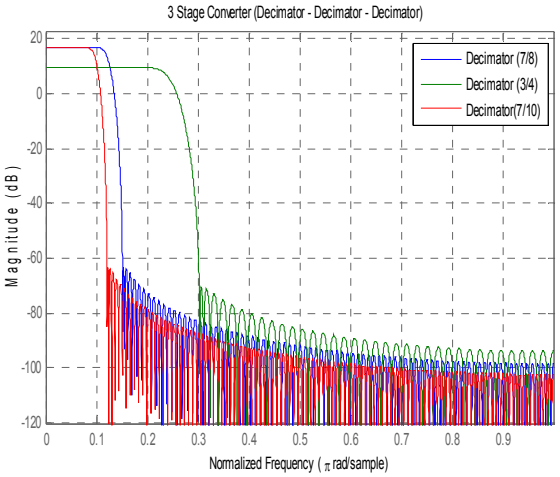


Fig. 5 Three Stage Sampling Rate Conversion – Case C

IV. Detailed Case Studies via Bode Plot

The Nyquist–Shannon sampling theorem is strictly followed upon the design of direct-form FIR polyphase sampling rate converter. Hence, the source signals can also be reconstructed as the bandwidth of a baseband signal is less than the Nyquist frequency. The right choice of the FIR structure ensures the stability of filter design. This fact is also clearly shown in the Bode diagrams as Figs. 6-8, where the stable gain margin and phase margin have been depicted in all cases, no matter it is a higher order filter or lower order FIR filter. Essentially, this sampling rate converter is a decimator instead of an interpolator. Therefore, as Case A has a separate stage of the interpolator design, it results in a relatively higher order design than two other cases with the cascade structure of three decimators. For Case B and Case C, further analysis is needed to compare the merit and drawback.

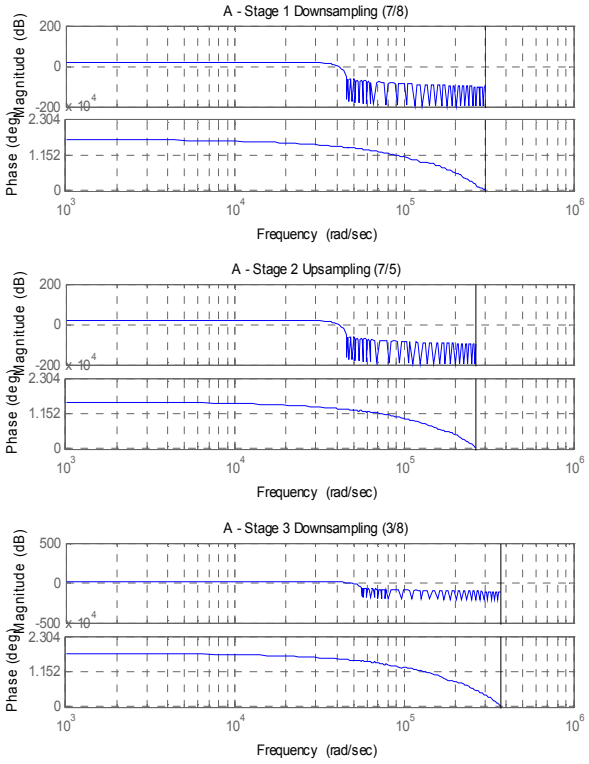


Fig. 6 Bode Plot of Multistage Interpolator and 2 Decimators – Case A

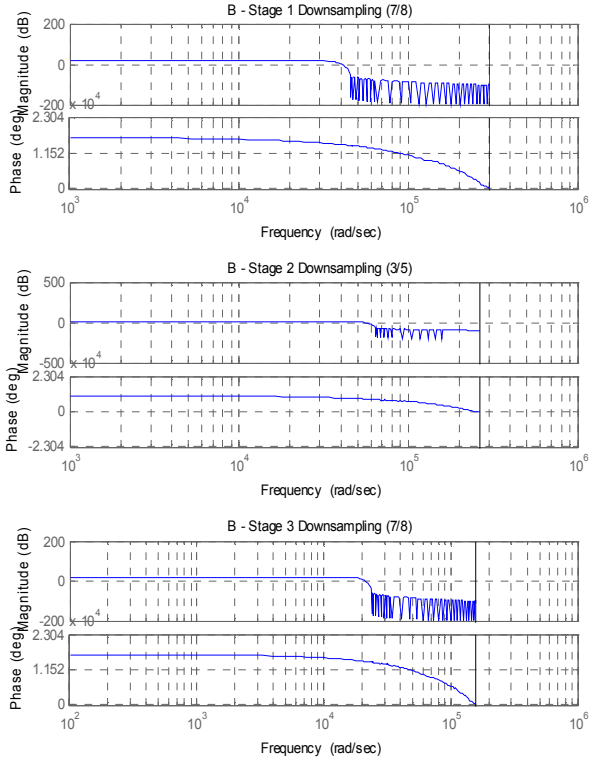


Fig. 7 Bode Plot of 3 Multistage Decimators – Case B

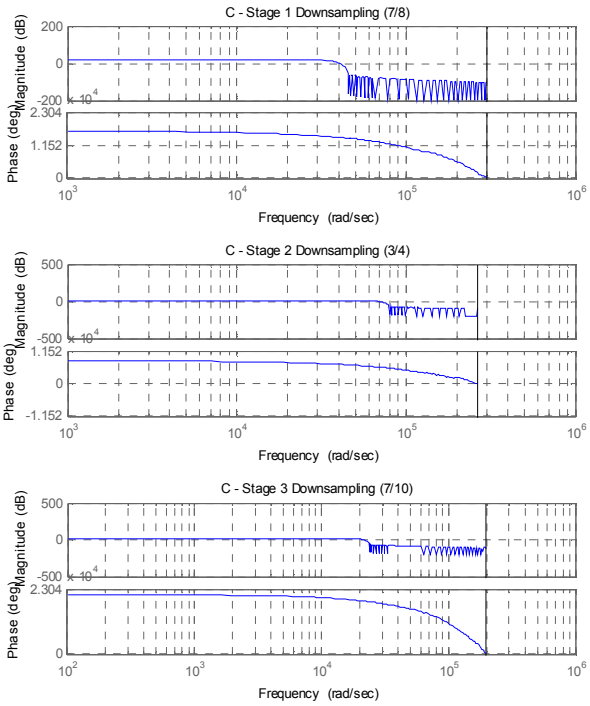


Fig. 8 Bode Plot of 3 Multistage Decimators – Case C

V. Complexity in Multistage Converter Design

To quantify the storage space and computation time needed for sampling rate conversion, a simple way could be achieved by comparing the actual amount of adders and multipliers being used as well as the number of states associated. For the high order single stage sampling rate converter and three low order multistage sampling rate converters being employed, the related expenses are listed in Table 2. Among 3 individual cases, when all three stages are counted together, Case A requires maximal numbers of states and electronic elements since a single stage interpolator has a negative effect on the overall decimator design. Case B requires a similar number of states to Case C but it requires less numbers of adders and multipliers than Case C. It results from the fact that more prime number factors of interpolation and decimation are used in Case B than Case C.

Table 2. Expense for Sampling Rate Converter Design

Expense (Quantity)	States	Adder	Multiplier	Adder + Multiplier
Single Stage	52	7510	7657	15167
A - Stage 1	27	162	169	331
A - Stage 2	23	138	144	282
A - Stage 3	63	166	169	335
B - Stage 1	27	162	169	331
B - Stage 2	38	94	97	191
B - Stage 3	27	162	169	331
C - Stage 1	27	162	169	331
C - Stage 2	31	70	73	143
C - Stage 3	34	210	217	427

Conclusions

Diverse types of data in the forms of audio, video or radio frequency signals are preferably processed in the digital domain. The conversion among different types of signals is frequently applied. Single stage sampling rate conversion is straightforward but the high order requirement represents a big challenge. The multistage sampling rate converter is a better solution. It is however still subject to optimization. There is no existing solid rule to reach the optimal number of stages and the optimal factor of L/M for the concern of computation time and memory storage. Case studies have been conducted in this article to seek for some feasible means towards the best solution of sampling rate conversion. If possible, each stage of multistage converters should be selected following the actual type of the single stage converter exactly, either decimator or interpolator. Also the prime number factors for interpolation and decimation are preferable throughout the upsampling and downsampling processes.

References

[1] R.Schilling, S.Harris, “Fundamental of Digital Signal Processing using MATLAB”, Cengage Learning, 2nd Edition, Cengage Learning, 2012

[2] Vinay K. Ingle, John G. Proakis, "Digital Signal Processing", 2nd Edition, Cengage Learning, 2007

[3] Y. Byon, B. Abdulhai, A. Shalaby, “Impact of Sampling Rate of GPS-Enabled Cell Phones on Mode Detection and GIS Map Matching Performance”, 2007 Transportation Research Board 86th Annual Meeting, Washington DC, Jan 21-25, United States

[4] S. Tassart, "Time-invariant context for sampling rate conversion systems", IEEE Transactions on Signal Processing, v 60, n 3, p 1098-1107, March 2012

[5] R. Bregovic, Y. Yu, T. Saramaki, C. Yong, "Implementation of linear-phase FIR filters for a rational sampling-rate conversion utilizing the coefficient symmetry", IEEE Transactions on Circuits and Systems, v 58, n 3, p 548-561, 2011

[6] G. Bi, S. Mitra, S. Li, "Sampling rate conversion based on DFT and DCT", Signal Processing, v 93, n 2, p 476-486, Feb. 2013

[7] H. Johansson, "Farrow-structure-based reconfigurable bandpass linear-phase FIR filters for integer sampling rate conversion", IEEE Transactions on Circuits and Systems II: Express Briefs, v 58, n 1, p 46-50, January 2011

[8] X. Chen, Z. Yu, S. Hoyos, B. Sadler, "A Sub-Nyquist Rate Sampling Receiver Exploiting Compressive Sensing", IEEE Transactions on Circuits and Systems, v 58, n 3, p 507-520, March 2011

[9] R. Gonzalez, R. Woods, “Digital Image Processing,” 3rd Edition, Prentice-Hall, 2007

[10] Z. Ye, H. Mohamadian and Y. Ye, "Information Measures for Biometric Identification via 2D Discrete Wavelet Transform", Proceedings of 2007 IEEE International Conference on Automation Science and Engineering (CASE 2007), pp. 835-840, Sept. 22-25, 2007, Scottsdale, Arizona, USA

Sparse 2D Fast Fourier Transform

André Rauh and Gonzalo R. Arce
Department of Electrical and Computer Engineering
University of Delaware
Newark, DE 19716
Email: rauh@udel.edu, arce@udel.edu

***Abstract*—This paper extends the concepts of the Sparse Fast Fourier Transform (sFFT) Algorithm introduced in [1] to work with two dimensional (2D) data. The 2D algorithm requires several generalizations to multiple key concepts of the 1D sparse Fourier transform algorithm. Furthermore, several parameters needed in the algorithm are optimized for the reconstruction of sparse image spectra. This paper addresses the case of the exact k-sparse Fourier transform but the underlying concepts can be applied to the general case of finding a k-sparse approximation of the Fourier transform of an arbitrary signal. The proposed algorithm can further be extended to even higher dimensions. Simulations illustrate the efficiency and accuracy of the proposed algorithm when applied to real images.**

I. Introduction

The Fast Fourier Transform (FFT) has become ubiquitous in signal processing applications. While the FFT does not make any assumptions about the structure of the signal, very often the signal of interest is obtained from a structured source resulting in a nearly sparse Fourier spectrum. Assuming that a signal of length N is k-sparse ($k < N$) in the Fourier domain, we can describe the signal with only these k coefficients. This fact is the basis for signal compression and is used among others in the popular MP3 codec. Due to the fact that the signal is accurately described with just k coefficients it seems natural that there should be a better performing algorithm that exploits this property of the signal. Several algorithms have been proposed with this goal [2], [3], [4], [5], [6]. One particular approach is the so called sparse FFT (sFFT) which lowered the computational complexity significantly was introduced recently in [1]. The authors focused on the one dimensional case and the extension to two or multi-dimensions is not straight forward. Since the Fourier transform is separable, it is tempting to sequentially apply the 1D sFFT algorithm separately on all rows and columns. This approach, however, would not be efficient as the algorithm would be of complexity at least $O(N)$ for a signal of NxN samples which is not nearly as good as the proposed sub-linear algorithm. In addition, this strategy does not exploit the intrinsic two dimensionality of the signal and leads to a sub-optimal algorithm. This paper aims to fill this gap and introduces the extensions that are necessary to the algorithm.

The paper is organised as follows: Section II introduces the main ideas and workings of the sparse Fourier transform. Section III describes the new concepts and necessary modifications of the algorithm to extend it to 2D. Due to the sensitivity of the parameter of the algorithm, Section IV lays out guidance as to how the parameters should be selected under the assumption of a natural image as the input. Simulations and the conclusion are provided in the last two sections.

II. Sparse Fourier Transform Algorithm

It would be infeasible for this paper to describe in detail the sFFT algorithm in its entirety. Instead we refer the reader to [1] (up to page 9) and only describe the principal components of the algorithm which are necessary to understand the proposed extension. First, the notation is introduced. Note however that the notation will be re-used for the 2D case in the next Section. Given a signal x of length N we denote its discrete Fourier transform as $\hat{x}$. A signal is considered to be k-sparse if there are only k non-zero components in $\hat{x}$. Furthermore we define $\omega = e^{-2\pi\imath/N}$.

The key idea of the sFFT algorithm is to hash the k coefficients into few buckets in sublinear time. This is achieved by using a carefully designed filter that is concentrated in time as well as in the frequency domain. Due to the sparsity of the signal and the careful selection of the number of bins, each bin is likely to only contain one coefficient. After the coefficients of each bin are obtained the actual positions in the frequency domain are recovered by *locating* and *estimating*. The algorithm does this hashing twice and "encodes" the frequency of the coefficient into the phase difference between the two hashed coefficients. This technique achieves the *locating* part of the algorithm by decoding the phase and obtaining the frequency. Before the coefficients are hashed into buckets, the procedure (HashToBins) permutes the signal x in the time domain by applying the permutation operator P which is defined as

$$(P_{\sigma,a,b}x)_i = x_{\sigma(i-a)}\omega^{\sigma bi}, \tag{1}$$

where the parameter b is uniformly random between 1 and N, σ is uniformly random odd between 1 and N, and a is 0 for one hashing operation and 1 for the other. With the use of some basic properties of the Fourier transform the following can be proved (page 5 of [1]):

$$\widehat{P_{\sigma,a,b}x}_{\sigma(i-b)} = \hat{x}_i\omega^{a\sigma i}. \tag{2}$$

Informally, this equation states the following: A permutation, defined by equidistant subsampling in the time domain in addition to a linear phase, results in a permutation in the frequency domain with a linear phase. By carefully choosing the parameters of (2) it is possible to design the permutation such

that the phase difference between the two hashed coefficients is linear in frequency which can then be recovered.

The previous paragraph describes the key ideas of one iteration of the algorithm. A high level overview which was taken from [1] is the following:

- HASHTOBINS permutes the spectrum of $\widehat{x-z}$, then hashes to B bins. Where z is the already recovered signal which is initially all zero.
- NOISELESSSPARSEFFTINNER runs HASHTOBINS twice and *estimates* and *locates* "most" of $\widehat{x-z}$'s coefficients.
- NOISELESSSPARSEFFT iterates NOISELESSSPARSEFFTINNER until it finds $\hat{x}$ exactly.

NOISELESSSPARSEFFTINNER generates the random parameters for the permutation (among others) and passes it to HASHTOBINS. The permutations are $P_{\sigma,0,b}$ for the first call of HASHTOBINS and $P_{\sigma,1,b}$ for the second call respectively. The number of bins is denoted by B and gradually reduced with each call of NOISELESSSPARSEFFTINNER. HASHTOBINS performs an FFT on B samples and thus has a complexity of $O(B \log B)$. By carefully reducing B per iteration the 1D sFFT algorithm runs in time $O(k \log N)$. Again, see [1] for a detailed descriptions of the 1D sFFT algorithm.

III. EXTENSION TO 2D

For simplicity we will reuse the symbols and redefine the notation for the two dimensional case. Let x be an NxN signal with sparsity k, and the number of bins be BxB. It is intuitive to extend the filtering and permutation to two dimensions. However, the fact that the phase difference between the two hashes is always a one dimensional entity even in a 2D sample poses a problem. To be able to recover the frequencies in both dimensions it is necessary to hash a total of three times and encode one dimension in the second and the other dimension in the third call of HASHTOBINS. This allows to *locate* the coefficient in two dimensions. Additionally it is necessary to extend the permutation to 2D which is done with the following definition:

$$\begin{aligned}&(P_{\sigma_x,\sigma_y,\tau_x,\tau_y,a_x,a_y,b_x,b_y}x)_{i_x,i_y}\\ &= x_{\sigma_x i_x+a_x+\tau_x i_y,\sigma_y i_y+a_y+\tau_y i_x}\omega^{-(b_x\sigma_x i_x+b_y\sigma_y i_y)}.\end{aligned} \quad (3)$$

Note that, in addition to extending the permutation to two dimensions a new parameter τ was introduced to allow more powerful permutations. A similar equation as (2), which provides a relationship between the time and frequency domain, can be obtained for the 2D case:

$$\begin{aligned}&\widehat{(P_{\sigma,\tau,\mathbf{a},\mathbf{b}}x)}_{\sigma_x(i_x-b_x)+\tau_x i_y,\sigma_y(i_y-b_y)+\tau_y i_x}\\ &\quad= \hat{x}_{i_x,i_y}\omega^{a_x\sigma_x i_x}\omega^{a_y\sigma_y i_y}.\end{aligned} \quad (4)$$

For the proposed algorithm the high level overview is similar to that previously introduced in Section II. The main difference is within the function NOISELESSSPARSEFFTINNER which needs to properly select the parameters $\sigma_x, \sigma_y, \tau_x, \tau_y$, b_x, b_y and a_x, a_y. For the three calls of HASHTOBINS, a_x, a_y are selected as follows:

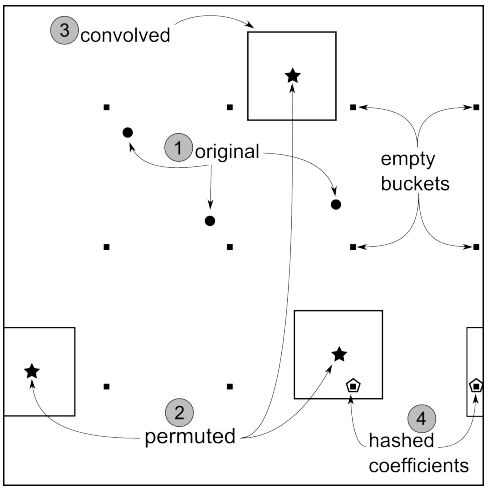


Fig. 1. Graphical depiction of the steps performed in HASHTOBINS. The original spectrum (1) has only three non-zero coefficients ($k = 3$) which are then permuted (2) and convolved with the low pass filter (3). Note that only two coefficients are hashed (4) and the third (a) is missed. There is no collisions in this particular example which could occur if the spectrum overlaps with neighboring coefficients and the area is hashed.

1) $a_x = 0$, $a_y = 0$
2) $a_x = 1$, $a_y = 0$
3) $a_x = 0$, $a_y = 1$.

This approach encodes the frequency of the first dimension in the phase difference of the first and second hashed coefficients and the frequency of the second dimension in the phase difference between the first and third hashed coefficient, respectively. In order to allow the reconstruction of the frequency by inverting the applied permutation, the parameter σ needs to be carefully chosen. In the one dimensional case the constraint for the parameter σ was for it to be odd. For the two dimensional case the following conditions are to be met:

$$\sigma_x \text{ odd}, \sigma_y \text{ odd}, \tau_x \text{ even}, \tau_y \text{ even}$$

or

$$\sigma_x \text{ even}, \sigma_y \text{ even}, \tau_x \text{ odd}, \tau_y \text{ odd}.$$

These constraints ensure that the permutation applied in HASHTOBINS is reversible which is necessary to decode the frequencies in NOISELESSSPARSEFFTINNER.

The newly proposed algorithm has similar parameters as the 1D algorithm. Though, at certain locations the parameters need to be changed two accommodate for two dimensions or extended to two dimensions. An example is the number of bins B which changes to B^2. Since the total number of non-zero efficient is still k, the parameter k occurring in NOISELESSSPARSEFFT is changed to $\sqrt{k}$.

The new procedure NOISELESSSPARSEFFTINNER generates the random parameters according to the constraints laid out above and calls HASHTOBINS three times after which the frequency locations can be recovered and $w_{i_x,i_y} = v$ is performed for "most" of $\widehat{x-z}$ where i_x and i_y are extracted from a combination of the three hashed coefficient and v is taken from the first hash as in the one dimensional case.

Fig. 1 depicts the concept of hashing the coefficients in two dimension.

The proposed algorithm uses the same filter as that introduced in [1] and extends it to two dimensions which is straight forward and is not discussed here.

IV. Optimal parameter selection

In [1], the authors only considered signals with random spectra. That is, spectra where the k non-zero coefficients have no structure. Often, however, signals encountered in real world applications are structured. For instance, audio signals often carry their majority of energy in harmonic frequencies. Additionally, an image often contains most of its energy in low frequency coefficients around the origin. This structure of Fourier coefficients is the foundation of signal compression where only the major coefficients are kept and low energy coefficients are discarded [7]. In many signal processing applications a randomized algorithm works extremely well [8]. Often, however, it is beneficial to exploit the inherent structure to obtain a better performing algorithm. In the proposed algorithm, if the parameters are chosen randomly, the performance can possibly be very poor which can be seen in Fig. 3.

In particular, in the 1D algorithm of [1] the parameters σ and b are chosen randomly as described after (1). In our proposed algorithm the parameters that need special attention are σ_x, σ_y and τ_x, τ_y. For the remaining of this paper we will assume that we deal with two dimensional data whose spectrum is concentrated around the origin.

The 2D permutation defined in (3) essentially performs a linear mapping of the following form:

$$\begin{pmatrix} i'_x \\ i'_y \end{pmatrix} \mapsto \begin{pmatrix} \sigma_x i_x + \tau_x i_y \\ \sigma_y i_y + \tau_y i_x \end{pmatrix} \tag{5}$$

In this form it is easy to see that σ and τ can be interpreted as scaling and shearing parameters. In particular the scaling is linear in σ_x and σ_y and the shear is linear in $S_x = \tau_x/\sigma_x$ and $S_y = \tau_y/\sigma_y$.

In order to optimize the parameters it is important to know that the low pass filter that is used in the algorithm has an approximately rectangular shape. It is also necessary to understand the inner workings of HashToBins:

First, the spectrum is permuted using the permutation in (3). Note that, if the permutation maps coefficients outside of the valid range of 1 to N, the number is automatically taken modulo N as the discrete Fourier transform is periodic with N. Next the permuted spectrum is convolved with the nearly rectangular two dimensional filter. Eventually the hashes are obtained by evenly subsampling the spectrum. It is important to note, that if the permuted samples are too close together the hashed coefficients can be errornous due to colliding filter windows after the convolution was applied.

Taking the above into consideration it is straight forward to see how the parameters σ and τ can be optimized such that the number of collisions are minimized:

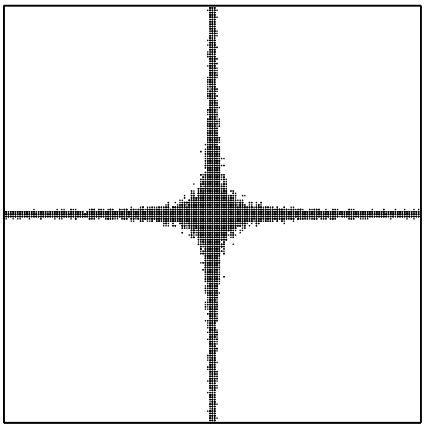


Fig. 2. A 3% sparse spectrum of a 2D Wafer. Note the coefficients are concentrated around the center and principal axes

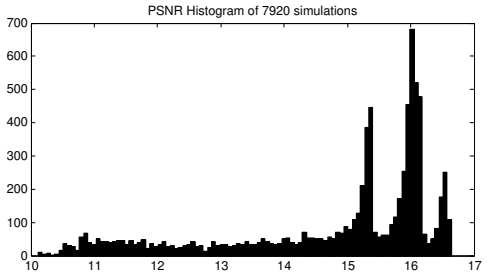


Fig. 3. Histogram of 7920 simulation of a 16384x16384 pixel Wafer with scale parameter S range from 3 to 4201

The first step is to minimize the number of parameters and to set the scale $C = \sigma_x = \sigma_y$ due to the fact that most natural occurring images have similar spectral characteristics along each dimension. An example spectrum is depicted in Fig. 2. Note that the coefficients are concentrated around the center and the principal axes. Secondly, the shear S_x and S_y are set such that $S_x \approx 1$ and $S_y \approx -1$ which can be achieved by setting $\tau_x = \sigma_x - 1$ and $\tau_y = -\tau_x$. This results in an approximately 45° rotation around the origin. This shear is crucial in achieving a low collision rate, as the coefficients along the principal axes would collide with each other without the shear. Furthermore, it is important to choose the scale parameter C carefully. Figure 3 depicts a histogram of a series of simulations where the scale S was swept from 3 to $N/2$. Note that, the PSNR can possibly be very poor if the scale parameter were to be chosen randomly. Instead, our proposed algorithm chooses the scale as $S^* = N/B - 1$ which more consistently results in a good performance in regards to PSNR. The scale S^* is chosen because each bin contains N/B samples and so that it is likely that only one coefficient falls into one bucket since the original coefficients are concentrated prior to the application of the permutation which in turn minimized collisions after the permutation.

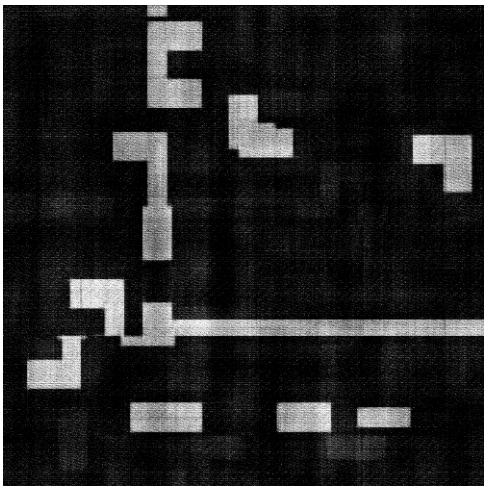


Fig. 4. A cropped 800x800px segment of a "reconstructed" 16384x16384px image of a Wafer. Input was 2% sparse.

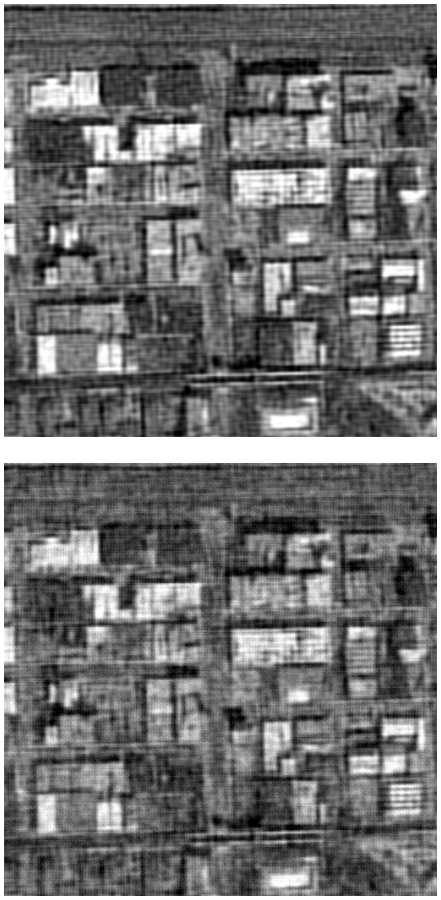


Fig. 5. Top: A 512x512 crop of a 2048x2048 image with a sparsity of 2%. Bottom: Image after running the proposed sFFT algorithm. The PSNR is 23.3dB when compared to the original sparse image.

V. Simulation results

We implemented the proposed algorithm in MATLAB and therefore only simulated the algorithm itself rather than implementing it in C/C++ and measuring real world speedup. Hence, no actual performance comparisons to a C/C++-implementation (such as FFTW) were carried out and the input signal size was limited to 32768^2 due to memory constraints. In order to compare the performance of different parameters, simulations terminated after one outer iteration of the algorithm. An example "reconstructed" image of a Wafer is depicted in Fig. 4. In this case, the Wafer image was sparsified to 3% of the coefficients and then the proposed algorithm was run on the sparse signal.

Figure 5 depicts a 512x512 crop of a 2048x2048 black and white image. The resulting bottom image shows that the proposed 2D sFFT algorithm successfully computed the sparse FFT. First, the original image was loaded, sparsified and then transformed to the spatial domain. This is the top image of Fig. 5. Then the 2D sFFT algorithm was applied to that image followed by an inverse FFT. This is the bottom image which has a PSNR of 23.3dB.

VI. Conclusion

In this paper a new sparse 2D Fourier transform algorithm was introduced. The proposed algorithm is based on the very efficient sFFT algorithm of [1]. The extension to 2D was done by hashing the coefficients into two dimensional buckets and decoding both frequencies from only three hashes. We showed that it is crucial to pay special attention to the parameters σ and τ of the newly introduced permutation, especially when dealing with natural images which usually have the main coefficients around the origin. The result is an algorithm with a time complexity of $O(k \log(N/k) \log^2 N)$ which is similar to the one dimensional algorithm of [1]. Even though, we only considered the optimization of the parameters of the 2D algorithm, the findings can be also be applied to the 1D algorithm when dealing with structured signals such as natural speech.

References

[1] H. Hassanieh, P. Indyk, D. Katabi, and E. Price, "Nearly optimal sparse fourier transform," *CoRR*, vol. abs/1201.2501, 2012.

[2] A. Akavia, S. Goldwasser, and S. Safra, "Proving hard-core predicates using list decoding," in *Annual Symposium on Foundations of Computer Science*, vol. 44. IEEE COMPUTER SOCIETY PRESS, 2003, pp. 146–159.

[3] A. Akavia, "Deterministic sparse fourier approximation via fooling arithmetic progressions," in *Proceedings of the 2010 Conference on Learning Theory, AT Kalai and M. Mohri, eds., Omnipress*, 2010, pp. 381–393.

[4] A. Gilbert, S. Guha, P. Indyk, S. Muthukrishnan, and M. Strauss, "Near-optimal sparse fourier representations via sampling," in *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*. ACM, 2002, pp. 152–161.

[5] M. Iwen, "Combinatorial sublinear-time fourier algorithms," *Foundations of Computational Mathematics*, vol. 10, no. 3, pp. 303–338, 2010.

[6] Y. Mansour, "Randomized interpolation and approximation of sparse polynomials," in *Automata, languages, and programming: 19th international colloquium, Wien, Austria, July 13-17, 1992: proceedings*, vol. 623. Springer, 1992, p. 261.

[7] A. Gersho and R. Gray, *Vector quantization and signal compression*. Springer, 1992, vol. 159.

[8] J. Tropp and A. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *Information Theory, IEEE Transactions on*, vol. 53, no. 12, pp. 4655–4666, 2007.

GESPAR:Efficient Sparse Phase Retrieval with Application to Optics

Yoav Shechtman
Physics Department
Technion, Israel Institute of Technology
joe@tx.technion.ac.il

Amir Beck
Department of Industrial Engineering
Technion, Israel Institute of Technology
becka@ie.technion.ac.il

Yonina C. Eldar
Department of Electrical Engineering
Technion, Israel Institute of Technology
yonina@ee.technion.ac.il

***Abstract*—The problem of phase retrieval, namely, recovery of a signal from the magnitude of its Fourier transform is ill-posed since the Fourier phase information is lost. Therefore, prior information on the signal is needed in order to recover it. In this work we consider the case in which the prior information on the signal is that it is sparse, i.e., it consists of a small number of nonzero elements. We propose GESPAR: A fast local search method for recovering a sparse signal from measurements of its Fourier transform magnitude. Our algorithm does not require matrix lifting, unlike previous approaches, and therefore is potentially suitable for large scale problems such as images. Simulation results indicate that the proposed algorithm is fast and more accurate than existing techniques. We demonstrate applications in optics where GESPAR is generalized and used for finding sparse solutions to sets of quadratic measurements.**

I. Introduction

Recovery of a signal from the magnitude of its Fourier transform, also known as phase retrieval, is of great interest in applications such as optical imaging [1], crystallography [2], and more [3]. Due to the loss of Fourier phase information, the problem (in 1D) is generally ill-posed. A common approach to overcome this ill-posedeness is to exploit prior information on the signal. A variety of methods have been developed that use such prior information, which may be the signal's support, non-negativity, or the signal's magnitude [4], [5]. A popular class of algorithms is based on the use of alternate projections between the different constraints. In order to increase the probability of correct recovery, these methods require the prior information to be very precise, for example, exact/or "almost" exact knowledge of the support set. Since the projections are generally not onto convex sets, convergence to a correct recovery is not guaranteed [6]. A more recent approach is to use matrix-lifting of the problem which allows to recast phase retrieval as a semi-definite programming (SDP) problem [7]. The algorithm developed in [7] does not require prior information about the signal but instead uses multiple signal measurements (e.g., using different illumination settings, in an optical setup).

In order to obtain more robust recovery without requiring multiple measurements, we develop a method that exploits signal sparsity. Existing approaches aimed at recovering sparse signals from their Fourier magnitude belong to two main categories: SDP-based techniques [8],[9],[10] and algorithms that use alternate projections (Fienup-type methods) [11]. Phase retrieval of sparse signals can be viewed as a special case of the more general quadratic compressed sensing (QCS) problem considered in [8]. Specifically, QCS treats recovery of sparse vectors from quadratic measurements of the form $y_i = \mathbf{x}^T\mathbf{A}_i\mathbf{x},\ \ i = 1,\ldots,N$, where $\mathbf{x}$ is the unknown sparse vector to be recovered, y_i are the measurements, and $\mathbf{A}_i$ are known matrices. In (discrete) phase retrieval, $\mathbf{A} = \mathbf{F}_i^T\mathbf{F}_i$ where $\mathbf{F}_i$ is the ith row of the discrete Fourier transform (DFT) matrix.

A general approach to QCS was developed in [8] , in the context of partially incoherent imaging, based on matrix lifting. More specifically, the quadratic constraints where lifted to a higher dimension by defining a matrix variable $\mathbf{X} = \mathbf{x}\mathbf{x}^T$. The problem was then recast as an SDP involving minimization of the rank of the lifted matrix subject to the recovery constraints as well as row sparsity constraints on $\mathbf{X}$. An iterative thresholding algorithm based on a sequence of SDPs was then proposed to recover a sparse solution. Similar SDP-based ideas were recently used in the context of phase retrieval [9],[10]. However, due to the increase in dimension created by the matrix lifting procedure, the SDP approach is not suitable for large-scale problems.

Another approach for phase retrieval of sparse signals is adding a sparsity constraint to the well-known iterative error reduction algorithm of Fienup [11]. In general, Fienup-type approaches are known to suffer from convergence issues and often do not lead to correct recovery especially in 1D problems; simulation results show that even with the additional information that the input is sparse, convergence is still problematic and the algorithm often recovers erroneous solutions.

In this paper we propose an efficient method for phase retrieval which also leads to good recovery performance. Our algorithm is based on a fast 2-opt local search method (see [12] for an excellent introduction to such techniques) applied to a sparsity constrained non-linear optimization formulation of the problem. We refer to our algorithm as GESPAR: GrEedy Sparse PhAse Retrieval. Sparsity constrained nonlinear optimization problems have been considered recently in [13]; the method derived in this paper is motivated – although different in many aspects – by the local search-type techniques of [13]. We demonstrate through numerical simulations that the

proposed algorithm is both efficient and more accurate than current techniques, and we present an example application in optical imaging where a modified version of GESPAR is used.

II. Problem Formulation

We are given a vector of measurements $\mathbf{y} \in \mathbb{R}^N$, that corresponds to the magnitude of an N point discrete Fourier transform of a vector $\mathbf{x} \in \mathbb{R}^N$, i.e.:

$$y_l = \left| \sum_{m=1}^{n} x_m e^{-\frac{2\pi j(m-1)(l-1)}{N}} \right|, \quad l = 1, \ldots, N, \qquad (1)$$

where $\mathbf{x}$ was constructed by zeros padding of a vector $\bar{\mathbf{x}} \in \mathbb{R}^n$ $(n < N)$ with elements x_i, $i = 1, 2, \ldots, n$. In the simulations section we considered the setting $N = 2n$ which corresponds to oversampling the DFT of $\bar{\mathbf{x}}$ by a factor of 2. In any case, we will assume that $N \geq 2n - 1$. This allows to determine the correlation sequence of $\mathbf{x}$ from the given measurements, as we elaborate on more below. Denoting by $\mathbf{F} \in \mathbb{C}^{N \times N}$ the DFT matrix with elements $e^{-\frac{2\pi j(m-1)(l-1)}{N}}$, we can express $\mathbf{y}$ as $\mathbf{y} = |\mathbf{F}\mathbf{x}|$, where $|\cdot|$ denotes the element-wise absolute value. The vector $\mathbf{x}$ is known to be s-sparse on its support, i.e., it contains at most s nonzero elements in the first n elements. Our goal is to recover $\mathbf{x}$ given the measurements $\mathbf{y}$ and the sparsity level s.

The mathematical formulation of the problem that we consider consists of minimizing the sum of squared errors subject to the sparsity constraint:

$$\begin{array}{ll} \min_{\mathbf{x}} & \sum_{i=1}^{N} (|\mathbf{F}_i \mathbf{x}|^2 - y_i^2)^2 \\ \text{s.t.} & \|\mathbf{x}\|_0 \leq s, \\ & supp(\mathbf{x}) \subseteq \{1, 2, \ldots, n\}, \\ & \mathbf{x} \in \mathbb{R}^N, \end{array} \qquad (2)$$

where $\mathbf{F}_i$ is the ith row of the matrix $\mathbf{F}$, $\|\cdot\|_0$ stands for the zero-"norm", that is, the number of nonzero elements. Note that the unknown vector $\mathbf{x}$ can only be found up to trivial degeneracies that are the result of the loss of Fourier phase information: circular shift, global phase, and signal "mirroring".

To aid in solving the phase retrieval problem we will rely on the fact that the correlation sequence of $\bar{\mathbf{x}}$ (the first n components of $\mathbf{x}$) can be determined from $\mathbf{y}$. Specifically, let $g_m = \sum_{i=1}^{n} x_i x_{i+m}, m = -(n-1), \ldots, n-1$ denote the correlation sequence. Note that $\{g_m\}$ is a sequence of length $2n - 1$. Since the DFT length N satisfies $N \geq 2n - 1$, we can obtain $\{g_m\}$ by the inverse DFT of the squared Fourier magnitude $\mathbf{y}$. Throughout the paper, we assume that no support cancellations occur in $\{g_m\}$, namely, if $x_i \neq 0$ and $x_j \neq 0$ for some i, j, then $g_{|i-j|} \neq 0$. When the values of $\mathbf{x}$ are random, this is true with probability 1. This fact is used in the proposed algorithm in order to obtain information on the support of $\mathbf{x}$.

The information on the support is used to derive two sets, J_1 and J_2 from the correlation sequence $\{g_m\}$ in the following manner. Let J_1 be the set of indices known in advance to be in the support, from the autocorrelation sequence. In the noiseless setting which we consider, J_1 comprises two indices:

$$J_1 = \{1, i_{\max}\}.$$

Due to the existing degree of freedom relating to shift-invariance of $\mathbf{x}$, the index 1 can be assumed to be in the support, thereby removing this degree of freedom; as a consequence, the index corresponding to the last nonzero element in the autocorrelation sequence is also in the support, i.e.

$$i_{max} = 1 + \operatorname*{argmax}_i \{i : g_i \neq 0\}.$$

We denote by J_2 the set of indices that are candidates for being in the support, meaning the indices that are *not* known in advance to be in the off-support (the complement of the support). In other words, J_2 contains the set of all indices $k \in \{1, 2, \ldots, n\}$ such that $g_{k-1} \neq 0$. Obviously, since we assume that $x_k = 0$ for $k > n$, we have $J_2 \subseteq \{1, 2, \ldots, n\}$. Defining $\mathbf{A}_i = \Re(\mathbf{F}_i)^T \Re(\mathbf{F}_i) + \Im(\mathbf{F}_i)^T \Im(\mathbf{F}_i) \in \mathbb{R}^{N \times N}$ and $c_i = y_i^2$ for $i = 1, 2, \ldots, N$, problem (2) along with the support information can be written as

$$\begin{array}{ll} \min_{\mathbf{x}} & f(\mathbf{x}) \equiv \sum_{i=1}^{N} (\mathbf{x}^T \mathbf{A}_i \mathbf{x} - c_i)^2 \\ \text{s.t.} & \|\mathbf{x}\|_0 \leq s, \\ & J_1 \subseteq supp(\mathbf{x}) \subseteq J_2, \\ & \mathbf{x} \in \mathbb{R}^N, \end{array} \qquad (3)$$

which will be the formulation to be studied.

In the next section, we propose an iterative local-search based algorithm for solving (3). We note that although in the context of phase retrieval the parameters $\mathbf{A}_i, J_1, J_2$ have special properties (e.g., $\mathbf{A}_i$ is positive semidefinite of at most rank 2, $|J_1| = 2$), we will not use these properties in the proposed method. Therefore, our approach is capable of handling general instances of (3) with the sole assumption that $\mathbf{A}_i$ is symmetric for any $i = 1, 2, \ldots, N$.

III. GReEdy Sparse PhAse Retrieval (GESPAR) Algorithm

In this section GESPAR is summarized. A more detailed description can be found in [14].

A. The Damped Gauss-Newton Method

Before describing the algorithm, we begin by presenting the damped Gauss-Newton (DGN) method [15],[16] that is in fact the core step of our approach. The DGN method is invoked in order to solve the problem of minimizing the objective function f over a *given* support $S \subseteq \{1, 2, \ldots, n\}$ $(|S| = s)$:

$$\min\{f(\mathbf{U}_S \mathbf{z}) : \mathbf{z} \in \mathbb{R}^s\}, \qquad (4)$$

where $\mathbf{U}_S \in \mathbb{R}^{n \times s}$ is the matrix consisting of the columns of the identity matrix $\mathbf{I}_N$ corresponding to the index set S. With this notation, (4) can be explicitly written as

$$\min \left\{ g(\mathbf{z}) \equiv \sum_{i=1}^{N} (\mathbf{z}^T \mathbf{U}_S^T \mathbf{A}_i \mathbf{U}_S \mathbf{z} - c_i)^2 : \mathbf{z} \in \mathbb{R}^s \right\}. \qquad (5)$$

Problem (5) is a nonlinear least-squares problem. A natural approach for tackling it is via the DGN iterations. This algorithm begins with an arbitrary vector $\mathbf{z}_0$. We choose it to be an uncorrelated random Gaussian vector with zero mean and unit variance. At each iteration, all the terms inside the squares in $g(\mathbf{z})$ are linearized around the previous guess. The linearized term is then minimized to determine the next approximation of the solution. Specifically, at each step we pick $\mathbf{y}_k$ to be the solution of

$$\underset{y}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} (\mathbf{z}_{k-1}^T \mathbf{B}_i \mathbf{z}_{k-1} - c_i + 2(\mathbf{B}_i \mathbf{z}_{k-1})^T (\mathbf{y} - \mathbf{z}_{k-1}))^2 \right\},$$

where $\mathbf{B}_i = \mathbf{U}_S^T \mathbf{A}_i \mathbf{U}_S$. This can be written as the linear least squares problem

$$\mathbf{y}_k = \arg\min \|\mathbf{M}\mathbf{y} - \mathbf{b}\|_2^2 \tag{6}$$

with the ith row of $\mathbf{M}$ being $\mathbf{M}_i = 2(\mathbf{B}_i \mathbf{z}_{k-1})^T$, and with $b_i = c_i + \mathbf{z}_{k-1}^T \mathbf{B}_i \mathbf{z}_{k-1}$ for $i = 1, 2, \ldots, N$. The solution $\mathbf{y}_k$ can therefore be calculated explicitly by the pseudo-inverse of $\mathbf{M}$, i.e. $\mathbf{y}_k = (\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^T\mathbf{b}$. We then define a direction vector as $\mathbf{d}_k = \mathbf{y}_k - \mathbf{z}_{k-1}$. This direction is used to update the solution with an appropriate stepsize designed to guarantee the convergence of the method to a stationary point of $g(\mathbf{z})$. The stepsize is chosen via a simple backtracking procedure.

B. The 2-opt Local Search Method

The GESPAR method consists of repeatedly invoking a local-search method on an initial random support set. In this section we describe the local search procedure. At the beginning, the support is chosen to be a set of s random indices chosen to satisfy the support constraints $J_1 \subseteq S \subseteq J_2$. Then, at each iteration a swap between a support and an off-support index is performed such that the resulting solution via the DGN method improves the objective function. Since at each iteration only two elements are changed (one in the support and one in the off-support), this is a so-called "2-opt" method (see [12]). The swaps are always chosen to be between support indices corresponding to components in the current iterate with small absolute value and off-support indices corresponding to large absolute value of ∇f. This process continues as long as the objective function decreases and stops when no improvement can be made.

C. The GESPAR Algorithm

The 2-opt method can have the tendency to get stuck at local optima points. Therefore, our final algorithm, which we call GESPAR, is a restarted version of 2-opt. The 2-opt method is repeatedly invoked with different initial random support sets until the resulting objective function value is smaller than a certain threshold (success) or the number of maximum allowed total number of swaps was passed (failure). A detailed description of the method is given in Algorithm 1. One element of our specific implementation that is not described in Algorithm 1 is the incorporation of random weights added to the objective function, giving randomly different weights to the different measurements.

Algorithm 1 GESPAR

Input: $(\mathbf{A}_i, c_i, \tau, \mathrm{ITER})$.
$\mathbf{A}_i \in \mathbb{R}^{N \times n}, i = 1, 2, \ldots, N$ - symmetric matrices.
$c_i \in \mathbb{R}, i = 1, 2, \ldots, N$.
τ - threshold parameter.
ITER - Maximum allowed total number of swaps.

Output: $\mathbf{x}$ - an optimal (or suboptimal) solution of (3).

Initialization. Set $C = 0, k = 0$.

- **Repeat**
  Invoke the 2-opt method with input $(\mathbf{A}_i, c_i, 4, 8)$ and obtain an output $\mathbf{x}$ and T. Set $\mathbf{x}_k = \mathbf{x}, C = C + T$ and advance k: $k \leftarrow k + 1$.
  Until $f(\mathbf{x}) < \tau$ or $C >$ ITER.
- The output is $\mathbf{x}_\ell$ where $\ell = \mathrm{argmin}_{m=0,1,\ldots,k-1} f(\mathbf{x}_m)$.

IV. Numerical Simulation

In order to demonstrate the performance of GESPAR, we conducted a numerical simulation. The algorithm is evaluated both in terms of signal-recovery accuracy and in terms of computational efficiency.

A. Simulation details

We choose $\bar{\mathbf{x}}$ as a random vector of length n. The vector contains uniformly distributed values in s randomly chosen elements. The N point DFT of the signal is calculated, and its magnitude is taken as $\mathbf{y}$, the vector of measurements. The $2n-1$ point correlation is also calculated. In order to recover the unknown vector $\mathbf{x}$, the GESPAR algorithm is used with $\tau = 10^{-4}$ and $T = 20000$, as well as two other algorithms for comparison purposes: An SDP based algorithm (Algorithm 2, [9].), and an iterative Fienup algorithm with a sparsity constraint [11]. In our simulation $n = 64$ and $N = 128$.

B. Simulation Results

Signal recovery results of the numerical simulation are shown in Fig. 1, where the probability for successful recovery is plotted for different sparsity levels. Successful recovery probability is defined as the ratio of correctly recovered signals $\mathbf{x}$ out of 100 signal-simulations. In each simulation both the support and the signal values are randomly selected. The three algorithms (GESPAR, SDP and Sparse-Fienup) are compared. The results clearly show that GESPAR outperforms the other algorithms in terms of probability of successful recovery - over 90% successful recovery up to $s = 15$, vs. $s = 8$ and $s = 5$ in the other two algorithms.

The average runtime performance of the three algorithms was also compared for several sparsity levels ($s = 3, 5, 8$), and the results are shown in table I. GESPAR is shown to perform much faster than the SDP based method, and comparable in time to the Sparse-Fienup method, while outperforming both in terms of signal recovery.

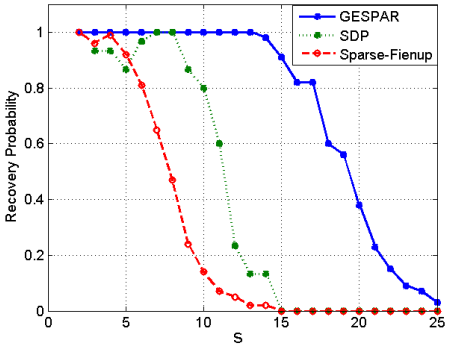


Fig. 1. Recovery probability vs. sparsity (s)

TABLE I
RUNTIME COMPARISON

	SDP	Sparse-Fienup	GESPAR
$s = 3$	1.32 sec	0.09 sec	0.12 sec
$s = 5$	1.78 sec	0.12 sec	0.12 sec
$s = 8$	3.85 sec	0.50 sec	0.23 sec

V. Applications in Optics

As an example of one of the recent applications of GESPAR in optical problems, where it is modified to handle more general quadratic problems, we present Coherent Diffractive Imaging (CDI) for sparsely varying objects. CDI [17] is an imaging method used usually in the x-ray domain, where a small object is illuminated by a coherent plane wave, and the far-field diffraction intensity pattern is measured. The measured intensity corresponds to the 2D Fourier transform of the object. Discretization of the problem followed by appropriate scaling of coordinates yields: $y_i = \mathbf{x}^T \mathbf{A}_i \mathbf{x}, \;\; i = 1, \ldots, N$, where y_i are the far-field intensity measurements, $\mathbf{x}$ is the object to be recovered, and as before - $\mathbf{A}_i = \mathbf{F}_i^T \mathbf{F}_i$. We shall now focus on an example where a dynamic scene is being imaged - e.g. a moving object - so that sequential intensity patterns are being captured at a certain frame rate. If the difference in the object between the consecutive frames $\Delta_k = \mathbf{x}_k - \mathbf{x}_{k-1}$ is sparse (even if the object itself is not) - then recovering the frame difference becomes the problem of finding a sparse solution Δ_k to $y_k^i = (\mathbf{x}_{k-1} + \Delta_k)^T \mathbf{A}^i (\mathbf{x}_{k-1} + \Delta_k)$. Given The result of the previous frame $\mathbf{x}_{k-1}$,this is a quadratic problem in Δ_k, and a modified version of GESPAR is used to solve it. An example recovery is shown in Figure 2- where a comparison to standard frame by frame Fienup HIO [4] recovery without using sparsity is made. In this example there is added noise (SNR=30) and the first frame is assumed to be known (e.g. $\mathbf{y}_0$ is measured with a sufficient number of measurements).

VI. Conclusion

We proposed and demonstrated GESPAR - a fast algorithm to recover a sparse vector from its Fourier magnitude. We showed via simulations that GESPAR outperforms alternative approaches suggested for this problem. The algorithm does not require matrix-lifting, and therefore is potentially suitable for large scale problems such as 2D images, and we demonstrate its application for a more general quadratic imaging problem.

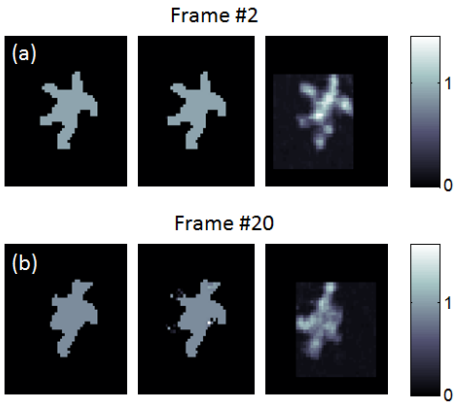


Fig. 2. Sparsely varying CDI example - True object (Left) is being recovered from noisy Fourier magnitude (SNR=30), using sparsity of frame differences (GESPAR - center) and without (Fienup HIO algorithm - right).

References

[1] A. Walther. "The question of phase retrieval in optics". Opt. Acta, 10:41–49, 1963.
[2] R.W. Harrison, "Phase problem in crystallography". J. Opt. Soc. Am. A, 10(5):1045–1055, 1993.
[3] N. Hurt. "Phase retrieval and zero crossings," Kluwer Academic Publishers, Norwell, MA, 1989.
[4] J.R. Fienup, "Phase retrieval algorithms: a comparison," Applied Optics 21, 2758-2769, 1982.
[5] R.W. Gerchberg and W.O. Saxton, "Phase retrieval by iterated projections," Optik 35, 237, 1972.
[6] H.H. Bauschke, P.L. Combettes, and D.R. Luke. "Phase retrieval, error reduction algorithm, and Fienup variants: a view from convex optimization," J. Opt. Soc. Am. A, 19(7):1334–1345, 2002.
[7] E. J. Candes, Y. C. Eldar, T. Strohmer and V. Voroninski, "Phase retrieval via matrix completion," arXiv:1109.0573, Sep. 2011.
[8] Y. Shechtman, Y.C. Eldar, A. Szameit, and M. Segev, "Sparsity based sub-wavelength imaging with partially incoherent light via quadratic compressed sensing," Optics Express 19, 16, 14807-24822, 2011.
[9] K. Jaganathan, S.Oymak, and B. Hassibi, "Recovery of sparse 1-D signals from the magnitudes of their Fourier transform," arXiv:1206.1405v1, June 2012.
[10] H. Ohlsson, A. Y. Yang, R. Dong, S. S. Sastry, "Compressive phase retrieval from squared output measurements via semidefinite programming," arXiv:1111.6323v3 , March 2012.
[11] S. Mukherjee and C. S. Seelamantula, "An iterative algorithm for phase retrieval with sparsity constaints: Application to frequency-domain optical-coherence tomography," ICASSP 2012
[12] C. H. Papadimitriou. *Combinatorial optimization: algorithms and complexity*. Prentice-Hall 1982, with Ken Steiglitz; second edition by Dover, 1998.
[13] A. Beck, Y. C. Eldar, "Sparsity constrained nonlinear optimization: Optimality conditions and algorithms," arXiv:1203.4580v1,
[14] Y. Shechtman et al., "GESPAR: Efficient phase retrieval of sparse signals", arXiv:1301.1018 (2013).
[15] D. P. Bertsekas. *Nonlinear Programming*. Belmont MA: Athena Scientific, second edition, 1999. March 2012.
[16] A. Björck, *Numerical Methods for Least-Squares Problems*, Philadelphia, PA: SIAM, 1996.
[17] J. Miao et al., "Extending the methodology of X-ray crystallography to allow imaging of micrometer-sized non-crystalline specimens," Nature, vol. 400, pp. 342–344, (1999).

Sparse Signal Reconstruction from Phase-only Measurements

Petros T. Boufounos
Mitsubishi Electric Research Laboratories,
Cambridge, MA, 02139
petrosb@merl.com

***Abstract*—We demonstrate that the phase of complex linear measurements of signals preserves significant information about the angles between those signals. We provide stable angle embedding guarantees, akin to the restricted isometry property in classical compressive sensing, that characterize how well the angle information is preserved. They also suggest that a number of measurements linear in the sparsity and logarithmic in the dimensionality of the signal contains sufficient information to acquire and reconstruct a sparse signal within a positive scalar factor. We further show that the reconstruction can be formulated and solved using standard convex and greedy algorithms taken directly from the CS literature. Even though the theoretical results only provide approximate reconstruction guarantees, our experiments suggest that exact reconstruction is possible.**

I. Introduction

The advent of compressive sensing (CS) has significantly improved our ability to sense a variety of signals. Classical CS theory reveals that it is possible to acquire signals at a rate dictated by the complexity of the signal model, rather than the signal dimensionality [1]–[3]. The acquisition is performed using incoherent measurements that preserve all the information in the signal. The signal is recovered from those measurements by exploiting a signal model such as sparsity. Computation—increasingly available thanks to Moore's law—plays an important role in this recovery. Thus it is possible to simplify sensing systems in a number of applications and substitute inexpensive computational complexity in place of frequently expensive sampling complexity.

In this paper we explore how compressive sensing can be used to reconstruct signals from phase-only measurements. Specifically, we demonstrate that the phase of linear complex measurements preserves information about angles of signals. This information can be sufficient to reconstruct the signal within a positive scaling factor. We further show that the measurements contain sufficient information to formulate a convex program or a greedy algorithm to recover the signal.

In many ways, this paper extends earlier work on 1-bit CS, in which a signal is acquired by quantizing the measurements to 1-bit per measurement, i.e. only preserving their signs [4]–[6]. Similar to phase measurements, this operation preserves the angles of signals but not amplitude information. Thus, the signal can only be reconstructed within a scaling factor and only approximated since the measurements are quantized. This paper extends 1-bit CS in the same way that phase/magnitude representations of complex numbers extend sign/magnitude representations of a real numbers.

This work also extends earlier results on the importance of phase information in recovering signals, with a number of practical applications [7]–[10]. In summary, the phase of a fully sampled Fourier transform of a signal contains, under a variety of conditions, sufficient information to uniquely specify the signal and enable its reconstruction within a scaling factor. Our results exploit sparse signal models to reduce the number of phase measurements required. In that sense they transfer classical CS results to phase measurements. While we establish the results using random matrices with i.i.d. normal entries, we conjecture that a large variety of distributions could be used, including subsampled Fourier transforms. Note that quantizing the phase, explored in [11], provides an alternative quantized representation to quantizing the linear measurements.

In the next section we provide a brief background on CS and 1-bit CS, which also partly serves to establish notation. Section III describes the problem, discusses the embedding properties of phase-only measurements and explores how to reconstruct the measured signal. Section IV provides experimental results, validating our approach. Finally, Section V provides some discussion and concludes.

II. Background

A. Compressive Sensing

Classical, by now, results in CS have established that it is possible to measure and successfully reconstruct a signal sparse in some basis using a number of linear measurements which is approximately proportional to the small number of non-zero components of the signal in that basis [1]–[3]. This acquisition can be expressed as the linear system

$$\mathbf{y} = \mathbf{A}\mathbf{x}, \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^N$ denotes the sparse signal, $\mathbf{y} \in \mathbb{R}^M$ denotes the measured data, $\mathbf{A} \in \mathbb{R}^{M \times N}$ denotes the measurement matrix representing the linear system, and M and N denote the dimensionality of the data and the acquired signal, respectively. The sparsity of $\mathbf{x}$, i.e., the number of non-zero coefficients, is denoted using K. We assume, without loss of generality, that the signal is sparse in the canonical basis.

A sufficient condition to recover the signal from the measurements, is the Restricted Isometry Property (RIP). The

matrix $\mathbf{A}$ satisfies the RIP of order K, with RIP constant δ_K if for all K-sparse vectors $\mathbf{x}$:

$$(1-\delta_K)\|\mathbf{x}\|_2 \leq \|\mathbf{A}\mathbf{x}\|_2 \leq (1+\delta_K)\|\mathbf{x}\|_2, \tag{2}$$

i.e., approximately preserves the norm of all K-sparse vectors. Thus, a matrix satisfying the RIP of order $2K$ describes an embedding of K-sparse vectors in N dimensions into an M-dimensional space. This embedding preserves the ℓ_2 distance.

If the RIP of order $2K$ holds with a small RIP constant, the signal can be exactly recovered using the convex program

$$\widehat{\mathbf{x}} = \arg\min_{\mathbf{x}} \|\mathbf{x}\|_1 \text{ s.t. } \mathbf{y} = \mathbf{A}\mathbf{x}, \tag{3}$$

or one of many available greedy algorithms [1], [12]–[16]. Variations of this program, as well as the recovery guarantees have also been developed for a variety of measurement noise conditions and relaxations of the strict sparsity requirement.

The RIP has been established for a variety of matrix classes. With high probability, a properly scaled random matrix with entries generated from an i.i.d. normal or subgaussian distribution satisfies the RIP as long as $M = O(K \log N)$. Similar results have been shown for other matrices, such as ones generated by randomly selecting rows of a DFT matrix.

B. 1-bit Compressive Sensing

Practical acquisition systems quantize their measurements. 1-bit CS examines extreme quantization to one bit per measurement, i.e., preserving only the sign of each measurement:

$$\mathbf{y} = \text{sign}(\mathbf{A}\mathbf{x}), \tag{4}$$

where $\text{sign}(\cdot)$ is applied element-wise to its argument. Since $\text{sign}(\mathbf{A}\mathbf{x}) = \text{sign}(\mathbf{A}c\mathbf{x})$ for all $c > 0$, 1-bit CS acquisition eliminates amplitude information about the signal. Thus, we can only hope to recover the signal within a scaling factor. Furthermore, the solution of an ℓ_1 minimization program similar to (3) degenerates to a zero $\mathbf{x}$. Some way to enforce a norm constrain is necessary [4].

The constraint proposed originally, $\|\mathbf{x}\|_2 = 1$, leads to non-convex program, difficult to analyze and provide guarantees for. More recently, [17] showed that a convex program can be formulated if we exploit the fact that the sign measurements of the signal reveal the hyperoctant in which the measurements lie. Thus a linear constraint can be used to enforce a non-trivial solution, resulting to the convex program

$$\widehat{\mathbf{x}} = \arg\min_{\mathbf{x}} \|\mathbf{x}\|_1 \text{ s.t. } \mathbf{y} = \text{sign}(\mathbf{A}\mathbf{x}) \text{ and } \mathbf{y}^T(\mathbf{A}\mathbf{x}) = 1. \tag{5}$$

This program enforces an ℓ_1 norm constraint by exploiting the fact that $\mathbf{y}^T(\mathbf{A}\mathbf{x}) = \|\mathbf{A}\mathbf{x}\|_1$ at the correct solution.

In the context of 1-bit CS, a condition similar to the RIP can be established, the Binary ϵ-Stable Embedding (BeSE) [6]. The BeSE guarantees the correctness of a sign-consistent reconstruction and characterizes the reconstruction error. The BeSE is in fact an angle embedding, which preserves the angles between signals, defined as

$$d_\angle(\mathbf{x}, \mathbf{x}') = \frac{1}{\pi} \arccos \frac{\langle \mathbf{x}, \mathbf{x}' \rangle}{\|\mathbf{x}\|_2 \|\mathbf{x}'\|_2} \tag{6}$$

for two signals $\mathbf{x}$ and $\mathbf{x}'$. The angle is preserved in the normalized Hamming distance between the measurements, defined as $d_H(\mathbf{y}, \mathbf{y}') = (\sum_i y_i \oplus y_i')/M$, according to

$$d_\angle(\mathbf{x}, \mathbf{x}') - \epsilon \leq d_H(\mathbf{y}, \mathbf{y}') \leq d_\angle(\mathbf{x}, \mathbf{x}') + \epsilon. \tag{7}$$

Thus, if a signal with consistent measurements is found, i.e., $d_H = 0$, it will be within angle ϵ of the measured signal. Similar to the RIP, the BeSE holds for measurement matrices with i.i.d. normal entries, although not in more general ensembles. Furthermore, successful signal recovery from 1-bit measurements with more general ensembles and without requiring the BeSE has also been shown in [18].

III. Phase-Only Compressive Sensing

A. Phase-Only Signal Acquisition

In this paper we consider the following acquisition model

$$\mathbf{z} = \mathbf{A}\mathbf{x}, \ \mathbf{y} = \angle(\mathbf{z}), \tag{8}$$

where $\mathbf{x} \in \mathbb{R}^N$ is a real signal, $\mathbf{A} \in \mathbb{C}^{M \times N}$, $\mathbf{z}$ represents the linear measurement, $\angle(\cdot)$ denotes the principal angle of a complex number, applied element-wise to each vector coefficient, and $\mathbf{y}$ represents the final phase measurements. We also use $\mathbf{a}_m$ to denote the m^{th} row of $\mathbf{A}$.

Obviously, $\angle(\mathbf{A}\mathbf{x}) = \angle(\mathbf{A}c\mathbf{x})$ for any $c > 0$. Thus, angle measurements are similar to sign measurements in 1-bit CS and eliminate any norm information on $\mathbf{x}$. Furthermore, if the acquisition matrix $\mathbf{A}$ only contains real elements, the information in $\mathbf{y}$ is essentially the sign of the measurement—0 and π for positive and negative measurements, respectively. In that case, the problem reverts to 1-bit CS. While complex signals $\mathbf{x}$ can also be considered in this formulation, we defer development of the theory to subsequent work.

B. Stable Angle Embedding

Similar to sign measurements, phase measurements also provide stable embeddings. If two signals $\mathbf{x}, \mathbf{x}'$ in a finite set $\mathcal{W}$ of size L are measured with a random Gaussian vector, the expected value of the measured phase difference is equal to

$$E\left\{\left|\angle\left(\frac{z_m}{z_m'}\right)\right|\right\} = E\left\{\left|\angle\left(e^{i(y_m - y_m')}\right)\right|\right\} = \pi d_\angle(\mathbf{x}, \mathbf{x}'). \tag{9}$$

Hoeffding's inequality bounds the probability that the average of M random variables $|\angle(e^{i(y_m - y_m')})|$ deviates from (9). Using the union bound on L^2 point pairs, a property reminiscent of Johnson-Lindenstrauss (JL) embeddings [19] follows.

Theorem 3.1: Consider a finite set $\mathcal{W} \subset \mathbb{R}^N$ of L points measured using (8), with $\mathbf{A} \in \mathbb{C}^{M \times N}$ consisting of i.i.d elements drawn from the standard complex normal distribution. With probability greater than $1 - 2e^{2\log L - 2\epsilon^2 M}$ the following holds for all $\mathbf{x}, \mathbf{x}' \in \mathcal{S}$ and corresponding measurements $\mathbf{y}, \mathbf{y}' \in \mathbb{R}^M$.

$$\left| \frac{1}{M} \sum_m \left| \frac{1}{\pi} \angle\left(e^{i(y_m - y_m')}\right) \right| - d_\angle(\mathbf{x}, \mathbf{x}') \right| \leq \epsilon \tag{10}$$

Furthermore, the absolute value of the phase difference $\left|\angle\left(e^{i(y_m - y'_m)}\right)\right|$ is Lipschitz continuous with Lipschitz constant equal to 1. Thus, an argument similar to [12] provides a continuous version of the embedding guarantees, similar to the BeSE and the RIP, which is appropriate for sparse signals.

Theorem 3.2: Consider the set $\mathcal{S}_K \subset \mathbb{R}^N$ of all K-sparse signals in $\mathbb{R}^N$, measured as in Thm. 3.1. Eq. (10) holds with probability greater than $1 - 2e^{2K\log\left(\frac{12e}{\epsilon}\frac{N}{K}\right) - \frac{\epsilon^2 M}{2}}$, for all $\mathbf{x}, \mathbf{x}' \in \mathcal{S}_K$ and corresponding measurements $\mathbf{y}, \mathbf{y}' \in \mathbb{R}^M$

These theorems demonstrate that if the mean phase difference between the embedding of two signals is small, then the angle between these signals is also very small. Their nature is similar to the JL lemma, the RIP and the BeSE. They suggest that, similar to classical CS, $M = O(K\log(N/K))$ measurements are sufficient to acquire and reconstruct a signal. The embedding guarantees can be extended to other structured signal sets, such as unions of subspaces or manifolds, using the Kolmogorov complexity of the set in a manner similar to [20].

Unfortunately, the additive form of (10) does not guarantee exact reconstruction. Even if we manage to determine a sparse signal estimate $\widehat{\mathbf{x}}$ that has the same embedding as the measured signal $\mathbf{x}$, Thm. 3.2 can only guarantee that we have identified the signal within an angle ϵ from $\mathbf{x}$, i.e., $|d_\angle(\mathbf{x}, \widehat{\mathbf{x}})| \leq \epsilon$. This behavior is similar to quantized embeddings, such as the BeSE, rather than continuous embeddings such as the RIP. Our experimental results suggest that exact reconstruction guarantees should be possible to derive—not necessarily provided in the form of a stable embedding. However, we do not attempt a proof in this paper.

C. Reconstruction

As discussed above, acquiring a signal using (8) eliminates all information on the total magnitude of the signal. Thus, a reconstruction algorithm, especially one based on ℓ_1-norm minimization, should use a norm constraint to avoid trivial solutions. The original 1-bit CS formulation uses $\|\mathbf{x}\|_2 = 1$, which seems like a natural constraint but leads to a non-convex problem [4]. Instead, we use an approach inspired by the convex formulation in [17].

Specifically, we use the phase of each measurement to rotate that measurement to a positive real number. To do so, we define a vector of unit-magnitude complex coefficients whose phase is equal to the phase of the measurements. Abusing notation, we denote it using $e^{i\mathbf{y}}$, i.e., $(e^{i\mathbf{y}})_m = e^{iy_m}$. Since $e^{-iy_m} z_m = |z_m|$, it follows that $(e^{i\mathbf{y}})^H \mathbf{z} = \|\mathbf{z}\|_1$, where $(\cdot)^H$ denotes the Hermitian (conjugate) transpose. Thus, the convex constraint $(e^{i\mathbf{y}})^H \mathbf{A}\mathbf{x} = 1$ can be used as a norm constraint to prevent degenerate solutions.

In addition to the norm constraint, the phase measurements of a solution should be the same as the original phase measurements. This means that when the linear measurements are properly rotated they should produce positive real numbers: $\Re\{e^{-iy_m} z_m\} \geq 0$ and $\Im\{e^{-iy_m} z_m\} = 0$, where $\Re\{\cdot\}$ and $\Im\{\cdot\}$ denotes the real and the imaginary part, respectively.

Combining all constraints we obtain the following program:

$$\begin{aligned} \widehat{\mathbf{x}} = \arg\min_{\mathbf{x}} & \|\mathbf{x}\|_0 \\ \text{s.t. } & (e^{i\mathbf{y}})^H \mathbf{A}\mathbf{x} = 1, \\ & \Re\left\{e^{-iy_m}\langle \mathbf{a}_m, \mathbf{x}\rangle\right\} \geq 0 \\ \text{and } & \Im\left\{e^{-iy_m}\langle \mathbf{a}_m, \mathbf{x}\rangle\right\} = 0. \end{aligned} \tag{11}$$

Of course, this ℓ_0 minimization can exhibit combinatorial complexity. Thus, (11) can be relaxed to the convex program:

$$\begin{aligned} \widehat{\mathbf{x}} = \arg\min_{\mathbf{x}} & \|\mathbf{x}\|_1 \\ \text{s.t. } & (e^{i\mathbf{y}})^H \mathbf{A}\mathbf{x} = 1, \\ & \Re\left\{e^{-iy_m}\langle \mathbf{a}_m, \mathbf{x}\rangle\right\} \geq 0 \\ \text{and } & \Im\left\{e^{-iy_m}\langle \mathbf{a}_m, \mathbf{x}\rangle\right\} = 0. \end{aligned} \tag{12}$$

Alternatively we can use a greedy algorithm that attempts to find a sparse vector satisfying the constraints. This is the approach we follow in this work. We first define a rotated matrix $\widetilde{\mathbf{A}}$ such that $\widetilde{\mathbf{a}}_m = e^{-iy_m}\mathbf{a}_m$, i.e., such that if the original signal was measured it would produce positive real measurements. This means that the signal should be in the nullspace of the imaginary part of $\widetilde{\mathbf{A}}$. Thus we can attempt to use a greedy algorithm to solve the following optimization:

$$\begin{aligned} \widehat{\mathbf{x}} = \arg\min_{\mathbf{x}} & \left\| \begin{bmatrix} \left(e^{-i\mathbf{y}}\right)^H \mathbf{A} \\ \Im\left\{\widetilde{\mathbf{A}}\right\} \end{bmatrix} \mathbf{x} - \begin{bmatrix} 1 \\ \mathbf{0} \end{bmatrix} \right\|_2^2 \\ \text{s.t. } & \|\mathbf{x}\|_0 \leq K \\ \text{and } & \Re\left\{e^{-iy_m}\langle \mathbf{a}_m, \mathbf{x}\rangle\right\} \geq 0. \end{aligned} \tag{13}$$

This can be solved with straightforward modifications to standard CS algorithms,such as CoSaMP [14], IHT [15], or ALPS [16], to incorporate the positivity constraint on the real part, in a manner similar to the constraints enforcing quantization consistency in [4]–[6]. However, our experimental results showed that the positivity constraint does not contribute significantly to the performance of the system and can be ignored. In this case, the program can be solved using the existing algorithms without any modification. Since a number of implementations of those algorithms expect real matrices as inputs, the complex constraint $(e^{i\mathbf{y}})^H \mathbf{A}\mathbf{x} = 1$ can be implemented as two real constraints $\Re\{(e^{i\mathbf{y}})^H \mathbf{A}\}\mathbf{x} = 1$ and $\Im\{(e^{i\mathbf{y}})^H \mathbf{A}\}\mathbf{x} = 0$. Similarly for the part of the cost function enforcing that constraint in (13).

IV. Experimental Results

To validate the theory we performed experiments in a range of conditions. The results presented are for $N = 1000$ and a variety of K and M, although different values of N exhibited similar behavior. The experiments examined the correlation of the recovered and the measured signals as well as the correct support recovery. Using $\mathbf{x}$ and $\widehat{\mathbf{x}}$ to denote the measured and recovered signals, respectively, the correlation coefficient is

$$\rho = \frac{\langle \mathbf{x}, \widehat{\mathbf{x}}\rangle}{\|\mathbf{x}\|_2 \|\widehat{\mathbf{x}}\|_2} \tag{14}$$

and is equal to 1 if and only if the signal is perfectly recovered. Similarly, using $\mathcal{T}(\cdot)$ to denote the support set, the support recovery can be measured using the ratio

$$P_s = \frac{|\mathcal{T}(\mathbf{x}) \cap \mathcal{T}(\widehat{\mathbf{x}})|}{|\mathcal{T}(\mathbf{x})|}. \quad (15)$$

Note that although perfect signal recovery implies perfect support recovery, the opposite is not true. The support could be perfectly recovered without perfect signal recovery.

For reconstruction we used the very efficient ALPS algorithm [16] to solve (13) without enforcing the positivity constraint $\Re\left\{e^{-iy_m}\langle \mathbf{a}_m, \mathbf{x}\rangle\right\} \geq 0$. The acquisition matrix $\mathbf{A}$ was generated randomly with coefficients drawn from a standard complex normal distribution. The signal $\mathbf{x}$ was generated by first selecting its support set uniformly from the $\binom{N}{K}$ possible sets and then drawing coefficients from a standard normal distribution. The results are averaged over 1500 trials, each with different draw of matrix and signal.

The results are illustrated in Fig. 1. The left plot shows the average correlation as a function of the number of measurements for different values of K. Similarly, the right plot shows the fraction of support recovered as a function of the number of measurements. As evident from the results, the recovery performance exhibits similar behavior to classical compressive sensing. The recovery fails if there is an insufficient number of measurements and the performance exhibits a rapid phase transition as the number of measurements increase. Once a sufficient number of measurements is obtained the signal is perfectly recovered.

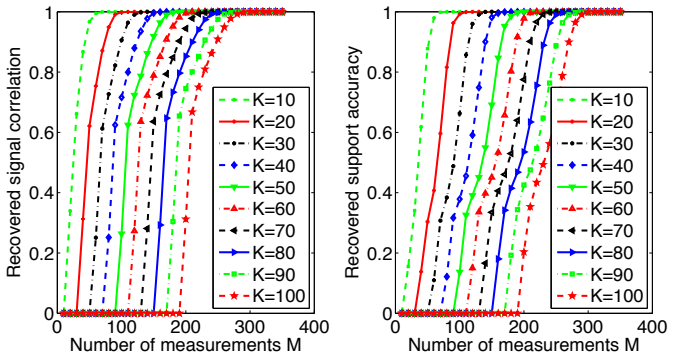


Fig. 1. Reconstruction from phase measurements. Left: average correlation of reconstructed signal with measured signal. Right: probability of correct support recovery in the reconstruction.

V. Discussion and Conclusion

In summary, we demonstrated that the phase of complex measurements contains sufficient information to fully reconstruct a sparse signal within a scaling factor. The theory we present demonstrates that two sparse signals with similar measurements also have very high correlation. Unfortunately, the stable angle embeddings we establish do not guarantee exact reconstruction, even if the phase measurements of the reconstructed signal are identical to those of the measured signal. The small error ϵ characterizes the worst-case reconstruction ambiguity. However, the experimental results suggest that Thm. 3.2 can be tightened to guarantee exact reconstruction

We should also note that the theorem does not guarantee that reconstruction is computationally tractable. The program in (11) will recover the signal if $\mathbf{A}$ provides a stable angle embedding. However, a stable angle embedding does not guarantee that the relaxations in (12) and (13) also recover the correct signal. In that sense, a stable angle embedding is not equivalent to the RIP. The latter has a dual role: In addition to its function as an embedding, the RIP also guarantees that ℓ_1 relaxation and greedy algorithms do provide an exact solution, robust to noise and sparsity level. Whether stable angle embeddings can provide such guarantees is still open.

References

[1] E. Candès, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Comm. Pure and Appl. Math.*, vol. 59, no. 8, pp. 1207–1223, 2006.

[2] D. Donoho, "Compressed sensing," *IEEE Trans. Inform. Theory*, vol. 6, no. 4, pp. 1289–1306, 2006.

[3] E. Candès, "Compressive sampling," in *Proc. Int. Congress Math.*, Madrid, Spain, Aug. 2006.

[4] P. Boufounos and R. Baraniuk, "1-bit compressive sensing," in *Proc. Conf. Inform. Science and Systems (CISS)*, Princeton, NJ, Mar. 2008.

[5] P. Boufounos, "Greedy sparse signal reconstruction from sign measurements," in *Proc. Asilomar Conf. on Signals Systems and Comput.*, Asilomar, California, Nov. 2009.

[6] L. Jacques, J. N. Laska, P. T. Boufounos, and R. G. Baraniuk, "Robust 1-bit compressive sensing via binary stable embeddings of sparse vectors," *Trans. Info. Thery*, 2013, to appear.

[7] M. Hayes, J. Lim, and A. Oppenheim, "Signal reconstruction from phase or magnitude," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 28, no. 6, pp. 672 – 680, dec 1980.

[8] J. Quatieri, T. and A. Oppenheim, "Iterative techniques for minimum phase signal reconstruction from phase or magnitude," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 29, no. 6, pp. 1187 – 1193, dec 1981.

[9] A. Oppenheim and J. Lim, "The importance of phase in signals," *Proceedings of the IEEE*, vol. 69, no. 5, pp. 529 – 541, may 1981.

[10] A. V. Oppenheim, M. H. Hayes, and J. S. Lim, "Iterative procedures for signal reconstruction from fourier transform phase," *Optical Engineering*, vol. 21, no. 1, pp. 122–127, feb 1982.

[11] P. Boufounos, "Angle-preserving quantized phase embeddings," in *Proc. SPIE Wavelets and Sparsity XV*, San Diego, CA, Aug 26-29 2013, to appear.

[12] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, "A simple proof of the restricted isometry property for random matrices," *Const. Approx.*, vol. 28, no. 3, pp. 253–263, 2008.

[13] E. Candès, "The restricted isometry property and its implications for compressed sensing," *Comptes rendus de l'Académie des Sciences, Série I*, vol. 346, no. 9-10, pp. 589–592, 2008.

[14] D. Needell and J. Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples," *Appl. Comput. Harmon. Anal.*, vol. 26, no. 3, pp. 301–321, 2009.

[15] T. Blumensath and M. Davies, "Iterative hard thresholding for compressive sensing," *Appl. Comput. Harmon. Anal.*, vol. 27, no. 3, pp. 265–274, 2009.

[16] V. Cevher, "On Accelerated Hard Thresholding Methods for Sparse Approximation," Tech. Rep., 2011.

[17] Y. Plan and R. Vershynin, "Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach," *Arxiv preprint arXiv:1202.1212*, 2012.

[18] A. Ai, L. A., Y. Plan, and R. Vershynin, "One-bit compressed sensing with non-gaussian measurements," *Arxiv preprint arXiv:1202.1212*, 2012.

[19] W. Johnson and J. Lindenstrauss, "Extensions of Lipschitz mappings into a Hilbert space," *Contemporary Mathematics*, vol. 26, pp. 189 – 206, 1984.

[20] P. T. Boufounos, "Universal rate-efficient scalar quantization," *IEEE Trans. Info. Theory*, vol. 58, no. 3, pp. 1861–1872, March 2012.

Optimal Sampling Rates in Infinite-Dimensional Compressed Sensing

Mitra Fatemi, Loic Baboulaz and Martin Vetterli
Laboratory of Audiovisual Communications
École Polytechnique Fédérale de Lausanne (EPFL)
Email: {mitra.fatemi, loic.baboulaz, martin.vetterli}@epfl.ch

***Abstract*—The theory of compressed sensing studies the problem of recovering a high dimensional sparse vector from its projections onto lower dimensional subspaces. The recently introduced framework of infinite-dimensional compressed sensing [1], to some extent generalizes these results to infinite-dimensional scenarios. In particular, it is shown that the continuous-time signals that have sparse representations in a known domain can be recovered from random samples in a different domain. The range M and the minimum number m of samples for perfect recovery are limited by a balancing property of the two bases. In this paper, by considering Fourier and Haar wavelet bases, we experimentally show that M can be optimally tuned to minimize the number of samples m that guarantee perfect recovery. This study does not have any parallel in the finite-dimensional CS.**

I. Introduction

Real-world signals are inherently analog or continuous-time and we often observe them through digital measuring devices. Imaging devices such as digital cameras and magnetic resonance imaging (MRI) machines are well known examples that measure light fields and brain signals, respectively. A linear measuring process consists of sampling the signal using certain sampling kernels. The samples of a continuous-time signal f can be regarded as its coefficients in an infinite-dimensional sampling domain $\mathcal{S}$ with a basis made of the sampling kernels. In general, infinite number of samples is required to precisely represent f. By adapting the sampling kernels to a specific type of signal, it is possible to reduce the infinite dimensional representation to a finite one. However, in most of the acquisition devices, the sampling kernels are limited by the physics of the device, and are rarely controllable. Therefore, it is very likely that a finite collection of samples captured by a measuring device result in a poor approximation of the signal.

An approach to reconstructing a satisfactory approximation of the signal is to calculate its coefficients in another domain $\mathcal{R}$ that efficiently represents the class of signals subject to the measurement. This means that any signal f in this class has sparse or fast decaying coefficients in $\mathcal{R}$ and N-term approximations of f in $\mathcal{R}$ rapidly converge to the signal. Wavelets are examples of the representation domains that provide fast converging approximations for piecewise continuous signals with pointwise singularities. Also, piecewise smooth images have compressible coefficients in the curvelet [2] and contourlet [3] domains.

First introduced in [4] and further improved in [5], consistent reconstruction is concerned with the problem of calculating the coefficients of a signal in a domain from its samples in a different domain. The consistent reconstruction method uses N samples in the sampling domain to calculate N coefficients in the reconstruction domain. Adcock and Hansen revisited this problem in [6], [7] and they argued that in general, N samples may not be enough to stably find N coefficients in $\mathcal{R}$. Also, they introduced a new *generalized sampling* (GS) approach to stably recover N coefficients in $\mathcal{R}$ from M samples in $\mathcal{S}$, where usually the stable sampling rate M is larger than N.

With the GS framework, we can perfectly reconstruct the signals that have sparse coefficients in a known domain $\mathcal{R}$ from a finite number of samples. However, similar to the finite-dimensional compressed sensing (CS) [8], [9], we are interested to take advantage of the sparsity of coefficients to reduce the number of samples. This problem can be considered as an infinite-dimensional variant of the CS problem where the goal is to recover a sparse vector $\mathbf{x}$ from linear measurements $\mathbf{y} = U\mathbf{x}$. It is shown that if the sensing matrix U has the so-called *restricted isometry property* (RIP) of order $2k$, any k-sparse vector $\mathbf{x}$ can be uniquely recovered from the measurements $\mathbf{y} = U_{m\times n}\mathbf{x}$ [10]. However, verifying the RIP condition for a matrix is computationally hard. In [11], Candès and Romberg considered orthonormal matrices $U \in \mathbb{R}^{n\times n}$ and they showed that in this case the coherence $\mu(U) = \max_{i,j} u_{i,j}$ can be used to determine the subsampling rate m.

Adcock and Hansen recently extended this idea to GS to address infinite-dimensional compressed sensing [1]. In this theory, a set of k-sparse coefficients in $\mathcal{R}$ with the support of nonzero coefficients in $\{1, ..., N\}$ are recovered with high probability from m samples in $\mathcal{S}$ chosen uniformly at random from the range $\{1, ..., M\}$ by solving the *basis pursuit* problem. The subsampling rate m depends on the coherence of the underlying sensing matrix. In addition, the parameters (N, k, M, m) should satisfy a *balancing condition* (refer to II-B).

The infinite-dimensional CS developed in [1] is a promising framework that allows us to obtain far better approximations of signals and images. However, it is not clear in this theory how the parameters (M, m) change with respect to (N, k) and what are the optimum values of the sampling rate m and

the support range M. In this paper, we study this problem. Specifically, we study the change of m as a function of M for some specific choices of sampling and reconstruction domains and find the optimum values of (M, m) for given values N and k, through the experiments.

The paper is organized as follows. In Section II, we define the problem and briefly review GS and infinite-dimensional CS theories. In Section III, we study the balancing condition in infinite-dimensional CS and discuss the optimum choices of sampling rate and support. Also, we present some experiment results to calculate the optimum values of (M, m) for some given pairs (N, k) when the sampling and reconstruction kernels are Fourier exponentials and Haar wavelets. We use the optimum values calculated in this section to recover the sparse coefficients of different signals in Section IV. Finally, we conclude the paper in Section V.

II. Problem Description

Let $\mathcal{H}$ be a Hilbert space and $\mathcal{S}, \mathcal{R} \subseteq \mathcal{H}$ represent the sampling and reconstruction spaces with the orthonormal bases $\{\psi_j\}_{j=1}^{\infty}$ and $\{\phi_i\}_{i=1}^{\infty}$, respectively. Let $f = \sum_{i=1}^{\infty} \alpha_i \phi_i$ be the signal we wish to recover and suppose that we have access to the collection of samples

$$\beta_1, \beta_2, ... \quad \text{with} \quad \beta_j = \langle f, \psi_j \rangle. \tag{1}$$

The problem throughout this paper is to recover the best approximation of f in terms of $\{\phi_j\}_{j=1}^{\infty}$ from the samples in (1). Equivalently, we seek the best approximation of the coefficients $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, ...]^T$ from measurements $\boldsymbol{\beta} = [\beta_1, \beta_2, ...]^T = U\boldsymbol{\alpha}$, with

$$U = \begin{pmatrix} \langle \phi_1, \psi_1 \rangle & \langle \phi_2, \psi_1 \rangle & \dots \\ \langle \phi_1, \psi_2 \rangle & \langle \phi_2, \psi_2 \rangle & \dots \\ \vdots & \vdots & \ddots \end{pmatrix}. \tag{2}$$

A. Consistent reconstruction and generalized sampling

The consistent reconstruction of f is a point $\hat{f} \in \mathcal{R}$ that generates the same samples $\langle \hat{f}, \psi_j \rangle = \beta_j, \quad j = 1, 2, ...$ If we represent the orthogonal projection of f onto $\mathcal{S}$ by $P_{\mathcal{S}} f = \sum_{j=1}^{\infty} \beta_j \psi_j$, this is equivalent to

$$\hat{f} \in \mathcal{R} : \quad P_{\mathcal{S}} f = P_{\mathcal{S}} \hat{f}. \tag{3}$$

When the two subspaces satisfy $\mathcal{R} \oplus \mathcal{S}^{\perp} = \mathcal{H}$, equation (3) has a unique solution that can be found by solving the infinite-dimensional system of linear equations $U\boldsymbol{\alpha} = \boldsymbol{\beta}$ [4]. Clearly in practice, we have access to a finite number of samples. Therefore, we must consider truncations of this linear system and seek the first N coefficients $\boldsymbol{\alpha}^N$ of $\boldsymbol{\alpha}$. This is equivalent to looking for the N-term approximation of f in $\mathcal{R}$, i.e. $P_{\mathcal{R}_N} f = \sum_{i=1}^{N} \alpha_i \phi_i$.

We may think of solving this problem by taking N samples in $\mathcal{S}$ and considering the consistency condition in the N-dimensional subspace $\mathcal{S}_N$:

$$\hat{f} \in \mathcal{R}_N \quad \text{s.t.} \quad P_{\mathcal{S}_N} \hat{f} = P_{\mathcal{S}_N} f.$$

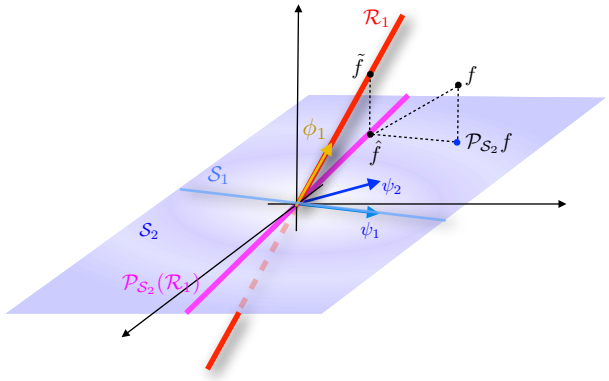


Fig. 1. Generalized sampling reconstruction $\tilde{f}$ of f in $\mathcal{R}_1$ from samples in $\mathcal{S}_2$.

The above equation has a stable solution only if

$$\mathcal{R}_N \oplus \mathcal{S}_N^{\perp} = \mathcal{H}. \tag{4}$$

If we define the angle between two subspaces $\mathcal{R}, \mathcal{S}$ as

$$\cos(\theta_{\mathcal{RS}}) = \inf_{\substack{\mathbf{r} \in \mathcal{R} \\ \|\mathbf{r}\|=1}} \|P_{\mathcal{S}} \mathbf{r}\|,$$

then the condition in (4) is equivalent to $\cos(\theta_{\mathcal{R}_N \mathcal{S}_N}) \neq 0$. In general, this condition may not hold for an arbitrary N, even if the infinite-dimensional spaces satisfy $\mathcal{R} \oplus \mathcal{S}^{\perp} = \mathcal{H}$ [6]. The generalized sampling approach to this problem is to increase the number of samples $M > N$ such that the condition $\cos(\theta_{\mathcal{R}_N \mathcal{S}_M}) \neq 0$ is met. In this case, the projection of $\mathcal{R}_N$ onto $\mathcal{S}_M$ is an N dimensional subspace $P_{\mathcal{S}_M}[\mathcal{R}_N] = span\{P_{\mathcal{S}_M} \phi_i\}_{i=1}^{N}$. Now, we find an approximation of $P_{\mathcal{R}_N} f$ by verifying the consistency condition in this subspace [7]

$$\hat{f} \in \mathcal{R}_N \quad \text{s.t.} \quad P_{P_{\mathcal{S}_M}[\mathcal{R}_N]} \hat{f} = P_{P_{\mathcal{S}_M}[\mathcal{R}_N]} f. \tag{5}$$

Note that $P_{P_{\mathcal{S}_M}[\mathcal{R}_N]} f = P_{P_{\mathcal{S}_M}[\mathcal{R}_N]} P_{\mathcal{S}_M} f$ can be derived from the samples.

In Figure 1, we try to explain the GS reconstruction through an example in $\mathbb{R}^3$. In this example, we find the approximation of f in $\mathcal{R}_1$ from two samples in $\mathcal{S}_2$. Note, that since $\mathcal{R}_1$ is orthogonal to $\mathcal{S}_1 = span\{\psi_1\}$, one sample of f in $\mathcal{S}_1$ is not sufficient for the stable approximation of f in $\mathcal{R}_1$.

The solution of the GS equation in (5) is a stable approximation of f in $\mathcal{R}_N$ and it satisfies

$$\|f - \hat{f}\| \leq \frac{1}{\cos(\theta_{\mathcal{R}_N \mathcal{S}_M})} \|f - P_{\mathcal{R}_N} f\|.$$

Also, the coefficients of $\hat{f}$ can be calculated as $\boldsymbol{\alpha}^N = ((U^{M,N})^* U^{M,N})^{-1} (U^{M,N})^* \boldsymbol{\beta}^M$, where $U^{M,N}$ is the $M \times N$ subsection of U.

B. Infinite-dimensional compressed sensing

Now, assume that the coefficients $\boldsymbol{\alpha}$ are k-sparse with a support $\Delta \in \{1, ..., N\}$. In this case, we can perfectly recover f from equation (5). The infinite-dimensional CS approach in [1] exploits the sparsity to reduce the number of samples.

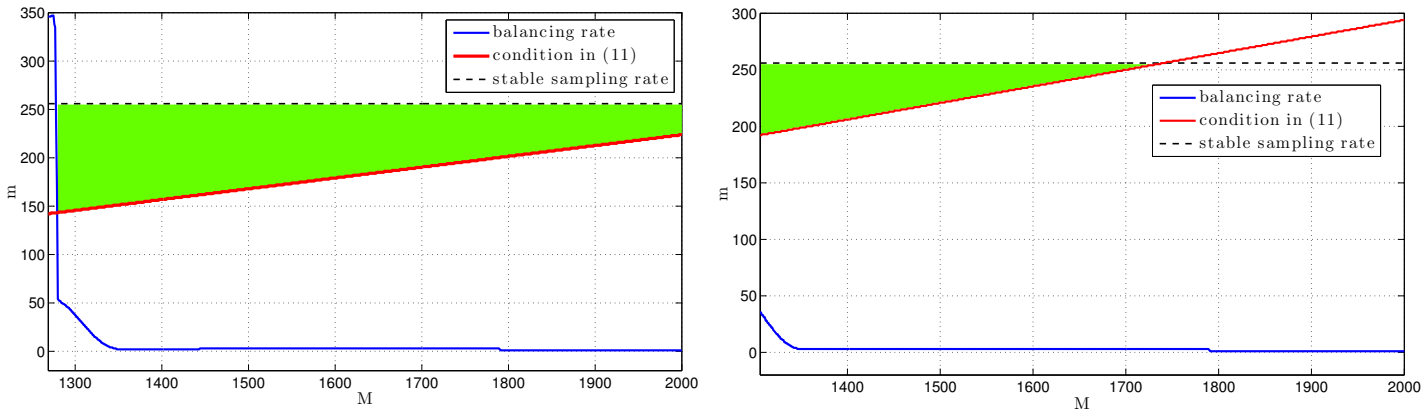


Fig. 2. The acceptable range of sampling rate m and sampling support M for samples in the Fourier domain and sparse coefficients in the Haar domain, $N = 200$ and (a) $k = 30$, (b) $k = 40$. The blue and red plots display the minimum values of m as a function of M that are dictated by the balancing property and the equation (7) with $\epsilon = 0.05$, respectively. The black lines show the stable sampling rate in GS. The green regions display the acceptable ranges of (M, m).

The price of the subsampling, however, is to trade the stable recovery in GS with a probabilistic recovery.

Before we recall the main results in [1] for recovery of sparse or compressible signals in $\mathcal{R}$, we need to define the balancing property.

Definition 1. *Let U be the isometry matrix in* (2)*. Then M and m satisfy the balancing property with respect to U, N and k if*

$$\|(U^{M\times N})^* U^{M\times N} - I_{N\times N}\| \le \left(4\sqrt{\log_2(4M\sqrt{k}/m)}\right)^{-1},$$

$$\|(U^{M\times N})^* U^{M\times N} - diag((U^{M\times N})^* U^{M\times N})\|_{mr} \le \frac{1}{8\sqrt{k}},$$

where $\|U\|_{mr}$ denotes the maximum ℓ^2 norm of different rows of U.

Theorem 1. *Let U be an isometry matrix with the coherence $\mu(U) = \max_{i,j}|u_{i,j}|$. Let the coefficients $\boldsymbol{\alpha} \in \ell^1(\mathbb{N})$ in $\mathcal{R}$ can be written as $\boldsymbol{\alpha} = \boldsymbol{\alpha}_0 + \boldsymbol{\alpha}_1$ with $\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1 \in \ell^1(\mathbb{N})$ and $supp(\boldsymbol{\alpha}_0) = \Delta \subset \{1, ..., N\}$ and $supp(\boldsymbol{\alpha}_1) = \{1, ..., N\}$. Also let $\epsilon > 0$ and $\Omega \subset \{1, ..., M\}$ be chosen uniformly at random with $|\Omega| = m$. If $\boldsymbol{\beta} = U\boldsymbol{\alpha}$ and $\hat{\boldsymbol{\alpha}}$ is a minimizer of*

$$\inf_{\boldsymbol{\eta}\in\ell^1(\mathbb{N})} \|\boldsymbol{\eta}\|_{\ell^1} \quad s.t. \quad U_\Omega^{M\times N}\boldsymbol{\eta}^N = \boldsymbol{\beta}_\Omega, \tag{6}$$

then with probability exceeding $1-\epsilon$ we have

$$\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\| \le (\frac{20M}{m} + 11 + \frac{m}{2M})\|\boldsymbol{\alpha}_1\|_{\ell^1},$$

given that $(N, |\Delta|, M, m)$ satisfy the balancing property and m satisfies

$$m \ge CM\mu^2(U)|\Delta|(\log(\epsilon^{-1}) + 1)\log\left(\frac{MN\sqrt{|\Delta|}}{m}\right), \tag{7}$$

for a universal constant C.

In case that $\boldsymbol{\alpha}_1 = 0$ and $\boldsymbol{\alpha}$ is a k-sparse vector with $k = |\Delta|$, the equation (6) *has a unique solution that coincides with $\boldsymbol{\alpha}$ with probability greater than $1-\epsilon$.*

III. Optimal Sampling Rate

Theorem 1 indicates that a signal with a k-sparse representation in $\mathcal{R}_N$ can be recovered with high probability from m random samples in $\mathcal{S}_M$, if m fulfills the condition in (7) and (N, k, M, m) satisfy the balancing property with respect to U. The condition (7) has a simple structure and we can easily track the change in m based on changes in M, N and k. On the contrary, it is not clear which values of (N, k, M, m) satisfy the balancing property with respect to a given U and how changes in (N, k) affect the sampling rate m and sampling support M. In other words, it is not clear what the subsampling gain of this theory is with respect to the stable sampling rate of GS, for a given sparsity.

In this section, we investigate the balancing property when the underlying sampling and reconstruction domains are formed by Fourier exponentials and Haar wavelet functions in $L^2[0, 1]$. This special choice of basis functions has applications in the MRI problem.

We use the following setup to find efficient sampling rates for fixed pairs of N and k. First, we find all values of M in the range $\{k, k+1, ..., M_{\max}\}$ such that the submatrix $U^{M\times N}$ satisfies the constraint

$$\|(U^{M\times N})^* U^{M\times N} - \text{diag}((U^{M\times N})^* U^{M\times N})\|_{mr} \le \frac{1}{8\sqrt{k}}.$$

The upper bound $M_{\max}$ on the range of samples is usually determined by the sampling device. We point out that in general, the maximum row norm in the above equation does not change monotonically with M. Thus, we should find the acceptable values of M by checking all numbers in $\{k, k+1, ..., M_{\max}\}$.

In the next step, for each verified M, we find the minimum m that satisfies (7) and the first constraint in Definition 1. Finally, we accept the pair (M, m) if $m < \min(M, M_1)$ where M_1 denotes the stable sampling rate in GS corresponding to N.

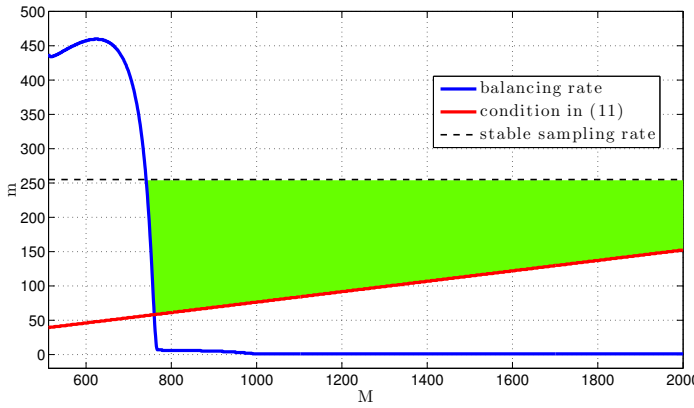


Fig. 3. The acceptable range of sampling rate m and sampling support M for samples in the Haar domain and sparse Fourier coefficients with $N = 200$ and $k = 20$.

Figures 2(a) and 2(b) display the acceptable pairs (M, m) for $N = 200, M_{\max} = 2000$ and two different sparsity values $k = 30, 40$, for sampling in Fourier and reconstruction in Haar domains. Figure 3 depicts the same variables for $k = 20$, when the sampling and sparsity domains are reversed. In these figures, the minimum values of m as a function of M satisfying the balancing property and the equation (7) are indicated in blue and red, respectively. The error probability is $\epsilon = 0.05$. Also, the black lines display the stable sampling rate corresponding to $N = 200$.

The green region in each figure shows the acceptable range of (M, m). The optimal sampling rate is determined by the point in this region that corresponds to the smallest m. For instance, Figure 3 shows that a signal with 20-sparse Fourier coefficients in the range $\{1, ..., 200\}$ can be recovered with probability greater than 0.95 from 58 samples that are chosen uniformly at random from the first 760 coefficients in the Haar domain. This means that we get a large subsampling gain by solving the basis pursuit problem in equation (6). On the contrary, Figure 2(b) illustrates that we do not get too much subsampling gain by replacing the basis pursuit problem in (6) with the stable reconstruction in GS for the specific values of the parameters in this plot.

IV. Numerical Experiments

In this section, we use the optimal values of (M, m) in Figure 2(a) to recover signals having sparse representations in the wavelet domain from randomly chosen Fourier coefficients.

In the first experiment, we consider signals of the form

$$f(t) = \sum_{i=1}^{200} \alpha_i \phi_i(t),$$

with only 20 nonzero coefficients, where $\{\phi_i(t)\}_{i\in\mathbb{N}}$ are Haar wavelets on $[0, 1]$. In the second experiment we consider signals of the form

$$f(t) = \sum_{i=1}^{200} \alpha_{0,i} \phi_i(t) + \sum_{i=1}^{200} \alpha_{1,i} \phi_i(t),$$

TABLE I
The approximation errors for the wavelet coefficients (avg. 100 trials)

	$\|\boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}}\|_{\ell_\infty} / \|\boldsymbol{\alpha}\|_{\ell_\infty}$	SNR
Noiseless coefficients	0.1024×10^{-6}	104 dB
Noisy coefficients	0.7921×10^{-3}	64.1 dB

where the coefficient vector $[\alpha_{0,1}, ..., \alpha_{0,200}]^T$ is 20-sparse and $[\alpha_{1,1}, ..., \alpha_{1,200}]^T$ has a small ℓ_1 norm. For each case, we take $m = 144$ Fourier samples chosen uniformly from the first 1280 Fourier coefficients and we recover the signal by finding the solution to (6). Table I summarizes the approximation errors in the wavelet coefficients. The results in this table are averages over 100 trials.

V. Conclusion

We studied the sampling problem of infinite-dimensional signals that have sparse representations in a known domain. We adopted the random sampling approach of compressed sensing. Unlike the finite-dimensional case, the sampling scheme involves a pair (M, m), where m samples are randomly chosen among a size M subset of possible sampling kernels. For a given setup, there are various pairs which provide high probability of reconstruction. A counter intuitive result is that the required number of samples m does not necessarily decrease as M increases. We experimentally showed that one can find the optimum M that results in the minimum number of samples. We also observed that by swapping the sampling and sparsity domains, the optimal sampling schemes drastically change.

References

[1] B. Adcock and A. Hansen, "Generalized sampling and infinite-dimensional compressed sensing," 2011, preprint.

[2] E. Candès and D. Donoho, "New tight frames of curvelets and optimal representations of objects with piecewise c^2 singularities," *Communications on Pure and Applied Mathematics*, vol. 57, no. 2, pp. 219–266, 2003.

[3] M. Do and M.Vetterli, "The contourlet transform: an efficient directional multiresolution image representation," *IEEE Trans. Image Proc.*, vol. 14, no. 12, pp. 2091–2106, 2005.

[4] M. Unser and A. Aldroubi, "A general sampling theory for nonideal acquisition devices," *IEEE Trans. Signal Proc.*, vol. 42, no. 11, pp. 2915–2925, 1994.

[5] M. Unser and J. Zerubia, "A generalized sampling theory without band-limiting constraints," *IEEE Trans. Circuits Syst. II.*, vol. 45, no. 8, pp. 959–969, 1998.

[6] B. Adcock and A. Hansen, "A generalized sampling theorem for stable reconstruction in arbitrary bases," *Journal of Fourier Analysis and Applications*, pp. 1–32, 2010.

[7] ——, "Sharp bounds, optimality and a geometric interpretation for generalized sampling in hilbert spaces," 2011, preprint.

[8] E. Candes and T. Tao, "Near optimal signal recovery from random projections: Universal encoding strategies," *IEEE Trans. Info Theory*, vol. 52, no. 12, pp. 5406–5425, Dec. 2006.

[9] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[10] E. Candes, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Info Theory*, vol. 52, no. 2, pp. 489–509, 2006.

[11] E. Candes and J. Romberg, "Sparsity and incoherence in compressive sampling," *Inverse Problems*, vol. 23, no. 3, pp. 969–985, 2007.

Deterministic Binary Sequences for Modulated Wideband Converter

Lu Gan
School of Engineering and Design
Brunel University
Uxbridge, UB8 3PH, UK
Lu.Gan@brunel.ac.uk

Huali Wang
Institute of Communication Engineering
PLA University of Science and Technology
Nanjing 210007, P. R. China
wanghl2008@gmail.com

***Abstract*—The modulated wideband converter (MWC) is a promising spectrum blind, sub-Nyquist multi-channel sampling scheme for sparse multi-band signals. In an MWC, the input analog signal is modulated by a bank of periodic binary waveforms, low-pass filtered and then down sampled uniformly at a low rate. One important issue in the design and implementation of an MWC system is the selection of binary waveforms, which impacts the stability of sparse reconstruction. In this paper, we propose to construct the binary pattern with a circulant structure, in which each row is a random cyclic shift of a single deterministic sequence or a pair of complementary sequences. Such operators have hardware friendly structures and fast computation in recovery. They are incoherent with the FFT matrix and the corresponding sampling operators satisfy the restricted isometry property with sub-optimal bounds. Some simulation results are included to demonstrate the validity of the proposed sampling operators.**

I. Introduction

The modulated wideband converter (MWC) proposed by Mishali and Eldar [1], [2] is a multi-channel, uniform sub-Nyquist sampling system for sparse multi-band signals. It holds great potential in applications such as communications, radar and sonar. Consider an analog signal $x(t)$ whose Fourier transform $X(f)$ is bandlimited in $\left[-\frac{f_{NYQ}}{2}, \frac{f_{NYQ}}{2}\right]$ Hz. Assume that $x(t)$ has only K active disjoint frequency bands, each of which has a maximum bandwidth of B Hz. $x(t)$ is said to be a *sparse multi-band signal* if $KB \ll f_{NYQ}$. Figure 1 shows the implementation diagram of an m-channel MWC. In each channel, the input signal is first modulated by a periodic waveform $p_i(t)$, $(i = 0, 1, \cdots m-1)$, low-pass filtered by $h(t)$ and then decimated at the rate of $1/T$ to produce $y_i[n]$. For ease of presentation, we consider the basic configuration of an MWC in which $p_i(t)$ is chosen as the sign alteration waveforms with period of T [1]. Within each sampling period T, there are M intervals of length T/M each and $p_i(t)$ takes the following form [2]

$$p_i(t) = s_{ik}, \quad k\frac{T}{M} \le t \le (k+1)\frac{T}{M} \tag{1}$$

with $s_{ik} \in \{1, -1\}$. Reconstruction of $x(t)$ from $y_i[n]$ $(0 \le i \le m-1)$ exploits the recently emerged compressed sensing theory [3], [4], which searches for the sparsest solution of a parameterized linear equation. Details can be found in [5].

The selection of an $m \times M$ binary pattern $\mathbf{S} = \{s_{ik}\}$ $(0 \le i \le m-1, 0 \le k \le M-1)$ is crucial to the performance of an MWC. From the theoretical perspective, $\mathbf{S}$ needs to offer stable reconstruction performance. From the implementation perspective, it is desirable that $\mathbf{S}$ requires the minimal number of hardware elements with flexible choice of m and M. In [2], $\mathbf{S}$ is constructed from a full random Bernoulli operator. Although such an operator offers near optimal theoretical guarantee, it requires mM flip-flops to implement [2]. To simplify the design, [2] proposed a mixed scheme, in which the first $r < m$ rows of $\mathbf{S}$ are Bernoulli matrices, and the remaining rows are cyclic shifts of them. Such a scheme needs only rM flip-flops. However, the theoretical performance guarantee of these operators is unknown. Besides, simulation results in [2] indicate performance degradation when r is small. In [6], deterministic operators using maximal, Gold and Kasami codes have been used. However, these codes only exist when $M = 2^\beta - 1$ $(\beta \in \mathbb{Z}^+)$, which is not flexible for practical applications.

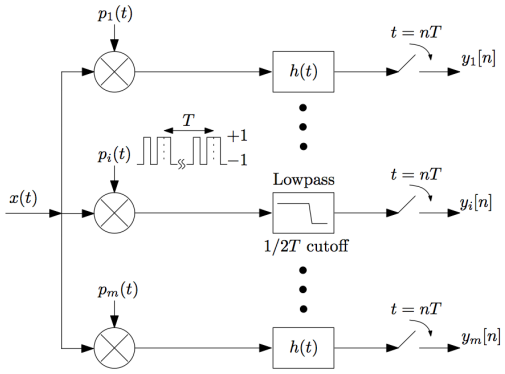


Fig. 1. Implementation diagram of the modulated wideband converter [2].

In this paper, we propose to construct $\mathbf{S}$ with a circulant structure, where each of its row is obtained by random cyclic shift of a single sequence (e.g., the m-sequence or the Legendre sequence) or a pair of cyclic complementary sequences. Due to their circulant structures, the proposed binary patterns are memory efficient with simple hardware implementation. They also offer fast calculation in reconstruction as the matrix multiplication requires only $\mathcal{O}(M \log M)$ operations. Moreover, they exist for a large choice of M. It can be shown that the corresponding sampling operator satisfies the restricted isometry property, which guarantees stable reconstruction in sparse optimization. Experimental results have shown that the proposed binary patterns can offer nearly the same performance as that of the random Bernoulli operator at much lower complexity.

The rest of the paper is organized as follows. In Section II, we briefly review mathematical formulation of the MWC system and related theory in compressed sensing. Section III presents our proposed binary patterns with circulant structure using a single sequence or a pair of complementary sequences. Their restricted isometry properties have been analyzed. Experimental results are shown in Section IV, followed by conclusions in Section V.

Notations: Throughout this paper, vectors are denoted by boldfaced lowercase letters and matrices by boldfaced uppercase characters. If their sizes are not clear from the context, subscripts are provided. For

a matrix $\mathbf{A}$, $\mathbf{A}(i,:)$ denotes its i-th row and $\mathbf{A}_{k,l}$ represent its (k,l)-th element. $\mathbf{A}^T$ and $\mathbf{A}^H$ denote the transpose and the Hermitian transpose of $\mathbf{A}$, respectively. $\mathbf{I}$ is the identity matrix and $\mathbf{F}_M$ is an $M \times M$ FFT matrix with $\mathbf{F}_{k,l} = e^{-j\frac{2\pi kl}{M}}$. For an $M \times M$ matrix $\mathbf{A}$, $\mu(\mathbf{A})$ denotes its coherence parameter, i.e., the maximum magnitude of its elements $\mu(\mathbf{A}) = \max_{0\le k,l\le M-1} |\mathbf{A}_{k,l}|$.

II. Review

Consider an m-channel MWC system in Figure 1. Let $\mathbf{y}[n]$ denote the $m \times 1$ sampled vector

$$\mathbf{y}[n] = \begin{bmatrix} y_0[n] & y_1[n] & \cdots & y_{M-1}[n] \end{bmatrix}^T.$$

Define $\mathbf{y}(f)$ as its discrete-time Fourier transform, i.e., $\mathbf{y}(f) = \sum_{n=-\infty}^{\infty} \mathbf{y}[n] e^{-j2\pi f n T}$. Also, define $z_i(f)$ $(i = 0, \cdots, M-1)$ as a slice of $X(f)$ with bandwidth of $\frac{1}{T}$

$$z_i(f) = X(f + (i - M_0)/T), \quad |f| \le \frac{1}{2T}$$

in which $M_0 = \lfloor M/2 \rfloor$. Let $\mathbf{z}(f)$ denote the $M \times 1$ vector $\mathbf{z}(f) = \begin{bmatrix} z_0(f) & z_1(f) & \cdots & z_{M-1}(f) \end{bmatrix}^T$, the input-output relation in an MWC system can be written as [2]

$$\mathbf{y}(f) = \mathbf{SF}\left(\mathbf{PDz}(f)\right), \quad |f| \le \frac{1}{2T}, \tag{2}$$

where $\mathbf{F}$ is an $M \times M$ FFT matrix, $\mathbf{P}$ is a permutation matrix and $\mathbf{D}$ is diagonal matrix which accounts for the decay of the Fourier transform of $p_i(t)$ at high frequencies. In general, (2) is an under-determined linear equation. But as $X(f)$ is a multi-band sparse signal, $\mathbf{z}(f)$ is a sparse vector with only $K \ll M$ active elements. Based on sparse reconstruction in compressed sensing theory [3], [4], $x(t)$ can be recovered from $\mathbf{y}[n]$ by first identifying the spectral support and then reconstructed using a close-form expression [5].

Note that as $\mathbf{PDz}(f)$ is also a sparse vector with K non-zero elements, we will only focus on the matrix product $\mathbf{SF}$ hereafter. Let us consider the following simplified equation

$$\mathbf{v} = \mathbf{SFu}, \tag{3}$$

in which $\mathbf{u}$ is an $M \times 1$ sparse vector with only K nonzero elements and $\mathbf{v}$ is an $m \times 1$ vector. According to the compressed sensing theory [3], [4], $\mathbf{u}$ can be reconstructed from $\mathbf{v}$ stably when the operator $\Phi = \frac{1}{\sqrt{mM}}\mathbf{SF}$ satisfies the restricted isometry property (RIP):

Definition 1 (RIP): An $m \times M$ matrix $\mathbf{\Phi}$ with normalized columns is said to satisfy the RIP with parameters (K, δ) $(\delta \in (0,1))$ if [3], [4]

$$(1-\delta)\|\mathbf{u}\|^2 \le \|\mathbf{\Phi u}\|^2 \le (1+\delta)\|\mathbf{u}\|^2 \tag{4}$$

for all K-sparse vectors of $\mathbf{u}$.

It is well known if $\mathbf{S}$ is a full-random Bernoulli matrix, then $\mathbf{\Phi} = \frac{1}{\sqrt{mM}}\mathbf{SF}$ satisfies the RIP when $m \ge \mathcal{O}(K \log(M/K))$ [3], [4]. However, full random matrix incurs large memory in storage and high cost in implementation. Another class of operators satisfying the RIP is the randomly subsampled unitary matrix, as presented in the following theorem [7].

Theorem 1 (RIP of a partial unitary matrix): Consider an $m \times M$ matrix $\mathbf{\Phi} = \frac{1}{\sqrt{m}}\mathbf{R}_\Omega \mathbf{U}$, where $\frac{1}{\sqrt{m}}$ is a normalizing coefficient, $\mathbf{R}_\Omega$ is a random sampling operator which selects m samples out of M ones uniformly at random, and $\mathbf{U}$ is an $M \times M$ unitary matrix satisfying $\mathbf{U}^*\mathbf{U} = M\mathbf{I}_M$. $\mathbf{\Phi}$ satisfies the RIP with high probability when [7]

$$M \ge \mathcal{O}\left(\mu^2(\mathbf{U}) K \log^4 M\right), \tag{5}$$

in which $\mu(\mathbf{U})$ represents the maximum magnitude of the elements in $\mathbf{U}$, i.e., $\mu(\mathbf{U}) = \max_{k,l} |\mathbf{U}_{k,l}|$.

Note that the unitary property of $\mathbf{U}$ implies that $1 \le \mu(\mathbf{U}) \le \sqrt{M}$. Hence, when $\mu(\mathbf{U}) = \mathcal{O}(1)$, we can get the sub-optimal bound $M \ge \mathcal{O}(K \log^4 M)$. In the next section, we will develop deterministic binary sequnces for the MWC system based on the above Theorem.

III. Binary Patterns constructed from deterministic sequnces

A. Construction from a single sequence

In this subsection, we consider $\mathbf{S}$ constructed from a partial circulant matrix with the following form

$$\mathbf{S} = \mathbf{R}_\Omega \mathbf{C} \tag{6}$$

where $\mathbf{R}_\Omega$ is a random subsampling operator, which selects m rows out of M ones uniformly at random. $\mathbf{C}$ is a circulant operator that can be expressed as

$$\mathbf{C} = \begin{bmatrix} c_0 & c_1 & \cdots & c_{M-1} \\ c_{M-1} & c_0 & \cdots & c_1 \\ \vdots & \vdots & \ddots & \vdots \\ c_1 & a_2 & \cdots & c_0 \end{bmatrix}, \tag{7}$$

in which $\mathbf{c} = \begin{bmatrix} c_0, c_1, \cdots, c_{M-1} \end{bmatrix}$ is a *deterministic* sequence. According to [2], such a sampling operator can be easily implemented in hardware with only M flip-flops.

It is well known that an $M \times M$ real-coefficient circulant matrix can be factorized into

$$\mathbf{C} = \frac{1}{M}\mathbf{F}\mathrm{diag}(\hat{\mathbf{c}})\mathbf{F}^H, \tag{8}$$

in which $\mathbf{F}$ is the $M \times M$ FFT matrix, and the $1 \times M$ row vector $\hat{\mathbf{c}} = \begin{bmatrix} \hat{c}_0, \hat{c}_1, \cdots, \hat{c}_{M-1} \end{bmatrix}$ is the IFFT of $\mathbf{c}$, i.e., $\hat{\mathbf{c}} = \mathbf{c}\mathbf{F}^H$. Hence, the matrix product $\mathbf{SF}$ can be expressed as

$$\mathbf{SF} = \mathbf{R}_\Omega \mathbf{F}\mathrm{diag}(\hat{\mathbf{c}}). \tag{9}$$

To make use of Theorem 1, $\frac{1}{\sqrt{M}}\mathbf{F}\mathrm{diag}(\hat{\mathbf{c}})$ needs to be a unitary matrix, which implies that each element of $\hat{\mathbf{c}}$ has the same magnitude, i.e., $|\hat{c}_i| = \sqrt{M}$. However, the only known binary sequence with constant FFT magnitudes is $\mathbf{c} = \begin{bmatrix} 1 & 1 & 1 & -1 \end{bmatrix}$ or its cyclic shift. Thus, we consider binary sequences whose FFT coefficients are *nearly* flat. Two popular choices are the maximum length sequence and the Legendre sequence [8]. Specifically,

- **m-sequence**: The maximum length sequence exists for $M = 2^\beta - 1$ $(\beta \in \mathbb{Z}^+)$. It can be easily implemented using β shift registers and has found wide applications in spread-spectrum communications and measurement of impulse response. If $\mathbf{c}$ is a maximum length sequence, then $|\hat{c}_i|$ can be expressed as

$$|\hat{c}_i| = \begin{cases} 1 & i = 0; \\ \sqrt{M+1} & 1 \le i \le M-1. \end{cases} \tag{10}$$

- **Legendre sequence**: A Legendre sequence $\mathbf{c}$ has length M (M prime) and is given by the Legendre symbol [8]

$$c_0 = 1,$$
$$c_i = \begin{cases} 1 & \textit{if } i \text{ is a square (mod } M) \\ -1 & \textit{if } i \text{ is a non-square (mod } M). \end{cases} \quad i > 1 \tag{11}$$

For such a sequence, its IFFT coefficients $\hat{c}_i$ $(0 \le i \le M-1)$ take the form of [8]

$$\hat{c}_0 = 1,$$
$$\hat{c}_i = \begin{cases} 1 + c_i\sqrt{M} & \textit{if } M = 1 \text{ (mod 4)} \\ 1 + jc_i\sqrt{M} & \textit{if } M = 3 \text{ (mod 4)} \end{cases} \tag{12}$$

It is clear that both the maximum length sequence and the Legendre sequence have a (nearly) flat spectrum except for $\hat{c}_0$. By exploiting such a property, we could arrive at the following theorem:

Theorem 2: Consider an sampling operator $\Phi = \frac{1}{\sqrt{mM}}\mathbf{SF}$, in which $\mathbf{F}$ is an $M \times M$ FFT matrix and $\mathbf{S}$ takes the form of (6), where $\mathbf{c}$ is a maximum length sequence or the Legendre sequence. For all K-sparse vector $\mathbf{u} = \left[u_0, u_1, \cdots, u_{M-1}\right]$ with $u_0 = 0$, Eq. (4) holds with high probability provided that $m \geq \mathcal{O}(K \log^4 M)$.

The proof of the above theorem can be achieved by using (10), (12) and Theorem 1. Details are omitted due to lack of space. Note that when $\mathbf{S}$ is constructed from the maximum-length sequence or the Legendre sequence, Theorem 2 implies that stable reconstruction can be achieved as long as $X(f) = 0$ in $|f| < \frac{1}{2T}$. When $X(f)$ is non-zero in $|f| < \frac{1}{2T}$, we can first apply a lowpass filter with cut-off frequency of $\frac{1}{2T}$ to $x(t)$ first and then sample it at the rate of $1/T$. Combined with the samples from MWC, $x(t)$ can then be recovered.

B. Construction from a periodic complementary pair

Both the maximum-length sequence and the Legendre sequence only exist for odd M. In this section, we consider the construction of $\mathbf{S}$ when M is even. To this end, we first present the definition of periodic complementary sequences (PCS) [9]–[11].

Definition 2: For a length-M, real-valued sequence $\mathbf{c} = \left[c_0, c_1, \cdots, c_{M-1}\right]$, its periodic autocorrelation $R_{\mathbf{c}}(l)$ $(0 \leq l \leq M-1)$ is given by

$$R_{\mathbf{c}}(l) = \sum_{k=0}^{M-1} c_k \cdot c_{\bmod (k+l,M)}. \tag{13}$$

Let $\mathbf{a}$ and $\mathbf{b}$ be a pair of length-M bipolar sequences. They are said to form a *periodic complementary pair* (PCP) [9], [11] if

$$R_{\mathbf{a}}(l) + R_{\mathbf{b}}(l) = 0, \quad 1 \leq l \leq M-1. \tag{14}$$

$\mathbf{a}$ (or $\mathbf{b}$) is called as a *periodic complementary sequence* (PCS).

It is known that periodic complementary sequences exist for $M = 2^{\kappa_1}10^{\kappa_2}26^{\kappa_3}$, $M = 2^{\kappa_1}34^{\kappa_2}$ or $M = 2^{\kappa_1}50^{\kappa_2}$ with κ_i $(1 \leq i \leq 3)$ being non-negative integers [11]. It is worth mentioning that a periodic complementary sequence is also *nearly flat* in the FFT domain. To see this, let $\mathbf{a}$ and $\mathbf{b}$ be a PCP and define $\hat{\mathbf{a}} = \mathbf{aF}$ and $\hat{\mathbf{b}} = \mathbf{bF}$. From (14), it can be shown that [9]

$$|\hat{a}_k|^2 + |\hat{b}_k|^2 = 2M, \quad 0 \leq k \leq M-1, \tag{15}$$

in which $\hat{a}_k$ and $\hat{b}_k$ represent the k-th element of $\hat{\mathbf{a}}$ and $\hat{\mathbf{b}}$, respectively. Therefore,

$$|\hat{a}_k| < \sqrt{2M} \quad \text{and} \quad |\hat{b}_k| < \sqrt{2M}, \quad 0 \leq k \leq M-1. \tag{16}$$

In Theorem 3, we will use this property to derive the coherence bound.

We now move on to consider the construction of $\mathbf{S}$ using two circulant cores. Let $\mathbf{a}$ and $\mathbf{b}$ be a PCP of length-$M/2$ and define $\mathbf{A}$ and $\mathbf{B}$ as two $\frac{M}{2} \times \frac{M}{2}$ circulant matrices whose first rows are $\mathbf{a}$ and $\mathbf{b}$, respectively. Eq. (14) implies that an $M \times M$ operator $\mathbf{G}$ given below is a binary orthogonal matrix [12]:

$$\mathbf{G} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & -\mathbf{A}^T \end{bmatrix}. \tag{17}$$

Based on (17), we propose the following binary pattern $\mathbf{S}$:

$$\mathbf{S} = \mathbf{R}_\Omega \mathbf{G} \tilde{\mathbf{P}}, \tag{18}$$

in which $\mathbf{R}_\Omega$ is the same as that in (6), $\mathbf{G}$ is given by (17) and $\tilde{\mathbf{P}}$ is a permutation matrix so that for a vector $\mathbf{c} = \left[c_0, c_1, , \cdots, c_{M-1}\right]$,

$$\begin{aligned} &\left[c_0, \cdots, c_{M/2-1}, c_{M/2}, \cdots c_{M-1}\right] \tilde{\mathbf{P}} \\ &= \left[c_0, c_{M/2}, c_1, c_{M/2+1}, \cdots, c_{M/2-1}, c_{M-1}\right], \end{aligned}$$

i.e., it interleaves the first $M/2$ elements and the last $M/2$ elements of $\mathbf{c}$. The following Lemma presents some properties of the product matrix $\tilde{\mathbf{G}} = \mathbf{G}\tilde{\mathbf{P}}$:

Lemma 1: Consider $\tilde{\mathbf{G}} = \mathbf{G}\tilde{\mathbf{P}}$, in which $\mathbf{G}$ and $\tilde{\mathbf{P}}$ are the same as in (18). $\tilde{\mathbf{G}}$ has the following properties:

- $\tilde{\mathbf{G}}$ is an orthogonal matrix satisfying $\tilde{\mathbf{G}}\tilde{\mathbf{G}} = M\mathbf{I}_M$.
- $\tilde{\mathbf{G}}$ has a circulant structure. Specifically, $\tilde{\mathbf{G}}(k,:)$ and $\tilde{\mathbf{G}}(k + M/2,:)$ $(1 \leq k \leq M/2-1)$ are respectively, the cyclic shift of $\tilde{\mathbf{G}}(0,:)$ and $\tilde{\mathbf{G}}(M/2,:)$ to the right by displacement of $2k$, i.e., the following relations hold

$$\tilde{\mathbf{G}}_{k,l} = \tilde{\mathbf{G}}_{0,\ \bmod\ (2k+l,M)} \tag{19}$$

$$\tilde{\mathbf{G}}_{k+M/2,l} = \tilde{\mathbf{G}}_{M/2,\ \bmod\ (2k+l,M)}. \tag{20}$$

- Each row of $\tilde{\mathbf{G}}$ is a periodic complementary sequence.

Sketch of the proof: The orthogonal property of $\tilde{\mathbf{G}}$ is straightforward due to the orthogonal property of $\mathbf{G}$ and $\tilde{\mathbf{P}}$. The circulant structure of $\tilde{\mathbf{G}}$ can be obtained from the definitions of $\mathbf{G}$ and $\tilde{\mathbf{P}}$. To prove that each row of $\tilde{\mathbf{G}}$ is a PCS, we need the following two facts [11]: (*i*) If $\mathbf{a}$ and $\mathbf{b}$ form a PCP, then their individual cyclic shifts by any displacement l will also produce a PCP; and (*ii*) If $\mathbf{a}$ and $\mathbf{b}$ form a PCP with length of $M/2$, by interleaving them, one can get a new PCS with length of M.

By exploiting Lemma 1, eq.(16) and Theorem 1, the following theorem can be derived:

Theorem 3: Consider an $m \times M$ matrix $\Phi = \frac{1}{\sqrt{mM}}\mathbf{SF}$, in which $\mathbf{S}$ is given by (18) and $\mathbf{F}$ is the $M \times M$ FFT matrix. Φ satisfies the RIP with high probability when $m \geq \mathcal{O}(K \log^4 M)$.

Detailed proof of Lemma 1 and Theorem 3 will be given in the journal version of this paper. Note that due to the structure of $\mathbf{G}$ in (17), only M flip-flops are required to implement $\mathbf{S}$ in (18). Besides, unlike the m-sequence and the Legendre sequence, there is no additional processing of the signal $X(f)$ in $|f| < \frac{1}{2T}$ when $\mathbf{S}$ is constructed from a PCP.

IV. Simulations Results

Extensive simulations have been carried out to evaluate the performance of the proposed binary patterns. Due to lack of space, we only present the results using the Legendre sequence. The experimental setup is very similar to that in [2]. Specifically, the signal $x(t)$ has $f_{NYQ} = 10$ GHz with 3 pairs of active bands (i.e, $K = 6$), each of width $B = 50$ MHz, constructed as follows

$$x(t) = \sum_{i=1}^{3} \sqrt{E_i B}\mathrm{sinc}(B(t-\tau_i))\cos(2\pi f_i(t-\tau_i)), \tag{21}$$

with $\mathrm{sinc}(x) = \sin(\pi x)/(\pi x)$. The energy coefficients are $E_i = \{1, 2, 3\}$ and the time offsets are $\tau_i = \{0.4, 0.7, 0.2\}$. The frequency components f_i is selected uinformly at random from $[f_{NYQ}/2, f_{NYQ}/2]$. In [2], M is selected as 195. Here, we choose $M = 197$, the smallest prime number greater than 195 so that the Legendre sequence can be used. Just as in [2], we assume that $x(t)$ is corrupted by white Gaussian noise and 500 test signals have been evaluated. The reconstruction algorithm is based on that proposed in [5].

We first present the performance of Legendre sequence-based sampling operators for different number of channels with m ranging

from 20 to 100, and different input signal to noise ratio (SNR), ranging from -20 dB to 30 dB. For comparison purposes, the results of full-random binary pattern (i.e., when $\mathbf{S}$ is a Bernoulli matrix) are also included, as shown in Figure 2. One can observe that the proposed sampling operator using the Legendre sequence offers very similar performance to that of the full binary sampling operator at much lower implementation cost. Next, we compare our proposed sampling operators with the mixed scheme proposed in [2]. Specifically, in the mixed scheme, the first r rows are full random Bernoulli operators. Then, the i-th row ($r \leq i \leq m-1$) is five cyclic shifts (to the right) of the $(i-r)$-th row. This mixed scheme requires rM flip-flops, while our proposed sampling operator needs only M ones. Figure 3 presents the reconstruction performance of different binary patterns with $m = 49$ and $M = 197$. As can be seen, the proposed Legendre sequence-based sampling operator provides slightly better reconstruction performance than the full-random sampling operator when the SNR is below 0 dB. On the other hand, the mixed scheme is inferior to the full random sampling operator. In fact, substantial performance loss can be observed when r is small (i.e., $r = 4$). These simulation results demonstrate the effectiveness of using deterministic sequences for an MWC system.

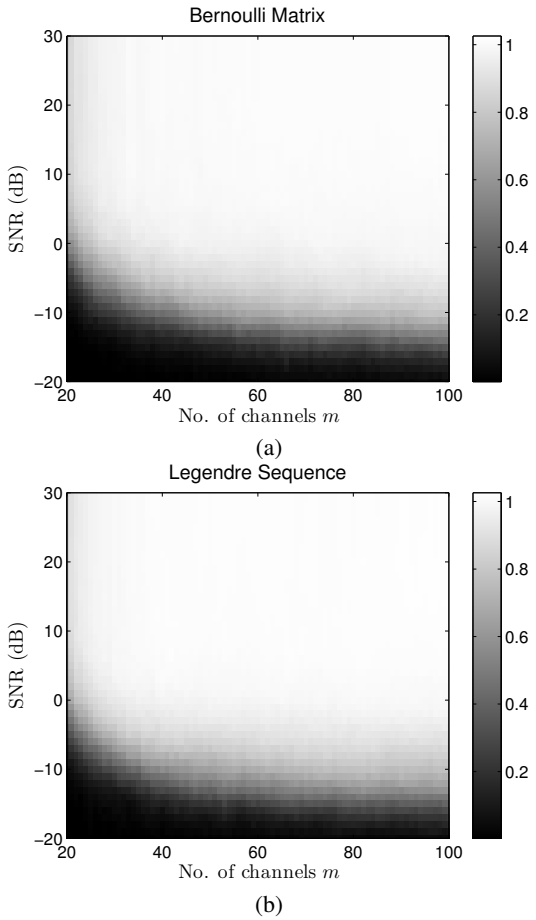


Fig. 2. Probabilities of successful support set recovery for different number of channels m and different SNR levels. (a) Results when $\mathbf{S}$ is a full-random Bernoulli operator. (b) Results when $\mathbf{S}$ is partial circulant matrix in (6) with $\mathbf{c}$ being the Legendre sequence.

V. Conclusions

In this paper, we have proposed to use deterministic sequences for modulated wideband converter in sub-Nyquist sampling of spectrally sparse signals. These include the maximum-length sequence, the Legendre sequence and periodic complementary sequences, all of which have nearly flat spectrum in the (I)FFT domain. The corresponding binary operator $\mathbf{S}$ features hardware friendly implementation, fast computation and near-optimal performance guarantees. Simulation results show that despite their simplicity, the proposed sampling operators can offer very similar performance as that of the full random sampling operators, which imply they are promising in practical applications of the MWC system.

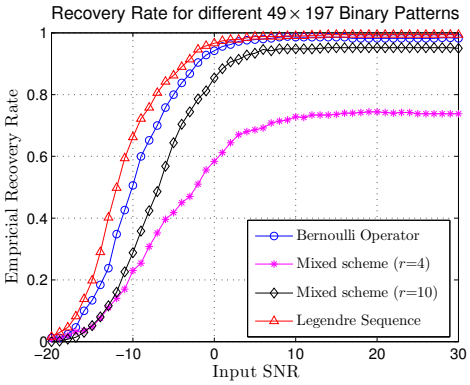


Fig. 3. Successful recovery rate using different 49×197 binary patterns under different input SNR.

References

[1] M. Mishali, Y. C. Eldar, and J. A. Tropp, "Efficient sampling and stable reconstruction of wide band sparse analog signals," in *Proc. 25th IEEE Conv. Electrical and Electronics Engineers in Israel (IEEEI)*, 2008, pp. 290–294.

[2] M. Mishali and Y. Eldar, "From theory to practice: Sub-nyquist sampling of sparse wideband analog signals," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 2, pp. 375–391, Apr. 2010.

[3] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inform. Theory*, vol. 52, pp. 1289–1306, Jul. 2006.

[4] E. Candès and T. Tao, "Near optimal signal recovery from random projections: Universal encoding strategies," *IEEE Trans. Inform. Theory*, vol. 52, pp. 5406–5425, 2006.

[5] M. Mishali and Y. Eldar, "Blind multiband signal reconstruction: Compressed sensing for analog signals," *IEEE Transactions on Signal Processing*, vol. 57, no. 3, pp. 993–1009, Mar. 2009.

[6] M. Mishali and Y. C. Eldar, "Expected RIP: Conditioning of the modulated wideband converter," in *Proc. of IEEE International Information theory workshop*, Oct 2009, pp. 343–347.

[7] M. Rudelson and R. Vershynin, "On sparse reconstruction from Fourier and Gaussian measurements," *Commun. Pure and Applied Math.*, vol. 61, pp. 1025–1045, Dec. 2008.

[8] J. Jensen, H. Jensen, and T. Hoholdt, "The merit factor of binary sequences related to difference sets," *IEEE Trans. Inform. Theory*, vol. 37, no. 3, pp. 617–626, May 1991.

[9] M. Golay, "Complementary series," *IEEE Trans. Inform. Theory*, vol. 7, no. 2, pp. 82 –87, Apr. 1961.

[10] ——, "Seives for low autocorrelation binary sequences," *IEEE Trans. Inform. Theory*, vol. 23, pp. 43–51, Jan. 1977.

[11] D. Ž. Doković, "Periodic complementary sets of binary sequences," in *International Mathematical Forum*, vol. 4, no. 15, 2009, pp. 717–725.

[12] J. Seberry and M.Yamada, "Hadamard matrices, sequences, and block designs," in *Contemporary Design Theory: A Collection of Surveys*, 1992, pp. 431–437.

Acknowledgment

This work was supported in part by the UK EPSRC under Grant EP/I038853/1. The work of Wang was supported by the National Science Foundation of China under Grant 61271354.

Generalized and Fractional Prolate Spheroidal Wave Functions

Ahmed I. Zayed
Department of Mathematical Sciences
DePaul University
Chicago, IL 60614
Email: azayed@depaul.edu

***Abstract*—An important problem in communication engineering is the energy concentration problem, that is the problem of finding a signal bandlimited to $[-\sigma, \sigma]$ with maximum energy concentration in the interval $[-\tau, \tau], 0 < \tau$, in the time domain, or equivalently, finding a signal that is time limited to the interval $[-\tau, \tau]$ with maximum energy concentration in $[-\sigma, \sigma]$ in the frequency domain. This problem was solved by a group of mathematicians at Bell Labs in the early 1960's. The solution involves the prolate spheroidal wave functions which are eigenfunctions of a differential and an integral equations.**

The main goal of this talk is to present a solution to the energy concentration problem in a Hilbert space of functions. This solution will contain as a special case the solution to the energy concentration problem in both the fractional Fourier transform and the linear canonical transform domains. The solution involves a generalization of the prolate spheroidal wave functions, which when restricted to the fractional Fourier transform domain, we may call fractional prolate spheroidal wave functions.

I. Introduction

One of the fundamental problems in communication engineering is the energy concentration problem, that is the problem of finding a signal bandlimited to $[-\sigma, \sigma]$ with maximum energy concentration in the interval $[-\tau, \tau], 0 < \tau$, in the time domain or equivalently, finding a signal that is time limited to the interval $[-\tau, \tau]$ with maximum energy concentration in $[-\sigma, \sigma]$ in the frequency domain. This problem was solved by a group of mathematicians, D. Slepian, H. Landau, and H. Pollak, at Bell Labs [6], [7], [12], [17] in the early 1960's. The solution involves the prolate spheroidal wave functions which are eigenfunctions of a differential and an integral equations.

Because bandlimited functions are entire functions, they cannot vanish outside any interval and as a result the energy concentration in any interval $[-\tau, \tau]$ cannot be 100%. The percentage of the energy concentration depends on σ and τ and involves the eigenvalues of a certain integral equation satisfied by the prolate spheroidal wave functions. The solution of the problem uses properties of the Fourier transform, among them is the fact that the Fourier transform of a prolate spheroidal wave function is a multiple of a scaled version of itself.

Recall that the energy concentration of f in $(-\tau, \tau)$ is given by $\int_{-\tau}^{\tau} |f(t)|^2 \, dt$; therefore, the solution of the concentration problem can be found by finding the function f that maximizes the ratio

$$\alpha^2(\tau) = \frac{\int_{-\tau}^{\tau} |f(t)|^2 \, dt}{\int_{-\infty}^{\infty} |f(t)|^2 \, dt}.$$

A more general problem to consider is the energy concentration problem in the fractional Fourier transform domain. That is to find a signal that is bandlimited to $[-\sigma, \sigma]$ in the fractional Fourier transform domain with maximum energy concentration in the interval $[-\tau, \tau], 0 < \tau$, in the time domain. This problem, in turn, is a special case of the energy concentration problem for the linear canonical transform. The latter problems were solved in [15] and discrete versions of them were solved in [22]. The solutions involved what the authors called the generalized prolate spheroidal wave functions. The generalized prolate spheroidal wave functions associated with the fractional Fourier transform and the linear canonical transform have interesting applications in the analysis of the status of energy preservation of optical systems, self-imaging phenomenon, and the resonance phenomenon of finite-sized one-stage and multiple-stage optical systems [15].

The main goal of this article is to solve the energy concentration problem in a Hilbert space of functions which will contain the fractional Fourier transform and the linear canonical transform as special cases.

A. The Fractional Fourier Transform

The fractional Fourier transform (or FrFT) was first introduced by Namias in 1980 in connection with an application in quantum mechanics [11]. But since its introduction to the signal processing community in the early 1990's, the transform has become an important tool in signal processing applications and signal representation in the fractional Fourier transform domain has been an active area of investigation [1], [3], [4], [5], [8], [9], [10], [13], [14], [19], [20], [21].

The fractional Fourier Transform or FrFT of a signal or a function, say $f(t) \in L^2(\mathbb{R})$, is defined by

$$\widehat{f}_\theta(\omega) = \int_{-\infty}^{\infty} f(t) k_\theta(t, \omega) \, dt \tag{1}$$

where

$$k_\theta(t, \omega) = \begin{cases} c(\theta) \cdot e^{ja(\theta)(t^2+\omega^2) - jb(\theta)\omega t}, \theta \neq p\pi \\ \delta(t-\omega), \theta = 2p\pi \\ \delta(t+\omega), \theta = (2p-1)\pi \end{cases}$$

is the transformation kernel with

$$c(\theta) = \sqrt{(1 - j\cot\theta)/2\pi}, \; a(\theta) = \cot\theta/2, \text{ and } \; b(\theta) = \csc\theta.$$

The kernel $k_\theta(t,\omega)$ is parameterized by an angle $\theta \in I\!R$ and p is some integer. For simplicity, we may write a, b, c instead of $a(\theta), b(\theta)$, and $c(\theta)$. The inverse-FrFT with respect to an angle θ is the FrFT with angle $-\theta$, given by

$$f(t) = \int_{-\infty}^{\infty} \widehat{f_\theta}\,(\omega)\, k_{-\theta}\,(t,\omega)\, d\omega. \tag{2}$$

When $\theta = \pi/2$, (1) reduces to the classical Fourier transform, which will be denoted by $\hat{f}_{\pi/2} = \hat{f}$

$$\hat{f}(\omega) = \frac{1}{\sqrt{2\pi}} \int_{I\!R} f(t) e^{-j\omega t} dt.$$

Let $k_\theta(t,\omega)$ be the kernel of FrFT and define the operator L_θ as

$$L_\theta\,[f]\,(\omega) = \int_{-\infty}^{\infty} f(t) k_\theta(t,\omega) dt.$$

It is easy to see that

$$L_\theta\,(L_\phi\,[f]\,(\omega)) = L_{\theta+\phi}\,[f]\,(\omega).$$

It can be shown that the solution of the concentration problem for the Fractional Fourier transform is the solution of the integral equation (3)

$$\int_{-\sigma}^{\sigma} F(\omega) K_\tau\,(\omega,\zeta)\, d\omega = \lambda F(\zeta), \tag{3}$$

that yields the maximum λ, where

$$K_\tau\,(\omega,\zeta) = \frac{e^{ja\left(t^2-\zeta^2\right)}}{\pi(\omega-\zeta)} \sin b\tau\,(t-\zeta).$$

The solutions of the integral equation (3) share similar properties with the prolate spheroidal wave functions, but satisfy more general differential and integral equations. For lack of better terminology, we shall call these new functions fractional prolate spheroidal wave functions.

B. The Linear Canonical Transform

The linear canonical transform $G_{(a,b,c,d)}(u)$ of a function $f(x)$, which depends on four parameters a, b, c, d, is defined as

$$G_{(a,b,c,d)}(u) \quad = \quad \begin{cases} \int_{-\infty}^{\infty} K_{(a,b,c,d)}(x,u) f(x) dx, & b \neq 0 \\ \sqrt{d} e^{(j/2)cdu^2} f(ud), & b = 0 \end{cases}$$

where

$$K_{(a,b,c,d)} = \frac{1}{\sqrt{2\pi j b}} \exp\left(\frac{j}{2b}\left[du^2 - 2ux + ax^2\right]\right),$$

with $ad - bc = 1$.

For $a = \cos\theta, b = \sin\theta, c = -\sin\theta, d = \cos\theta$, the linear canonical transform reduces to the fractional Fourier transform.

C. The Prolate Spheroidal Wave Functions

The prolate spheroidal wave functions (PSWF), were first discovered in [12] as the bounded eigenfunctions of the following differential operator L_c,

$$L_c\varphi(x) = (1-x^2)\frac{d^2}{dx^2}\varphi(x) - 2x\frac{d}{dx}\varphi(x) - c^2x^2\varphi(x), \tag{4}$$

where $c > 0$ is a real number. In the 1960's, the group at Bell Labs discovered that the following integral operator

$$F_c(\varphi_{n,c})(x) = \int_{-1}^{1} \varphi_{n,c}(t) \frac{\sin(c(x-t))}{\pi(x-t)} dt, \tag{5}$$

commutes with L_c, where $\varphi_{n,c}$ are the eigenfunctions of the operator (4). This commutation relation was termed "a lucky accident" by David Slepian. In a series of papers, the group at Bell Labs employed the commutation relation to derive several properties of the prolate spheroidal wave functions, see [6], [7], [17]. For example, they have showed that the PSWFs satisfy the following integral equation

$$\int_{-1}^{1} \varphi_{n,c}(x) e^{icwx} dx = \mu_n(c)\varphi_{n,c}(w). \tag{6}$$

The PSWFs are normalized so that

$$\|\varphi_{n,c}\|_2^2 = \int_{-\infty}^{+\infty} |\varphi_{n,c}(x)|^2\, dx = 1, \tag{7}$$

or equivalently,

$$\|\varphi_{n,c}\chi_{(-1,1)}\|_2^2 = \int_{-1}^{1} |\varphi_{n,c}(x)|^2\, dx = \lambda_n(c), \tag{8}$$

where $\lambda_n(c)$ is the nth eigenvalue of F_c. The most important properties of the PSWFs are:

(P_1) The set of PSWFs $\{\varphi_{n,c}, n \in I\!N\}$ is an orthogonal basis of $L^2([-1,1])$. More precisely, we have

$$\int_{-1}^{1} \varphi_{n,c}(x)\varphi_{m,c}(x)\, dx = \lambda_n(c)\delta_{mn}.$$

(P_2) The Fourier transform of $\varphi_{n,c}$ is given by :

$$\hat{\varphi}_{n,c}(w) = (-i)^n \sqrt{\frac{2\pi}{c\lambda_n(c)}} \varphi_{n,c}\left(\frac{w}{c}\right) \chi_{[-c,c]}(w).$$

(P_3) The set of PSWFs $\{\varphi_{n,c}, n \in I\!N\}$ is an orthonormal set of $L^2(I\!R)$ and also an orthonormal basis of B_c, where

$$B_c = \{f \in L^2(I\!R) :\ supp(\widehat{f}) \subset [-c,c]\},$$

is the space of functions bandlimited to $[-c,c]$.

II. Conclusion

The main goal of this talk is to show that the energy concentration problem can be solved in a general Hilbert space of functions using the theory of reproducing-kernel Hilbert spaces. To outline the setting in which the problem will be solved, let us introduce the following notation.

Let $\mathcal{E}$ be an arbitrary set and $\mathcal{F}(\mathcal{E})$ be the linear space of all complex-valued functions defined on $\mathcal{E}$. Let $d\mu$ be a σ-finite positive measure and $\mathcal{T}$ be a $d\mu$-measurable set in $I\!R^N$.

Consider the Hilbert space $\mathcal{H} = L^2(\mathcal{T}, d\mu)$ consisting of all complex-valued functions F such that

$$\|F\|^2_{L^2(\mathcal{T},d\mu)} = \int_{\mathcal{T}} |F(t)|^2 d\mu(t) < \infty.$$

Let $h(t,p)$ denote a complex-valued function on $\mathcal{T} \times \mathcal{E}$, such that

$$h(t,p) \in L^2(\mathcal{T}, d\mu) \text{ for any } p \in \mathcal{E}.$$

Let L be the linear mapping $L : L^2(\mathcal{T}, d\mu) \to \mathcal{F}(\mathcal{E})$ defined by

$$f(p) = (LF)(p) = \int_{\mathcal{T}} F(t)\overline{h}(t,p)d\mu(t), \quad F \in L^2(\mathcal{T}, d\mu). \tag{9}$$

It is not difficult to see that the the function

$$K(p,q) = \int_{\mathcal{T}} h(t,q)\overline{h}(t,p)d\mu(t), \tag{10}$$

is positive definite on $\mathcal{E}$, i.e.,

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i\overline{a_j}K(p_i,p_j) \geq 0,$$

for any finite set $\{p_i\}$ of $\mathcal{E}$. Then it follows from [2] that $K(p,q)$ is a reproducing kernel for some Hilbert space of functions defined on $\mathcal{E}$. In fact, the set of all f's given by (9), i.e., the range of the operator L, is a reproducing-kernel Hilbert space $\widetilde{\mathcal{H}}$ whose reproducing kernel is given by (10) so that $f(q) =< f, K(.,q) >_{\widetilde{\mathcal{H}}}$; see [16].

Hereafter, all functions of the form (9) will be called K-bandlimited functions. In this talk we will show that the energy concentration problem can be solved for the class of K-bandlimited functions, but the details will be published somewhere else. The problem will be solved by constructing a sequence of functions that share similar properties to those of the PSWF, in particular Equations (5), (6), and properties P_1 and P_2.

ACKNOWLEDGMENT

The authors would like to thank one of the referees for his constructive comments and bringing to my attention references [15], [22].

REFERENCES

[1] L. B. Almeida, The fractional Fourier transform and time- frequency representations, IEEE Trans. on signal processing, vol.42, no.11 (1994) pp.3084-3091.

[2] N. Aronszajn, Theory of reproducing kernels, *Trans. Amer. Math. Soc.,* Vol. 68 (1950), pp. 337-404.

[3] L. Auslander and R. Tolimier, Radar amiguity functions and group theory, SIAM J. Math. Anal., Vol. 16, pp. 577-601, 1985.

[4] C. Candan, M. A. Kutay, H. M. Ozakdas, The discrete fractional Fourier transform, IEEE Trans. on signal processing, vol. 48. no. 5, 2000.

[5] M. A. Kutay, H. M. Ozakdas, O. Arikan, and L. Onural , Optimal filtering in fractional Fourier domains, IEEE Trans. Signal Processing, vol. 45. pp. 1129-1143, 1997.

[6] H. J. Landau and H. O. Pollak, *Prolate spheroidal wave functions, Fourier analysis and uncertainty-II*, Bell Syst. Techn. J. (1960). pp.65-84.

[7] H. J. Landau and H. O. Pollak, *Prolate spheroidal wave functions, Fourier analysis and uncertainty-III; The dimension of the space of essentially tine and band-limited signals.*, Bell Syst. Techn. J. (1962). pp.1295-1336.

[8] A. W. Lohmann, Image rotation, Wigner rotation and the fractional Fourier transform, J. Opt. Soc. Amer. A. , vol. 10, pp. 2181-2186, 1993.

[9] D. Mendlovic, Z. Zalevsky, and H. M. Ozakdas, The applications of the fractional Fourier transform to optical pattern recognition, in Optical Pattern Recognition, New York: Academic, 1998, Ch. 3.

[10] D. Mendlovich and H. M. Ozaktas, Fractional Fourier transforms and their optical implementation 1, J. Opt. Soc. Amer. A. , vol.10. pp. 1875-1881, 1993.

[11] V. Namias, The fractional order Fourier transforms and its application to quantum mechanics, J. Inst. Math. Appl. , vol. 25, pp. 241-265, 1980.

[12] C. Niven, *On the conduction of heat in ellipsoids of revolution,* Phylosophical Trans. Roy. Soc. Lond., (1880), 171, pp: 117-151.

[13] H. M. Ozakdas, M. A. Kutay, and D. Mendlovic, Introduction to the fractional Fourier transform and its applications, Advances in Imaging Electronics and Physics. New York: Academic, 1999, ch. 4.

[14] H. M. Ozaktas, B. Barshan, D. Mendlovic, and L. Onural, Convolution filtering, and multiplexing in fractional Fourier domains and their relation to chirp and wavelet transforms, J. Opt. Soc. Amer. A. , vol. 11, pp. 547-559, 1994.

[15] S. Pei, J.J Ding, Generalized prolate spheroidal wave functions for optical finite fractional Fourier and linear canonical transform, J. Opt.Soc. Amer. A, Vol. 22, March 2005.

[16] S. Saitoh, *The Theory of Reproducing Kernels and its Applications,* Pitman Research Notes in Math., Vol. 189 (1988), Longman Scientific & Technical, United Kingdom.

[17] D. Slepian and H. O. Pollak, *Prolate spheroidal wave functions, Fourier analysis and uncertainty-I,* Bell System Tech. J., (1961), pp.379-393.

[18] D. Slepian, *Prolate spheroidal wave functions, Fourier analysis and uncertainty–IV: Extensions to many dimensions; generalized prolate spheroidal functions,* Bell System Tech. J. **43** (1964), pp. 3009–3057.

[19] A. I. Zayed, Hilbert transform associated with the fractional Fourier transform, IEEE Signal Processing Lett., vol. 5, pp. 206-208, 1998.

[20] A. I. Zayed, Convolution and product theorem for the fractional Fourier transform, IEEE Signal Processing Lett. , vol. 5, pp. 101-103, 1998.

[21] A. I. Zayed, On the relationship between the Fourier and fractional Fourier transforms, IEEE Signal Processing Lett. , vol. 3, pp. 310-311, 1996.

[22] H. Zhao, R. Wang, D. Song, and D. Wu, Maximally concentrated sequences in both time and linear canonical transform domains, Signal Image Video Process, DOI 10.1007/s11760-012-0309-1.

Absolute Convergence of the Series of Fourier–Haar Coefficients

Boris Golubov
Moscow Institute of Physical Technologies
Institute Lane 9, Dolgoproudny, Moscow region, 141700, Russia
Email: golubov@mail.mipt.ru

Sergey Volosivets
Saratov State University
Astrakhanskaya Str. 83, Saratov, 410012, Russia
Email: VolosivetsSS@mail.ru

***Abstract*—We give some sharp statements on absolute convergence of the series of Fourier-Haar coefficients in terms of L_p- and p-variation best approximations by Haar polynomials.**

INTRODUCTION.

The Haar orthonormal system $\{\chi_n\}_{n=1}^{\infty}$ had been constructed in 1909 (see [1]). By this system A. Haar gave positive answer on the question of D. Hilbert: is there an orthogonal system such that Fourier series with respect to this system of any continuous function converges uniformly to that function?

Let us recall the definition of Haar system. We set $\chi_1(x) = 1$ on $[0,1]$. After that we introduce the open dyadic intervals $I_i^k = \left(2^{-k}(i-1), 2^{-k}i\right)$, $i = 1, ..., 2^k$, $k = 0, 1, ...$, and represent the natural number $n \geq 2$ in the form $n = 2^k + i$, $i = 1, ..., 2^k$, $k = 0, 1, ...$ Then we set $\chi_n(x) = 2^{k/2}$ for $x \in I_{2i-1}^{k+1}$, $\chi_n(x) = -2^{k/2}$ for $x \in I_{2i}^{k+1}$ and $\chi_n(x) = 0$ for $x \in [0,1] \backslash \overline{I_i^k}$, where $\overline{I_i^k}$ is closure of the interval I_i^k. If the Haar function $\chi_n(x)$ has a jump in some poimt $x \in (0,1)$, then $\chi_n(x) = [\chi_n(x-0) + \chi_n(x+0)]/2$. In the end points of interval $[0,1]$ we set $\chi_n(0) = \chi_n(0+0)$ and $\chi_n(1) = \chi_n(1-0)$. The Haar functions $\chi_n(x)$ are step functions.

The principal information on Fourier-Haar series may be found in the book [2].

For a function $f \in L_p[0,1]$, $1 \leq p < \infty$, we introduce the integral modulus of continuity

$$\omega(\delta, f)_p = \sup_{0 \leq h \leq \delta} \left(\int_0^{1-h} |f(t+h) - f(t)|^p \, dt \right)^{1/p}, \quad (1)$$

$0 \leq \delta \leq 1$, and the Fourier-Haar coefficients

$$\hat{f}(n) = \int_0^1 f(x)\chi_n(x)dx, \qquad n \in \mathbf{N}.$$

Z. Ciesielski and J. Musielak [3] proved the following

Theorem A. *Let $\beta > 0$, $\gamma \geq 0$, $p = \max(\beta, 1)$, $f \in L_p[0,1)$, and $\sum_{n=1}^{\infty} n^{\gamma - \beta/2} \omega^{\beta}(1/n, f)_p < \infty$. Then the series $\sum_{n=1}^{\infty} n^{\gamma} \left|\hat{f}(n)\right|^{\beta}$ converges.*

Let us observe that in the paper [3] the authors introduced a slightly different definition of the integral modulus of continuity in the space $L^p[0,1]$, $1 \leq p < \infty$. They extended the function f to the real axis by setting $f(x) = 0$ for $x \notin [0,1]$, and evaluated the integral in the right-hand side of (1) over the interval $[0,1]$. But if we analyze the proof of Theorem 2 from [3], we see that the statement of Theorem A is valid.

Let us define the Wiener's class $V_p[0,1]$, $1 \leq p < \infty$, of functions of bounded p^{th}-power variation on the interval $[0,1]$ (see [4]). We set $f \in V_p[0,1]$, if

$$V(f)_p = \sup_{\tau} \left\{ \sum_{i=1}^{n} |f(x_i) - f(x_{i-1})|^p \right\}^{1/p} < \infty,$$

where $\tau = \{0 = x_0 < x_1 < ... < x_n = 1\}$ is arbitrary partition of the interval $[0,1]$. Let us note that the inclusion $Lip(1/p) \subset V_p[0,1]$ holds for $1 \leq p < \infty$.

P. L. Ulyanov [5] proved the following theorem.

Theorem B. *For the function $f \in V_1[0,1]$ the series*

$$\sum_{n=1}^{\infty} \left|\hat{f}(n)\right|^{\beta} \quad \text{or} \quad \sum_{n=1}^{\infty} n^{\gamma - 1/2} \left|\hat{f}(n)\right| \qquad (2)$$

converge, if $\beta > 2/3$ or $\gamma < 1$ respectively. But this statement does not true for $\beta = 2/3$ or $\gamma = 1$ respectively.

The first author (see [6]) generalized Theorem B to functions $f \in V_p[0,1]$, $1 \leq p < \infty$.

Theorem C. *For the function $f \in V_p[0,1]$, $1 \leq p < \infty$, the series (2) converge, if $\beta > 2p/(p+2)$ or $\gamma < 1/p$. But this statement is not true for $\beta = 2p/(p+2)$ or $\gamma = 1/p$ respectively.*

In the paper [7] a two-dimensional analog of this theorem was proved.

In our paper we give some sharp generalizations of Theorems A and C. We use the weight sequences belonging to the classes $A(\alpha)$, $\alpha \geq 1$. These classes were introduced by L. Gogoladze and R. Meskhia [8].

MAIN RESULTS.

We shall say that the positive sequence $\gamma = \{\gamma_k\}_{k=1}^{\infty}$ belongs to the class $A(\alpha)$, $\alpha \geq 1$, if there is a constant $C > 0$ such that

$$\left(\sum_{k=2^n+1}^{2^{n+1}} \gamma_k^{\alpha} \right)^{1/\alpha} \leq C 2^{n(1-\alpha)/\alpha} \Gamma_n,$$

where

$$\Gamma_n = \sum_{k=2^{n-1}+1}^{2^n} \gamma_k, \quad n \in \mathbf{N}.$$

For $n = 0$ we assume that the above inequality holds for $\Gamma_0 = \gamma_1$.

This definition is a partial case of one introduced by L. Gogoladze and R. Meskhia [8]. It is easy to prove that $A(\alpha_1) \subset A(\alpha_2)$ for $\alpha_1 > \alpha_2 \geq 1$.

Let us recall that for a function $f \in L_p[0,1)$, $1 \leq p < \infty$, the norm is defined by the equality $\|f\|_p = \left(\int_0^1 |f(x)|^p \, dx\right)^{1/p}$. Below we shall use the best approximation $E_n(f)_p = \inf_{\{a_k\}} \|f - t_n\|_p$ of the function $f \in L_p[0,1)$ by Haar polynomials $t_n(x) = \sum_{k=1}^{n} a_k \chi_k(x)$ of order n.

Theorem 1. *Let $f \in L_p[0,1)$, $1 \leq p < \infty$, and*

$$\sum_{k=1}^{\infty} \gamma_k \left[k^{-1/2} E_k(f)_p\right]^{\beta} < \infty, \tag{3}$$

where $0 < \beta < p$, $\gamma \in A(p/(p-\beta))$. Then the series

$$\sum_{n=1}^{\infty} n^{\gamma} \left|\hat{f}(n)\right|^{\beta} \tag{4}$$

converges.

From the Theorem 1 and the inequality $E_n(f)_p \leq 2^{1+1/p}\omega(n^{-1}, f)_p$, $1 \leq p < \infty$, $n \in \mathbf{N}$, (see [5]) it follows

Theorem 2. *The assertion of the Theorem 1 is also valid, if instead of the condition* (3) *we assume* $\sum_{k=1}^{\infty} \gamma_k \left[k^{-1/2}\omega\left(k^{-1}, f\right)_p\right]^{\beta} < \infty$.

For the function $f \in V_p[0,1]$, $1 \leq p < \infty$, we set $\omega_{1-1/p}(\delta, f) = \sup_{\lambda(\tau) \leq \delta} \left\{\sum_{i=1}^{n} |f(x_i) - f(x_{i-1})|^p\right\}^{1/p}$, where $\tau = \{0 = x_0 < x_1 < ... < x_n = 1\}$ is a partition of interval $[0,1]$ and $\lambda(\tau) = \max_{1 \leq i \leq n}(x_i - x_{i-1})$. This notation was introduced in [9]. It is known the inequality $\omega(\delta, f)_p \leq \delta^{1/p}\omega_{1-1/p}(\delta, f)$ for the function $f \in V_p[0,1]$, $1 < p < \infty$ (see [6], Lemma 2 and [9], Theorem 2.5). Therefore from the Theorem 1 it follows

Corollary 1. *If $f \in V_p[0,1]$, $1 < p < \infty$, and*

$$\sum_{k=1}^{\infty} \gamma_k \left[k^{-1/2-1/p}\omega_{1-1/p}(1/k, f)\right]^{\beta} < \infty, \tag{5}$$

where $0 < \beta < p$, $\gamma \in A(p/(p-\beta))$, then the series (4) *converges.*

For the function $f \in V_p[0,1]$, $1 \leq p < \infty$, we define the norm $\|f\|_{V_p} = \max(V_p(f), \|f\|_\infty)$, where $\|f\|_\infty = \sup\{|f(x)| : x \in [0,1]\}$. Let us define the best approximation $E_n(f)_{V_p} = \inf_{\{a_k\}} \|f - t_n\|_{V_p}$ of the function $f \in V_p[0,1]$, $1 \leq p < \infty$, by Haar polynomials $t_n = \sum_{k=1}^{n} a_k \chi_k(x)$ of order n. It is easy to prove the inequality $E_n(f)_p \leq C_p n^{-1/p} E_n(f)_{V_p}$. Therefore from the Theorem 1 it follows

Corollary 2. *The assertion of the Corollary 1 is valid, if instead of the condition* (5) *we assume*

$$\sum_{k=1}^{\infty} \gamma_k \left[k^{-1/2-1/p} E_k(f)_{V_p}\right]^{\beta} < \infty.$$

The following two theorems show that under some conditions the statement of Theorem 1 is sharp.

Theorem 3. *Let $1 \leq p < \infty$, $0 < \beta < p$, $\gamma \in A(p/(p-\beta))$, and let be given some decreasing and tending to zero sequence $\varepsilon = \{\varepsilon_i\}_{i=1}^{\infty}$ satisfying Bary condition*

$$\sum_{i=k}^{\infty} \varepsilon_i / i = O(\varepsilon_k), \quad k \in \mathbf{N}, \tag{6}$$

and $\sum_{k=1}^{\infty} \gamma_k \left(k^{-1/2}\varepsilon_k\right)^{\beta} = \infty$. Then there exists a function $f \in L_p[0,1]$ such that $E_n(f)_p \leq \varepsilon_n$, $n \in \mathbf{N}$, and the series (4) *diverges.*

Theorem 4. *Let $0 < \beta \leq 1$ for $1 < p < \infty$ and $0 < \beta < 1$ for $p = 1$. Moreover, let be given the sequence $\gamma \in A(p/(p-\beta))$ such that $(1-\alpha)2^{\beta/2}\Gamma_{n+1} \geq \Gamma_n$ for some $\alpha \in (0,1)$ and a decreasing and tending to zero sequence $\varepsilon = \{\varepsilon_i\}_{i=1}^{\infty}$ such that $\sum_{k=1}^{\infty} \gamma_k \left(k^{-1/2}\varepsilon_k\right)^{\beta} = \infty$. Then there exists a function $f \in L_p[0,1)$ such that $E_n(f)_p \leq \varepsilon_n$, $n \in \mathbf{N}$, and the series* (4) *diverges.*

The following theorem shows that under some conditions the statement of the Corollary 2 is sharp.

Theorem 5. *Let $1 < p < \infty$, $0 < \beta < p$, $\gamma \in A(p/(p-\beta))$, and let be given some decreasing and tending to zero sequence $\varepsilon = \{\varepsilon_i\}_{i=1}^{\infty}$ satisfying Bary condition* (6) *and such that $\sum_{k=1}^{\infty} \gamma_k \left(k^{-1/2-1/p}\varepsilon_k\right)^{\beta} = \infty$. Then there exists a function $f \in V_p[0,1]$ such that $E_n(f)_{V_p} \leq \varepsilon_n$, $n \in \mathbf{N}$, and the series* (4) *diverges.*

Remark. Theorem 2 and Corollaries 1 and 2 have two-dimensional analogs which will appear elsewhere.

ACKNOWLEDGMENT

The work of the first author is supported by the Russian Foundation for Basic Research under Grant No 11-01-00321. The work of the second author is supported by the Russian Foundation for Basic Research under Grant No 13-01-00238.

REFERENCES

[1] A.Haar, Theorie der orthogonalen Functionensysteme, Math. Ann., **69** (1910), 331-371.

[2] B.S. Kashin, A.A. Sahakyan, Orthogonal series (Russian), Moscow, Nauka, 1984, 496p.

[3] Z. Ciesielski, J. Musielak, On absolute convergence of Haar series, Colloquium Math., **7**:1 (1959), 61-65.

[4] N. Wiener, The quadratic variation of a function and its Fourier coefficients, J. Math. and Phys., **3** (1924), 72-94.

[5] P.L. Ul'yanov, On the series with respect to Haar system (Russian), Mat. Sb., **63**: 3 (1964), 356-391.
[6] B.I. Golubov, On the Fourier series of continuous functions with respect to the Haar system (Russian), Izv. Akad. Nauk SSSR, Ser. mat., **28**:6 (1964), 1271-1296.
[7] B.I. Golubov, Absolute convergence of double series of Fourier-Haar coefficients for functions of bounded p-Variation (Russian), Izv. Vyssh. Ucheb. Zaved. Matematika, **6** (2012), pp. 3-13. Engl.transl.: Russian Math. (Iz. VUZ), **56**:6 (2012), 1-10.
[8] L. Gogoladze, R. Meskhia, On the absolute convergence of trigonometric Fourier series, Proc. Razmadze Math. Institute, **141** (2006), 29-40.
[9] A.P. Terekhin, Approximation of functions of bounded p-variation (Russian), Izv. Vyssh. Ucheb. Zaved. Matematika, **2** (1965), 171-187.

Mellin analysis and exponential sampling. Part I: Mellin fractional integrals

Paul L. Butzer
Lehrstuhl A fuer Mathematik
RWTH Aachen University
Aachen, Germany
Email: butzer@rwth-aachen.de

Carlo Bardaro
and Ilaria Mantellini
Department of Mathematics and Computer Science
University of Perugia, Perugia, Italy
Email: bardaro@unipg.it, mantell@dmi.unipg.it

***Abstract*—The Mellin transform and the associated convolution integrals are intimately connected with the exponential sampling theorem. Thus it is very important to develop the various tools of Mellin analysis. In this part we pave the way to sampling analysis by studying basic theoretical properties, including Mellin-type fractional integrals, and give a new approach and version for these integrals, specifying their basic semigroup property. Especially their domain and range need be studied in detail.**

I. Introduction

The theory of Mellin transforms and Mellin approximation theory was introduced in a systematic form, fully independent of Fourier analysis in [6], papers on the present line of research being [1], [2], [3], [4]. Mellin transform theory is intimately connected with the exponential sampling theorem, stating that

$$f(x) = \sum_{k=-\infty}^{+\infty} f(e^{k/T})\mathrm{lin}_{c/T}(e^{-k}x^T) \quad (x \in I\!R^+),$$

where f is a function which is Mellin-bandlimited to the interval $[-\pi T, \pi T]$, and

$$\mathrm{lin}_c(x) := x^{-c}\mathrm{sinc}(\log x), \quad \mathrm{lin}_c(1) = 1,$$

(see [8]). This version of the Shannon sampling theorem has many applications in optical physics and engineering ([13], [16], [5], [14]). Here the samples are not equally spaced apart as in the case of the Whittaker-Kotel'nikov-Shannon sampling theorem, but exponentially spaced; such spacing is needed in those applications where independent pieces of information accumulates near time $t = 0$.

The aim of this research is to put into a rigorous framework such applications, making use only of results from the Mellin transform theory. In [6] the following sentence is written: *The proofs of the Mellin results are mostly said to follow by a change of variable and a change of function from the corresponding Fourier or Laplace results. In fact one expresses it as follows: "It is a matter of using the theory of the Fourier or Laplace transform to derive what one needs concerning the Mellin transform". However, the hypotheses upon which the Mellin theory lies are often considered quite uncritically, and certainly by no means in a unified, systematic fashion.*

While the classical proof of the Shannon sampling theorem is based on the Poisson summation formula, the exponential version is established via the Mellin-Poisson summation formula, which connects the classical Mellin transform with the finite Mellin transform. Variuos fundamental facts in exponential sampling theory must be developed and, in this direction, a deep study of Mellin analysis appears necessary. In particular, the properties of Mellin convolution integrals and the Mellin differential operators are fundamental tools. In papers [9], [10], [11] certain Mellin convolution integrals, namely the so-called Hadamard- type fractional integrals, were developed: these integrals represent the appropriate extensions of the classical Riemann-Liouville and Weyl fractional integrals and also lead to definitions of certain Mellin fractional differential operators, (see also the book [15]). The purpose of this article is a continuation of these topics. As remarked in [9] the natural operator of Mellin fractional integration is not the classical Riemann-Liouville fractional integral of order $\alpha > 0$ on $I\!R^+$, (see [17], [12]) but the integral

$$(J^{\alpha}_{0+}f)(x) = \frac{1}{\Gamma(\alpha)}\int_0^x \left(\log\frac{x}{u}\right)^{\alpha-1} f(u)\frac{du}{u} \quad (x > 0). \quad (1)$$

Thus the operator of integration (anti-differentiation) is not the integral $\int_0^x f(u)du$, as used throughout the literature in matters Mellin transforms, including its table volumes, but the integral $\int_0^x f(u)du/u$. The use of the latter makes calculations not only much simpler but also more elegant.

For the development of the theory, it is important to consider the following generalization of (1), for $\mu \in I\!R,\ x > 0$ namely (see [9], [10], [11])

$$(J^{\alpha}_{0+,\mu}f)(x) = \frac{1}{\Gamma(\alpha)}\int_0^x \left(\frac{u}{x}\right)^{\mu}\left(\log\frac{x}{u}\right)^{\alpha-1} f(u)\frac{du}{u} \quad (2)$$

for functions belonging to the space X_c of all measurable complex-valued functions defined on $I\!R^+$, such that $(\cdot)^{c-1}f(\cdot) \in L^1(I\!R^+)$.

Since the definition of pointwise fractional derivatives, as defined in [10], is based on the Hadamard type integral, it is important to study in depth the domain and the range of these integrals. Here we first introduce the local spaces $X_{c,loc}$ and furnish some results concerning both the domain and the range. In this respect, a fundamental role is played by a new version of the basic semigroup property, which is proved here

in a direct way, as an extension of a corresponding property for spaces X_c^p given in [10].

The theory of Hadamard fractional integrals is one of the topics which reveals the importance of a direct approach via Mellin transforms. For example, while the domain of the classical Riemann-Liouville fractional operators of any order α contains all the locally integrable functions over the positive real line, this is no longer true for Hadamard operators. Indeed, for $\alpha > 1$, the domain of $J^{\alpha}_{0+,c}$ is strictly contained in the space $X_{c,loc}$. This implies that the Hadamard integrals and the corresponding notion of pointwise Mellin fractional derivative, which we develop in the second part of this study, represent new types of integro-differential operators which must be properly treated using Mellin transform theory.

In the second part we apply these results to the exponential sampling.

II. Mellin fractional integrals

Let $L^1 = L^1(\mathbb{R}^+)$ be the space of all the Lebesgue measurable and integrable complex valued functions defined on $\mathbb{R}^+$, endowed with the usual norm.

Let us consider the space, for some $c \in \mathbb{R}$,

$$X_c = \{f : \mathbb{R}^+ \to \mathbb{C} : f(x)x^{c-1} \in L^1(\mathbb{R}^+)\}$$

endowed with the norm

$$\|f\|_{X_c} = \|f(\cdot)(\cdot)^{c-1}\|_{L^1} = \int_0^{\infty} |f(u)|u^{c-1}du.$$

For $a, b \in \mathbb{R}$ we define the spaces $X_{(a,b)}$, $X_{[a,b]}$ by

$$X_{(a,b)} = \bigcap_{c\in]a,b[} X_c, \quad X_{[a,b]} = \bigcap_{c\in[a,b]} X_c$$

and, for every c in the given intervals, $\|f\|_{X_c}$ is a norm on them.

We define for every $f \in X_c$ the Mellin transform, with $s = c + it \in \mathbb{C}$, $c, t \in \mathbb{R}$, by

$$M[f](s) \equiv [f]^{\wedge}_M(s) = \int_0^{\infty} u^{s-1}f(u)du.$$

Thus $M : X_c \to C(\{c\} \times i\mathbb{R})$, $f \to M[f] = [f]^{\wedge}_M$, (see [6]). A boundedness property for $J^{\alpha}_{0+,\mu}$ in the space X_c, is needed when the coefficient μ is greater than c. This is due to the fact that only for $\mu > c$, we can view $J^{\alpha}_{0+,\mu}f$ as a Mellin convolution between two functions in X_c. However, we are interested here in the domain and the range of these fractional operators when $\mu = c$. We will show that for any non-trivial function f in the domain of $J^{\alpha}_{0+,c}$ the image $J^{\alpha}_{0+,c}f$ cannot be in X_c. This implies that we cannot compute its Mellin transform in the space X_c.

We define the domain of $J^{\alpha}_{0+,c}$, for $\alpha > 0$ and $c \in \mathbb{R}$, as the class of all the functions such that

$$\int_0^x u^c \left(\log \frac{x}{u}\right)^{\alpha-1} |f(u)| \frac{du}{u} < +\infty,$$

for a.e. $x \in \mathbb{R}^+$, denoted by $DomJ^{\alpha}_{0+,c}$.

Let $X_{c,loc}$ be the space of all the functions such that $(\cdot)^{c-1}f(\cdot) \in L^1(]0, a[)$ for every $a > 0$.

Proposition 1: If $f \in X_{c,loc}$, then the function $(\cdot)^c f(\cdot) \in X_{1,loc}$. Moreover, if $c < c'$, then $X_{c,loc} \subset X_{c',loc}$.

Note that the above inclusion does not hold for spaces X_c.

Concerning the domain of the operator $J^{\alpha}_{0+,c}$, we begin with

Proposition 2: Let $\alpha > 1$, $c \in \mathbb{R}$ be fixed. Then $DomJ^{\alpha}_{0+,c} \subset X_{c,loc}$.

For $\alpha = 1$ we have immediately $DomJ^{1}_{0+,c} = X_{c,loc}$. The case $0 < \alpha < 1$ is more delicate. In this instance $X_{c,loc} \subset DomJ^{\alpha}_{0+,c}$, due to the following "local" version of the semigroup property of $J^{\alpha}_{0+,c}$:

Theorem 1: Let $\alpha, \beta > 0$, $c \in \mathbb{R}$ be fixed. Let $f \in DomJ^{\alpha+\beta}_{0+,c}$. Then

(i) $f \in DomJ^{\alpha}_{0+,c} \cap DomJ^{\beta}_{0+,c}$
(ii) $J^{\alpha}_{0+,c}f \in DomJ^{\beta}_{0+,c}$ and $J^{\beta}_{0+,c}f \in DomJ^{\alpha}_{0+,c}$.
(iii) $(J^{\alpha+\beta}_{0+,c}f)(x) = (J^{\alpha}_{0+,c}(J^{\beta}_{0+,c}f))(x)$, a.e. $x \in \mathbb{R}^+$.
(iv) If $\alpha < \beta$ then $DomJ^{\beta}_{0+,c} \subset DomJ^{\alpha}_{0+,c}$.

Thus if $0 < \alpha \leq 1$, $c \in \mathbb{R}$, then $X_{c,loc} \subset DomJ^{\alpha}_{0+,c}$.

The inclusion in (iv) of Theorem 1 is strict for any choice of α and β. It is sufficient to consider the function, with $\alpha < \gamma < \beta$,

$$f(x) = \frac{x^{-c}}{|\log x|^{\gamma}} \chi_{]0,1/2[}(x),$$

$\chi_{]0,1/2[}$ being the characteristic function of interval $]0, 1/2[$.

A sufficient condition in order that a function f belongs to $DomJ^{\alpha}_{0+,c}$ for $\alpha > 1$, is

Proposition 3: Let $\alpha > 1$. If $f \in X_{c,loc}$ is such that $f(u) = \mathcal{O}(u^{-(r+c-1)})$ for $u \to 0^+$ and $0 < r < 1$, then $f \in DomJ^{\alpha}_{0+,c}$.

As a consequence, for $c \in \mathbb{R}$ fixed, we have

$$\widetilde{X}_{c,loc} \subset \bigcap_{\alpha>0} DomJ^{\alpha}_{0+,c}.$$

Concerning the range of the operators $J^{\alpha}_{0+,c}$ we need the following important propositions.

Proposition 4: Let $\alpha > 0$, $c \in \mathbb{R}$ be fixed. If $f \in DomJ^{\alpha+1}_{0+,c}$, then $J^{\alpha}_{0+,c}f \in X_{c,loc}$.

As a consequence we can deduce that if $f \in DomJ^{\alpha}_{0+,c}$, not necessarily does $J^{\alpha}_{0+,c}f \in X_{c,loc}$.

For spaces X_c we have the following

Proposition 5: Let $\alpha > 0$, $c \in \mathbb{R}$ be fixed. If $f \in DomJ^{\alpha}_{0+,c}$, then $J^{\alpha}_{0+,c}f \notin X_c$, unless $f = 0$ a.e. in $\mathbb{R}^+$.

However we have the following property.

Proposition 6: Let $\alpha > 0,\ c, \nu \in \mathbb{R},\ \nu < c$, being fixed. If $f \in Dom J^{\alpha}_{0+,c} \cap X_{[\nu,c]}$, then $J^{\alpha}_{0+,c} f \in X_\nu$ and

$$\|J^{\alpha}_{0+,c} f\|_{X_\nu} = \frac{\|f\|_{X_\nu}}{(c-\nu)^\alpha}.$$

Moreover, for any $s = \nu + it$,

$$|M[J^{\alpha}_{0+,c} f](s)| \leq \frac{\|f\|_{X_\nu}}{(c-\nu)^\alpha}.$$

REFERENCES

[1] C. Bardaro and I. Mantellini, *Voronovskaya-type estimates for Mellin convolution operators*, Result Math., 50, (2007), 1-16.

[2] C. Bardaro and I. Mantellini, *Quantitative Voronovskaja formula for Mellin convolution operators*, Mediterr. J. Math., 7(4), (2010), 483-501.

[3] C. Bardaro and I. Mantellini, *Approximation properties for linear combinations of moment type operators*, Comput. Math. Appl., 62, (2011), 213-229.

[4] C. Bardaro and I. Mantellini, *On the iterates of Mellin-Fejer convolution operators*, Acta Appl. Math., 121, (2012), 2304-2313.

[5] M. Bertero and E.R. Pike, *Exponential sampling method for Laplace and other dilationally invariant transforms I. Singular-system analysis. II. Examples in photon correction spectroscopy and Frauenhofer diffraction*, Inverse Problems, 7 (1991), 1-20; 21-41.

[6] P.L. Butzer and S. Jansche, *A direct approach to the Mellin transform*, J. Fourier Anal. Appl., **3**, (1997), 325-375.

[7] P.L. Butzer and S. Jansche, *The finite Mellin transform, Mellin-Fourier series, and the Mellin-Posson summation formula*, Proc. 3rd int. conf. on Functional Analysis and Approximation Theory, Maratea, Sept. 1996, Rend. Circ. Mat. Palermo,

[8] P.L. Butzer and S. Jansche, *The exponential sampling theorem of signal analysis*, Atti Sem. mat. Fis. Univ. Modena, Suppl. Vol. 46, (1998), 99-122.

[9] P.L. Butzer, A.A. Kilbas and J.J. Trujillo, *Fractional calculus in the Mellin setting and Hadamard-type fractional integral*, J. Mat. Anal. Appl., 269, (2002), 1-27

[10] P.L. Butzer, A.A. Kilbas and J.J. Trujillo, *Compositions of Hadamard-type fractional integration operators and the semigroup property*, J. Mat. Anal. Appl., 269, (2002), 387-400.

[11] P.L. Butzer, A.A. Kilbas and J.J. Trujillo, *Mellin transform analysis and integration by parts for hadamard-type fractional integrals*, J. Mat. Anal. Appl., 270, (2002), 1-15.

[12] P.L. Butzer and U. Westphal, *An introduction to fractional calculus*, In: Hifler, H., Ed; Applications of Fractional Calculus in Physics, Singapore, Wordl Scientific Publ. (2000), 1-85.

[13] D. Casasent (Ed), *Optical Data Processing*, Springer, Berlin (1978); 241-282.

[14] F. Gori, *Sampling in optics*, in "Advances Topics in Shannon Sampling and Interpolation Theory (R.J. Marks II Ed), Springer, New York (1993), 37-83.

[15] A.A. Kilbas, H.M. Srivastava and J.J. Trujillo, *Theory and applications of fractional differential equations*, Elsevier, Amsterdam, 2006.

[16] N. Ostrowksy, D. Sornette, P. Parker and E.R. Pike, *Exponential sampling method for light scattering polydispersity analysis*, Opt. Acta, 28, (1994), 1059-1070.

[17] S.G. Samko, A.A. Kilbas and O.I. Marichev, *Fractional Integrals and Derivatives. Theory and Applications*, Yverdon: Gordon and Breach, Amsterdam, (1993).

Mellin analysis and exponential sampling. Part II: Mellin differential operators and sampling

Paul L. Butzer
Lehrstuhl A fuer Mathematik
RWTH Aachen University
Aachen, Germany
Email: butzer@rwth-aachen.de

Carlo Bardaro
and Ilaria Mantellini
Department of Mathematics and Computer Science
University of Perugia, Perugia, Italy
Email: bardaro@unipg.it, mantell@dmi.unipg.it

***Abstract*—Here, we introduce a notion of strong fractional derivative and we study the connection with the pointwise fractional derivative, which is defined by means of Hadamard-type integrals. The main result is a fractional version of the fundamental theorem of integral and differential calculus in Mellin frame. Finally there follow the first of several theorems in the sampling area, the highlight being the reproducing kernel theorem as well as its approximate version for non-bandlimited functions in the Mellin sense, both being new.**

I. Introduction

This article is the continuation of the previous one devoted to the study of Mellin fractional integrals. Here we apply the results concerning the Hadamard type integrals in order to define an appropriate notion of the associated pointwise fractional derivative. Moreover we will introduce a notion of a strong fractional derivative in spaces X_c, as an extension to the Mellin setting of the notion of classical strong derivatives in L^p-spaces (see [6]). The pointwise fractional derivative of order $\alpha > 0$, is defined by the Hadamard integrals formally as follows:

$$(D^{\alpha}_{0+,c}f)(x) = x^{-c}\delta^m x^c (J^{m-\alpha}_{0+,c}f)(x),$$

where $m = [\alpha] + 1$ and $\delta := (x\frac{d}{dx})$. The above definition, introduced in [9, Part I], originates from the theory of the classical Mellin differential operator, studied in [6, Part I]. The main result here is an equivalence theorem which strictly connects the two notions of fractional derivatives and the Hadamard integrals. As far as we are aware this kind of equivalence was never stated explicitily in the setting of Fourier transform theory. This is also related to the fundamental theorem of integral and differential calculus in the fractional Mellin setting. For usual Mellin derivatives, this was described in [6, Part I], where, in particular, the representation of the Mellin derivatives in terms of the Stirling numbers of the second kind is discussed in depth. Finally there follow the first of several theorems in the sampling area.

One of the new and important applications regarding the exponential sampling is an error estimate giving the fast rate of approximation depending on the order of the fractional derivative (see Corollary 2 below).

II. The strong and pointwise Mellin fractional differential operators

We recall that X_c denotes the space of all the measurable functions $f : \mathbb{R}^+ \to \mathbb{C}$ such that $f(\cdot)(\cdot)^{c-1} \in L^1(\mathbb{R}^+)$. The Mellin transform of a function $f \in X_c$ is defined by

$$M[f](s) \equiv [f]^{\wedge}_M(s) = \int_0^{\infty} u^{s-1} f(u)du$$

where $s = c + it, t \in \mathbb{R}$, and the Mellin translation operator τ^c_h, for $h \in \mathbb{R}^+,\ c \in \mathbb{R},\ f : \mathbb{R}^+ \to \mathbb{C}$, by

$$(\tau^c_h f)(x) := h^c f(hx) \quad (x \in \mathbb{R}^+).$$

Setting $\tau_h := \tau^0_h$, then $(\tau^c_h f)(x) = h^c(\tau_h f)(x)$, $\|\tau^c_h f\|_{X_c} = \|f\|_{X_c}$. The Mellin fractional difference of $f \in X_c$ of order $\alpha > 0$, defined by

$$\Delta^{\alpha,c}_h f(x) := (\tau^c_h - I)^{\alpha} f(x) = \sum_{j=0}^{\infty} \binom{\alpha}{j} (-1)^{\alpha-j} \tau^c_{h^j} f(x).$$

for $h > 0$, I being the identity operator over the space of all measurable functions on $\mathbb{R}^+$, and

$$\binom{\alpha}{j} = \frac{\alpha(\alpha-1)\cdots(\alpha-j+1)}{j!},$$

has the following properties

Proposition 1: For $f \in X_c$ the difference $\Delta^{\alpha,c}_h f(x)$ exists a.e. for $h > 0$, with

i) $\|\Delta^{\alpha,c}_h f\|_{X_c} \le \|f\|_{X_c} \sum_{j=0}^{\infty} \left| \binom{\alpha}{j} \right|$
ii) $M[\Delta^{\alpha,c}_h f](c+it) = (h^{-it} - 1)^{\alpha} M[f](c+it)$.

Proof. As to (ii) it follows by taking the Mellin transforms on the left, thus

$$\sum_{j=0}^{\infty} \binom{\alpha}{j} (-1)^{\alpha-j} h^{-itj} M[f](c+it).$$

For spaces $X_{[a,b]}$, we have the following Proposition.

Proposition 2: Let $f \in X_{[a,b]}$, and let $c \in]a, b[$.

(i) If $0 < h \le 1$, then $\Delta^{\alpha,c}_h f \in X_{[a,c]}$, and for every $\nu \in [a, c]$

$$\|\Delta^{\alpha,c}_h f\|_{X\nu} \le \|f\|_{X_\nu} \sum_{j=0}^{\infty} \left| \binom{\alpha}{j} \right| h^{(c-\nu)j}.$$

Moreover for $t \in I\!R$,

$$M[\Delta_h^{\alpha,c} f](\nu + it) = (h^{c-\nu-it} - 1)^\alpha M[f](\nu + it).$$

(ii) If $h \geq 1$, then $\Delta_h^{\alpha,c} f \in X_{[c,b]}$, and for every $\mu \in [c, b]$

$$\|\Delta_h^{\alpha,c} f\|_{X\mu} \leq \|f\|_{X_\mu} \sum_{j=0}^{\infty} \left| \begin{pmatrix} \alpha \\ j \end{pmatrix} \right| h^{(c-\mu)j}.$$

Moreover for $t \in I\!R$,

$$M[\Delta_h^{\alpha,c} f](\mu + it) = (h^{c-\mu-it} - 1)^\alpha M[f](\mu + it).$$

Definition. If for $f \in X_c$ there exists $g \in X_c$ such that

$$\lim_{h\to 1} \left\| \frac{\Delta_h^{\alpha,c} f(x)}{(h-1)^\alpha} - g(x) \right\|_{X_c} = 0$$

then g is called the strong fractional derivative of f of order α and it is denoted by $g(x) = \text{s-}\Theta_c^\alpha f(x)$, and

$$W_{X_c}^\alpha := \{f \in X_c : \text{s-}\Theta_c^\alpha f \text{ exists and s-}\Theta_c^\alpha f \in X_c\},$$

with $W_{X_c}^0 = X_c$, is the Mellin Sobolev space. Analogously we define the spaces $W_{X_{[a,b]}}^\alpha, W_{X_{]a,b[}}^\alpha$.

Now to our several basic theorems of the two-parts paper.

Theorem 1: (i) If $f \in W_{X_c}^\alpha$, then for $s = c + it, t \in I\!R$,

$$M[\text{s-}\Theta_c^\alpha f](s) = (-it)^\alpha M[f](s).$$

(ii) If $f \in W_{X_{[a,b]}}^\alpha$, then for every $\nu, c \in [a, b]$,

$$M[\text{s-}\Theta_c^\alpha f](\nu+it) = (c-\nu-it)^\alpha M[f](\nu+it), \quad t \in I\!R.$$

Proof. As to (i), it can be shown in view of

$$\lim_{h\to 1} \left(\frac{h^{-it} - 1}{h - 1} \right)^\alpha = (-it)^\alpha,$$

that

$$\lim_{h\to 1} \left| (-it)^\alpha [f]^\wedge_M(s) - [\text{s-}\Theta_c^\alpha f]^\wedge_M(s) \right| = 0.$$

The pointwise fractional derivative of order α, associated with the integral $J_{0+,c}^\alpha f$, $c \in I\!R$, and $f \in Dom J_{0+,c}^{m-\alpha}$, is given by (see e.g. [9, Part I], [5], [15, Part I])

$$(D_{0+,c}^\alpha f)(x) = x^{-c} \delta^m x^c (J_{0+,c}^{m-\alpha} f)(x)$$

where $\alpha > 0, m = [\alpha] + 1$ and $\delta = (x\frac{d}{dx})$. The (classical) pointwise Mellin derivative of integral order, is defined by

$$\begin{aligned} & \lim_{h\to 1} \frac{\tau_h^c f(x) - f(x)}{h - 1} \\ = & \lim_{h\to 1} \left[h^c x \frac{f(hx) - f(x)}{hx - x} + \frac{h^c - 1}{h - 1} f(x) \right] \\ = & \; x f'(x) + c f(x), \end{aligned}$$

provided f' exists a.e. on $I\!R^+$, and the Mellin differential operator of order $r \in I\!N$ iteratively by $\Theta_c^1 := \Theta_c$, $\Theta_c^r := \Theta_c(\Theta_c^{r-1})$.

The following proposition gives the connection between Mellin and ordinary derivatives.

Proposition 3: For the pointwise derivative of order $r \in I\!N$, we have

$$(D_{0+,c}^r f)(x) = (\Theta_c^r f)(x) = \sum_{k=0}^{r} S_c(r, k) x^k f^{(k)}(x),$$

where $S_c(r, k)$, $0 \leq k \leq r$, denote the generalized Stirling numbers of second kind, defined recursively by

$$S_c(r, 0) := c^r, \; S_c(r, r) := 1,$$

$$S_c(r + 1, k) = S_c(r, k - 1) + (c + k) S_c(r, k).$$

In the fractional case, for a given $\alpha > 0$, we define the space $X_{c,loc}^\alpha$ by

$$\{f \in X_{c,loc} : \exists (D_{0+,c}^\alpha f)(x) \text{ a.e, } D_{0+,c}^\alpha f \in X_{c,loc}\}.$$

Proposition 4: Let $f \in X_{c,loc}^\alpha$ be such that $\Theta_c^m f \in X_{c,loc}$, where $m = [\alpha] + 1$. Then

$$(D_{0+,c}^\alpha f)(x) = \Theta_c^m (J_{0+,c}^{m-\alpha} f)(x) = J_{0+,c}^{m-\alpha} (\Theta_c^m f)(x).$$

Now to the fundamental theorem of the fractional differential and integral calculus in the Mellin frame.

Theorem 2: Let $\alpha > 0$ be fixed. Let $f \in X_{c,loc}^\alpha$, be such that $D_{0+,c}^\alpha f \in Dom J_{0+,c}^m$ and $\Theta_c^m f \in Dom J_{0+,c}^m$. If $\Theta^{m-1} f \in \widetilde{X}_{c,loc}$, then

$$(J_{0+,c}^\alpha (D_{0+,c}^\alpha f))(x) = f(x), \; a.e. \; x \in I\!R^+.$$

Moreover, let $f \in Dom J_{0+,c}^m$ be such that $J_{0+,c}^\alpha f \in X_{c,loc}$. Then

$$(D_{0+,c}^\alpha (J_{0+,c}^\alpha f))(x) = f(x), \quad a.e. \; x \in I\!R^+.$$

Concerning the connections between the strong and the pointwise Mellin derivatives, we have the following

Theorem 3: Let $\alpha > 0$ and $c \in]a, b]$ be fixed, and $f \in X_{[a,b]}^\alpha$ be such that $\Theta_c^m f \in X_{[a,b]}$. Then $f \in W_{[a,b]}^\alpha$ and

$$(D_{0+,c}^\alpha f)(x) = \text{s-}\Theta_c^\alpha f(x), \quad a.e. \; x \in I\!R^+.$$

Proof. By Proposition 4 we have

$$(D_{0+,c}^\alpha f)(x) = (J_{0+,c}^{m-\alpha} (\Theta_c^m f))(x).$$

Thus passing to Mellin transforms, we have, for $t \in I\!R$,

$$\begin{aligned} & [D_{0+,c}^\alpha f]^\wedge_M(\nu + it) = [(J_{0+,c}^{m-\alpha} (\Theta_c^m f))]^\wedge_M(\nu + it) \\ & = (c - \nu - it)^{\alpha-m} [\Theta_c^m f]^\wedge_M(\nu + it) \\ = & \; (c - \nu - it)^\alpha [f]^\wedge_M(\nu + it). \end{aligned}$$

Hence, $D_{0+,c}^\alpha f$ and $\text{s-}\Theta_c^\alpha f$ have the same Mellin transform along the line $\nu + it$, and so the assertion follows by the identity theorem (see [6, Part I]).

Using the previous results, we give the following equivalence theorem which is the fractional version of Theorem 10 in [6, Part I].

Theorem 4: Let $f \in X_{[a,b]}$, $\alpha > 0$. The following four assertions are equivalent

(i) $f \in W^{\alpha}_{X_{[a,b]}}$.

(ii) There is a function $g_1 \in X_{[a,b]}$ such that, for every $c \in]a,b[$,

$$\lim_{h\to 1}\left\|\frac{\Delta_h^{\alpha,c} f}{(h-1)^{\alpha}} - g_1\right\|_{X_c} = 0.$$

(iii) There is $g_2 \in X_{[a,b]}$ such that, for every $\nu, c \in]a,b[$,

$$(c-\nu-it)^{\alpha} M[f](\nu+it) = M[g_2](\nu+it).$$

(iv) There is $g_3 \in X_{[a,b]}$ such that for $c \in]a,b[$ and $x \in I\!R^+$,

$$f(x) = \frac{1}{\Gamma(\alpha)}\int_0^x \left(\frac{u}{x}\right)^c \left(\log\frac{x}{u}\right)^{\alpha-1} g_3(u)\frac{du}{u} \quad a.e.$$

If one of the above assertions is satisfied, then $D^{\alpha}_{0+,c} f(x) = \text{s-}\Theta^{\alpha}_c f(x) = g_1 = g_2 = g_3$ a.e. $x \in I\!R^+$.

Proof. It is easy to see that (i) implies (ii), and (ii) implies (iii) by Theorem 1. As to (iii) implies (iv), observe

$$M[J^{\alpha}_{0+,c} g_2](\nu+it) = (c-\nu-it)^{-\alpha} M[g_3](\nu+it).$$

As far as we know, a fundamental theorem with four equivalent assertions in the form presented above for the Mellin transform in the fractional case has never been stated for the Fourier transform. As a fundamental theorem in the present sense it was first established for $2\pi-$periodic functions via the finite Fourier transform in [10], and for the Chebyshev transform in [7], [8]. Fractional Chebyshev derivatives were there defined in terms of fractional order differences of the Chebyshev translation operator, the Chebyshev integral by an associate convolution product. The next fundamental theorem, after that for Legendre transforms (see e,g. [2]), was the one concerned with the Jacobi transform, see e.g. [9].

III. The exponential sampling theorem

Let B^T_c denote the class of functions $f \in X_c,\ f \in C(I\!R^+)$, $c \in I\!R$, which are Mellin band-limited in the interval $[-T,T]$, $T \in I\!R^+$, thus for which $[f]^{\wedge}_M(c+it) = 0$ for all $|t| > T$. A mathematician's version of the exponential sampling theorem introduced by the electrical engineers/physicists M.Bertero, E.R. Pike [5, Part I] and F. Gori [14, Part I], reads as follows

Theorem 5: If $f \in B^{\pi T}_c$ for some $c \in I\!R$, and $T > 0$, then the series

$$x^c \sum_{k=-\infty}^{\infty} f(e^{k/T})\text{lin}_{c/T}(e^{-k}x^T)$$

is uniformly convergent in $I\!R^+$, and one has the representation

$$f(x) = \sum_{k=-\infty}^{\infty} f(e^{k/T})\text{lin}_{c/T}(e^{-k}x^T) \equiv E^c_T f(x) \quad (x \in I\!R^+).$$

The lin_c-function for $c \in I\!R$, $\text{lin}_c : I\!R^+ \to I\!R$, is defined, for $x \in I\!R^+ \setminus \{1\}$, by

$$\text{lin}_c(x) = \frac{x^{-c}}{2\pi i}\frac{x^{\pi i} - x^{-\pi i}}{\log x} = \frac{x^{-c}}{2\pi}\int_{-\pi}^{\pi} x^{-it}dt,$$

with the continuous extension $\text{lin}_c(1) := 1$, thus $\text{lin}_c(x) = x^{-c}\text{sinc}(\log x)$.

As we all know, bandlimitation in the classical Fourier version of the Whittaker-Kotel'nikov-Shannon sampling theorem is a restriction we try to avoid. Likewise it is so in the Mellin setting. In this respect we have the following approximate version.

Theorem 6: Let $f \in X_c \cap C(I\!R^+)$, $c \in I\!R$, be such that $M[f] \in L^1(\{c\} \times i I\!R)$. Then there holds the error estimate

$$\left|f(x) - \sum_{k=-\infty}^{\infty} f(e^{k/T})\text{lin}_{c/T}(e^{-k}x^T)\right| \le \frac{x^{-c}}{\pi}\int_{|t|>\pi T} |M[f](c+it)|dt \quad (x \in I\!R^+,\ T > 0).$$

Corollary 1: Let $f \in X_c \cap C(I\!R^+)$, $c \in I\!R$, be such that $M[f] \in L^1(\{c\} \times i I\!R)$. Then

$$\lim_{T\to+\infty} |f(x) - E^c_T f(x)| = 0, \quad x \in I\!R^+.$$

Further, if $f \in B^{\pi \overline{T}}_c$ for some $\overline{T} > 0$, then, for all $T \ge \overline{T}$,

$$f(x) = E^c_T f(x), \quad x \in I\!R^+.$$

The operator $\text{s-}\Theta^{\alpha}_c f$, $\alpha > 0$, plays the basic role in the following corollary, giving the fast rate of approximation of $f(x)$, depending on its order α, by the exponential sampling sum $E^c_T f(x)$.

Corollary 2: If $f \in W^{\alpha}_{X_c}$, $c \in I\!R, \alpha > 0$, is continuous on $I\!R^+$ such that $M[\text{s-}\Theta^{\alpha}_c f] \in L^1(\{c\} \times i I\!R)$, then

$$|f(x) - E^c_T f(x)| = o(T^{-\alpha}), \quad (x \in I\!R^+; T \to +\infty).$$

Proof. According to Theorem 1, $|[f]^{\wedge}_M(c+it)| = |t|^{-\alpha}|[\text{s-}\Theta^{\alpha}_c f]^{\wedge}_M(c+it)|$, $t \in I\!R$, so that:

$$\int_{|t|>\pi T} |[f]^{\wedge}_M(c+it)|dt \le \frac{1}{\pi^{\alpha}T^{\alpha}}\int_{|t|>\pi T} |[\text{s-}\Theta^{\alpha}_c f]^{\wedge}_M(c+it)|dt = o(T^{-\alpha}),$$

so the assertion follows by Theorem 6.

In the previous new corollary, we can consider the pointwise derivative $D^{\alpha}_{0+,c}$ with the assumptions of Theorem 3.

One of the several theorems which are equivalent to the classical Whittaker-Kotel'nikov-Shannon sampling theorem is the well known reproducing kernel formula. In the Mellin setting, it reads as follows, for functions in $B^{T\pi}_c$, $T > 0$

Theorem 7: Let $f \in B^{T\pi}_c$, $c \in I\!R$, $T > 0$, be fixed. Then we have

$$f(x) = T\int_0^{\infty} f(y)\text{lin}_{c/T}\left((\frac{x}{y})^T\right)\frac{dy}{y} \quad (x \in I\!R^+).$$

Proof: Putting $h(y) = f(y^{1/T})$, we have $h \in B^{\pi}_{c/T}$. Then using the reasoning of Lemma 6.3 in [8, Part I] we can write

$$[h(y)\text{lin}_{c/T}(x/y)]^{\wedge}_M(it) = \frac{x^{-c/T}}{2\pi}\int_{-\pi}^{\pi}[h]^{\wedge}_M(c/T + i(t+v))x^{-iv}dv \quad (t \in I\!R).$$

Then for $t = 0$ we get

$$\int_0^\infty h(y)\mathrm{lin}_{c/T}(x/y)\frac{dy}{y}$$
$$= \frac{x^{-c/T}}{2\pi}\int_{-\pi}^{\pi}[h]^\wedge_M(c/T+iv)x^{-iv}dv = h(x)$$

by Theorem 2.4 in [8, Part I]. Thus we have

$$f(x^{1/T}) = T\int_0^\infty f(y)\mathrm{lin}_{c/T}(x/y^T)\frac{dy}{y},$$

and putting $x^{1/T} = z$ we have the assertion.

A further new result is the approximate reproducing kernel theorem, namely its version for not necessarily Mellin-bandilimited functions. It states that

Theorem 8: Let $f \in X_c$, $c \in I\!R$, be continuous on $I\!R^+$ such that $M[f] \in L^1(\{c\} \times iI\!R)$. Then there holds, for $y \in I\!R^+$ and $T > 0$, the error estimate

$$\left|f(x) - T\int_0^\infty f(y)\mathrm{lin}_{c/T}\left((\frac{x}{y})^T\right)\frac{dy}{y}\right|$$
$$\leq \frac{x^{-c}}{2\pi}\int_{|v|\geq\pi T}|[f]^\wedge_M(c+iv)|dv.$$

IV. The finite Mellin transform and the Mellin-Poisson summation formula

The Poisson summation formula in the classical frame of Fourier analysis is one the cornestones of all mathematical analysis. To formulate and establish it in the Mellin setting one needs further concepts, namely Mellin Fourier series (introduced in [7, Part I]) and the associated finite Mellin transform, since this Poisson summation connects the Mellin transform with its finite version.

Definitions.

(i) A function $f : I\!R^+ \to \boldsymbol{C}$ will be called *recurrent*, if $f(x) = f(e^{2\pi}x)$ for all $x \in I\!R^+$. The function f is called $c-$*recurrent* for $c \in I\!R$, if $x^c f(x)$ is recurrent, i.e., if $f(x) = e^{2\pi c}f(e^{2\pi}x)$ for all $x \in I\!R^+$.

(ii) The space Y_c of $c-$recurrent functions $f : I\!R^+ \to \boldsymbol{C}$ is defined for $c \in I\!R$, by

$$Y_c := \{f \in L^1_{loc}(I\!R^+) : f \text{ c-recurrent}, \ \|f\|_{Y_c} < +\infty\},$$

$\|f\|_{Y_c} = \int_{e^{-\pi}}^{e^\pi}|f(u)|u^{s-1}du$, with $s = c + it$.

(iii) The finite Mellin transform of $f \in Y_c$, $c \in I\!R$ is

$$\mathcal{M}_c[f](k) \equiv [f]^\wedge_{\mathcal{M}_c}(k) = \int_{e^{-\pi}}^{e^\pi} f(u)u^{c+it-1}du, \ (k \in \boldsymbol{Z}),$$

$\mathcal{M}_c : Y_c \to L^\infty(\boldsymbol{Z})$, $f \mapsto \{[f]^\wedge_{\mathcal{M}_c}(k)\}_{k\in\boldsymbol{Z}}$, with $\|\mathcal{M}_c\|_{[Y_c,L^\infty]} = 1$.

(iv) The associated Mellin-Fourier series of $f \in Y_c$ is

$$f(x) \sim \frac{1}{2\pi}\sum_{k=-\infty}^{\infty}[f]^\wedge_{\mathcal{M}_c}(k)x^{-c-ik}, \quad (x \in I\!R^+).$$

Theorem 9: Let $f \in X_c$, $c \in I\!R$, be continuous on $I\!R^+$ such that $\sum_{k=-\infty}^{\infty}|M[f](c+ik)| < \infty$. If the series

$$f^c(x) := \sum_{k=-\infty}^{\infty} f(e^{2\pi k}x)e^{2\pi kc}, \quad (x \in I\!R^+)$$

which is $c-$recurrent and absolutely convergent a.e. on the interval $[e^{-\pi}, e^\pi]$ is also uniformly convergent there, then

$$M[f](c+ik) = \mathcal{M}_c[f^c](k), \quad (k \in \boldsymbol{Z}),$$

and especially there holds

$$f^c(x) = \frac{1}{2\pi}\sum_{k=-\infty}^{\infty}M[f](c+ik)x^{-c-ik}, \quad (x \in I\!R^+).$$

It is the strong feeling of the authors that not only the Mellin-sampling theorem is equivalent to its approximate version, but also the reproducing kernel theorem and its approximate version are equivalent. Even more so all these theorems are equivalent among themselves, and under suitable conditions, are equivalent to the Mellin-Poisson summation formula. Equivalence is understood in the sense that each is a corollary of the others. This is indeed the situation in the non-fractional version of the Fourier case as recently proved in [3], [4].

References

[1] L. V. Ahlfors, *Complex Analysis*, McGraw-Hill Int. Eds, Third Edition, 1979.

[2] P.L. Butzer, *Legendre transform method in the solution of basic problems in algebraic approximation*, In: *Functions, Series, Operators*, (Proc. Conf. Budapest, 1980, dedicated to to L. Fejer and F. Riesz on their hundredth birthday), Coll. Math. Coc. Janos Bolyai, 35, North-Holland, 1883, Vol. I, 277-301.

[3] P.L. Butzer, P.J.S.G. Ferreira, R.J.Higgins, G.Schmeisser, R.L. Stens, *The sampling theorem, Poisson's summation formula, general Parseval formula, reproducing kernel formula and the Paley-Wiener theorem for bandlimited signals-their interconnections*, Appl. Anal., 90, (3-4), (2011), 431-461.

[4] P.L. Butzer, M. Dodson, P.J.S.G. Ferreira, R.J.Higgins, G.Schmeisser and R.L. Stens *The generalized Parseval decomposition formula, the approximate sampling theorem, the approximate reproducing kernel formula, Poisson's summation formula, and Riemann's Zeta functionthe; their interconnections for nonbandilimited functions*, to appear.

[5] P.L. Butzer, A.A. Kilbas and J.J. Trujillo, *Stirling functions of the second kind in the setting of difference and fractional calculus*, Numer. Funct. Anal. Optimiz., 4(7-8), (2003), 673-711.

[6] P.L. Butzer and R.J. Nessel, *Fourier Analysis and Approximation*. Vol.I, Academic Press, New York (1971).

[7] P.L. Butzer and R.L. Stens,*The operational properties of the Chebyshev transform. II. Fractional derivatives*, in "The theory of the approximation of functions, (Proc. Intern. Conf., Kaluga, 1975)" (Russian), 49-61, "Nauka", Moscow, 1977.

[8] P.L. Butzer and R.L. Stens *Chebyshev transform methods in the theory of best algebraic approximation*. Abh.Math. Sem.Hamburg 45, (1976), 165-190.

[9] P.L. Butzer, R.L. Stens and M. Wehrens, *Higher moduli of continuity based on the Jacobi translation operator and best approximation*. C.R. Math. Rep. Acad.Sci.Canada 2(1980), 83-87.

[10] P.L. Butzer and U. Westphal, *An access to fractional differentiation via fractional differece quotients*, "Fractional calculus and its Applications", Proc. conf. New Haven, Lecture Notes in Math, 457, (1975), 116-145, Springer, Heidelberg.

[11] S.G. Samko, A.A. Kilbas and O.I. Marichev, *Fractional Integrals and Derivatives. Theory and Applications*, Yverdon: Gordon and Breach, Amsterdam, (1993).

Optimisation and control of sampling rate in localisation microscopy

Seamus J. Holden, Thomas Pengo and Suliana Manley
Laboratory of Experimental Biophysics
Ecole Polytechnique Fédérale de Lausanne (EPFL)
CH-1015 Lausanne, Switzerland
Email: suliana.manley@epfl.ch

***Abstract*—Localisation microscopy (PALM/ STORM) involves sampling sparse subsets of fluorescently labelled molecules, so that the density of bright fluorophores in a single frame is low enough to allow single molecule sub-diffraction limited localisation. The sampling rate, i.e. the density of bright fluorophores per unit time, is key to both the temporal and spatial resolution of localization microscopy. Here we present DAOSTORM, an image analysis algorithm allowing increased sampling rate, and AutoLase, an algorithm for measurement and closed-loop feedback control of sampling rate.**

I. Introduction

Localisation microscopy (PALM [1]/ STORM [2], etc.) involves two key insights. Firstly, that the positions of well separated point sources can be localized to sub-diffraction limited accuracy. Secondly, that fluorescent molecules can be made to blink in a controlled fashion under appropriate experimental conditions. By adjusting the blinking of a fluorophore such that it spends most of its time in a dark inactive state, and only a tiny fraction of its time in a bright, photon-emitting state, a single image of even a densely-labelled structure will show only a few active, well separated point sources within the image. Repeated imaging of the sample records the position of different subsets of fluorophores; by combining the many subsets of localizations obtained from multiple images, a single super-resolved image of all fluorophores within the sample may be constructed.

One of the most important parameters in localization microscopy is the sampling rate, i.e. the density of bright fluorophores per unit time. If the sampling rate is too high, the bright fluorophores will no longer be well separated, and the spatial resolution of the image will be degraded. If the sampling rate is too low, an unnecessarily large number of raw images will be required to reconstruct a single super-resolved image, reducing the temporal resolution of the measurement. Sampling rate is thus key to both the temporal and spatial resolution of localization microscopy.

Here, we focus on two key sampling problems in localization microscopy: how to increase the maximal sampling rate, and how to maintain optimal sampling rate during data acquisition.

II. Increased sampling rate by high density localization

Until recently, algorithms for localization microscopy took the following simplistic approach. All bright fluorophores within a sample are assumed to be well separated (separation much greater than FWHM of the point spread function, PSF). Then bright spots in the image are identified and fitted with a single model PSF (usually a 2D Gaussian). However, if two spots overlap even slightly, this approach fails due to the inadequacy of the fitting model, producing a single localization which is in-between the two overlapping spots. This approach only works when the imaging density (the density of bright fluorophores in a single image) is very low, severely limiting the sampling rate of the technique.

We developed DAOSTORM [3], which is capable of single molecule localization at much higher imaging density. This is achieved by simultaneously fitting multiple model PSFs to bright regions of the image, instead of just one model PSF. This simple improvement over previous algorithms allows localization at much higher imaging density, increasing the sampling rate and temporal resolution of the technique.

We compared DAOSTORM to two common "sparse" localization algorithms. "Sparse Algorithm 1" (SA1) [2] fits candidate molecules with a single Gaussian PSF of variable size and ellipticity. Localizations arising from overlapping molecules are rejected if the fitted PSF appears too elliptical ("shape-based filtering"), or too large/ small ("size-based filtering"). "Sparse Algorithm 2" (SA2) [4] fits candidate molecules with a single Gaussian PSF of fixed shape and size, without shape/ size-based filtering.

We first investigated the qualitative performance of each algorithm for images of Alexa647-labelled microtubules in fixed COS-7 cells in dSTORM photoswitching conditions [4]. The results of each algorithm on single raw images, illustrates the characteristic performance of each algorithm (Fig. 1a-c). SA1 only localized isolated molecules, which were fitted with small localization error. SA2 localized a larger fraction of the molecules, but showed large localization errors for overlapping molecules. DAOSTORM outperformed both sparse algorithms, successfully identifying almost all molecules with small localization error.

We quantified the performance of each algorithm by analyz-

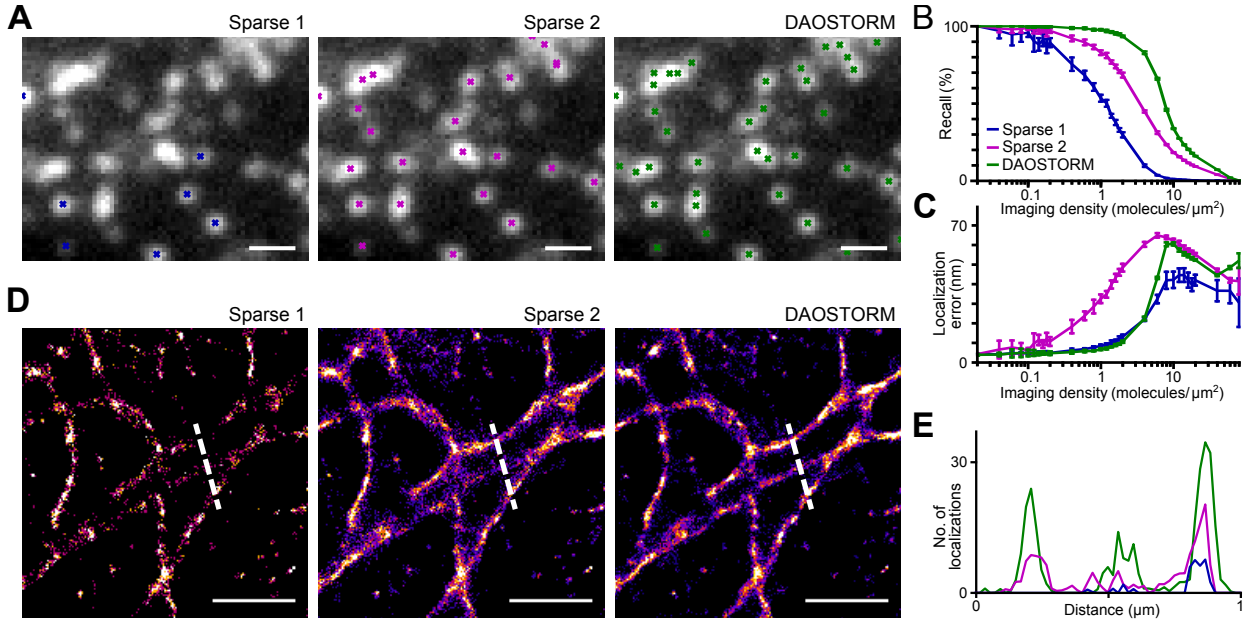


Fig. 1. *Comparison of DAOSTORM to existing super resolution localization algorithms. A.* A single image of fluorescently labeled microtubules was analyzed using SA1, SA2 and DAOSTORM. Crosses represent localizations for each algorithm. *B, C.* Recall (B) and localization error (C) of the algorithms used in a measured for simulated images of randomly distributed surface-immobilized molecules. Error bars, s.d. (n = 10). *D.* Super-resolved microtubule images from a 2000-frame data series. *E.* Line plots of cross-section indicated by dashed lines in D. Scale bars, 1 μm. Reproduced from [3]

ing simulations of randomly distributed surface-immobilized fluorophores. We compared observed localizations to simulated positions, calculating the recall and localization error at different imaging densities. The recall is the percentage of simulated fluorophores successfully detected. The localization error is the root-mean-square distance between a localization and the simulated position.

DAOSTORM substantially outperformed the sparse algorithms in simulations at high signal-to-noise ratio (SNR) typical of STORM data (Fig. 1d). SA2 gave large localization errors even at low imaging density. In contrast, DAOSTORM gave small localization errors similar to the other "precise" algorithm, SA1, together with a 6-fold improvement in recall performance.

Next, we recorded dSTORM images of the microtubule network described above, and used each algorithm to obtain super-resolved images (Fig. 1e-g). SA1 showed low recall, producing poorly sampled STORM images, while SA2 achieved higher recall, but with large localization error, leading to poorly-defined, noisy images. DAOSTORM showed high recall and small localization error, producing well-defined, low-noise images. A line-plot across three parallel microtubules demonstrates the performance difference among the algorithms (Fig. 1h): DAOSTORM resolved all three microtubules, SA2 detected two, and SA1 detected only one.

These results demonstrate the ability of DAOSTORM to increase the maximum sampling rate in localization microscopy, and thus increase temporal resolution. DAOSTORM can also increase quality of super-resolved images of biological samples in situations where control of imaging density is poor.

III. Measurement and Closed-Loop Feedback Control of Sampling Rate

During data acquisition, a careful balance in sampling rate is required: if sampling rate is too high, spatial resolution is reduced; if sampling rate is too low, temporal resolution is reduced. In both PALM and STORM, the sampling rate is usually sensitive to the illumination intensity of a "photoactivation" UV laser [1], [2]. The sampling rate can thus be adjusted to its optimal level by changing the photoactivation laser power (hereafter, UV power). However, the UV power required to maintain optimal sampling rate will vary significantly during a measurement, due to irreversible photobleaching of an increasing fraction of the fluorophores as the experiment progresses. It will also vary significantly between different fields of view within a sample, e.g. due to variations in the morphology and labelling density of the labelled structure.

Sampling rate is usually controlled by continuous manual assessment of the density of molecules in any single frame, and manual adjustment of UV power. This is tedious, and most importantly, is incompatible with automation. To resolve this, we present AutoLase, an algorithm for measurement and closed-loop feedback control of sampling rate.

A conceptually straightforwards approach [5] is to perform real-time localization analysis as the data is acquired and optimise the UV power based on the observed number of localizations. However, this approach has two serious limitations. Firstly, real-time localization is computationally intensive; this approach will therefore be difficult to implement for high frame-rate imaging and/ or for large field of view cameras

(e.g. sCMOS cameras). Secondly, and most importantly, this approach will fail at high imaging density, since multiple overlapping PSFs will be erroneously grouped together.

The design requirements for AutoLase are thus low computational burden and good performance at high imaging density. Instead of trying to optimise sampling rate using only the information from an individual frame, the problem can be significantly simplified by including temporal information from multiple frames. The amount of time that individual fluorophores remain in a bright, photon emitting state is Poisson distributed about a mean lifetime τ_{on}. Therefore, any region of a sample which remains continuously bright for significantly longer than τ_{on} very likely contains multiple bright fluorophores instead of just one. We devised an image-based estimator of τ_{on}, by estimating the amount of time that each pixel in an image has been continuously bright. This allows us to estimate the number of bright molecules, without the need for real-time localization. Since τ_{on} will increase with the number of active bright molecules, this estimator will be robust at high imaging density.

For each frame k, and for each pixel i, we define the estimated on-time, $\tau_{i,k}$,

$$\tau_{i,0} = 0,$$
$$\tau_{i,k} = (\tau_{i,k-1} + \Delta t)\ M_{th}(I_i),$$

where Δt is the interval between each frame, I_i is the intensity at the current pixel, th is an intensity threshold, and $M_{th}(I)$ is the binary threshold operator,

$$M_{th}(I) = \begin{cases} 1 & \text{if } I \geq th, \\ 0 & \text{otherwise.} \end{cases}$$

Each time the pixel intensity I_i falls below th, $\tau_{i,k}$ is set to 0. If I_i is above threshold, $\tau_{i,k}$ is equal to the duration for which that pixel has been above threshold at frame k. τ is thus a measure of how long each pixel has been *continuously* bright.

We implemented closed-loop feedback control of τ. The maximum value of τ at each frame K is smoothed via a running mean

$$\tau_{max} = \frac{1}{N} \sum_{K-N+1}^{K} \max_i\ \tau_{i,k},$$

and compared to a target value T. If the observed value of τ is above or below T by more than x %, then the UV power is reduced or increased, respectively. We calculated the image maximum of τ rather than an average, since we reasoned that the key criterion is that no region of the image contains too many active molecules.

Closed-loop feedback control was implemented on a home-built microscope, controlled using the open-source instrument control software, Micromanager [6]. We wrote a plugin to Micromanager, called *AutoLase*, to perform the feedback control, which we will shortly release as open-source software. Because Micromanager is open-source and works for a large variety of instruments, and because AutoLase is not computationally intensive and does not require real time localization analysis, it should be straightforward for researchers to implement feedback control on their own systems using our software.

The performance of the AutoLase algorithm in estimating τ_{on} is shown in Fig. 2. Live *C. crescentus* bacteria expressing FtsZ-Dendra2 [7] were imaged at a frame rate of 100 Hz using AutoLase to control the imaging density. An exemplar subset of frames (Fig. 2A) shows the blinking behaviour of the labelled molecules. Most molecules remain on for less than 100 ms, however, two molecules (top middle and top right of images) remain on for greater than 200 ms. The on-time estimator τ successfully captures this behaviour (Fig. 2B), showing only two regions active for greater than 200 ms, consistent with the visual interpretation of the raw data.

AutoLase feedback control is shown in Fig. 2C-D. With feedback control (Fig. 2C), the laser power was initially 0 %, and AutoLase was turned on at t=0 s. The raw τ_{max} data is quite noisy (grey line), but clear trends are visible in the smoothed data (black line). Before t=0 s, most molecules are in their dark state, with only occasional spikes in τ_{max} due to autofluorescence or photoactivation by the imaging laser. When AutoLase is turned on at t=0 s, the laser power (blue line) is rapidly increased and stabilises at $\sim$10 % for the first 50 s of imaging, after which point it increases in approximately exponential form to the maximum power. This produces observed on-times stable around the target value of 400 ms for nearly 100 s, after which τ_{max} gradually decreases because very few unbleached molecules remain.

Without feedback control (Fig. 2D), a new field of view (FOV) was chosen, and the laser power was set to 10 % of maximum power at t=0 s, since this was observed to be the stable initial value for the previous FOV. Interestingly, this power level produces an observed τ_{max} well above the target value of 400 ms for the first 50 s of imaging. This is presumably due to variation in density of labelled molecules between different FOVs. Between 50–100 s, τ_{max} is near the target value, after which it decreases rapidly.

These results show that AutoLase can rapidly and accurately optimise τ_{max} to a given target value, and that this value can be maintained for extended periods of time. By contrast, setting the power to a constant value without feedback control is sensitive to variations in density of labelled molecules, which occurs even between adjacent FOVs (e.g. due to variation in morphology of the labelled structure), and significantly reduces the period for which τ_{max} is close to the target value. In practice, we have found that AutoLase gives performance at least as good as manual optimisation of the UV power, while being compatible with automated imaging.

IV. Conclusions

Sampling rate is a key parameter for localization microscopy. Our algorithm, DAOSTORM, is capable of analysing localization microscopy data even at high imaging density, where many fluorescent molecules are simultaneously

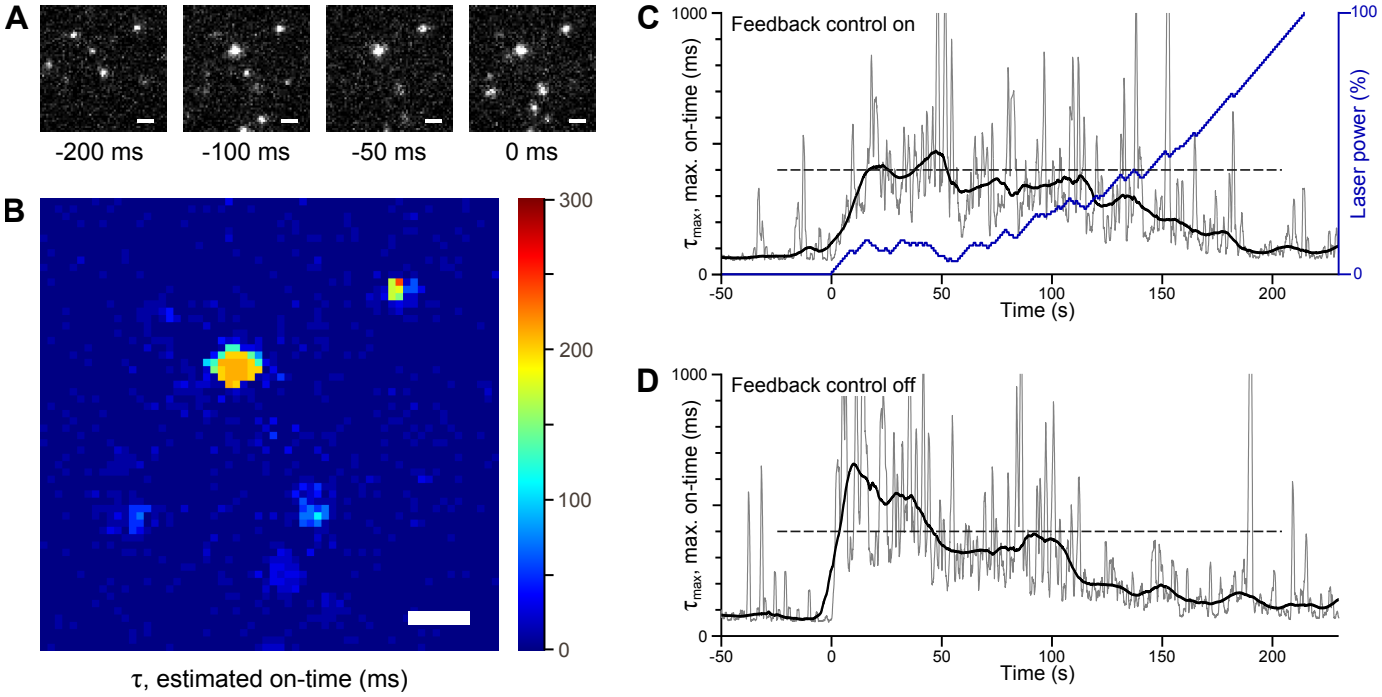


Fig. 2. *Measurement and control of molecule on-time using AutoLase*. *A.* Exemplar subset of images of live *C. crescentus* expressing FtsZ-Dendra2, under photoswithing conditions. *B.* Estimated on-time for each pixel, for frame corresponding to t= 0 ms in A. *C-D.* Observed maximum single-pixel on-time (τ_{max}) , with (C) and without (D) feedback control. Raw data, gray line; smoothed data, black line; laser power, blue line. For the case with feedback control turned off (D), UV power was set to 10 % at t= 0 s.

active. In high signal-to-noise conditions, a sixfold increase in maximum imaging density is obtained. This allows increased sampling rate with minimal loss of spatial resolution. In practice this allows super-resolved images to be constructed from fewer frames of raw data, significantly increasing the temporal resolution of the technique. These improvements are particularly useful for challenging applications such as live-cell super-resolution imaging [8].

We also presented AutoLase, an algorithm for measurement and closed-loop feedback control of sampling rate. Our algorithm is computationally non-intensive and is designed to give good performance even at high imaging density. By allowing automatic optimisation of photoactivation laser intensity, AutoLase facilitates automated localization microscopy measurements.

Acknowledgments

S.J.H and S.M. were supported by the European Research Council (grant 243016-PALMassembly). T.P. was supported by the Brazilian-Swiss Joint Research Program (grant 011004). S.J.H. was supported by a Marie Curie Intra-European Fellowship (grant 297918). We thank Erin Goley (John Hopkins University, USA) for the FtsZ-Dendra2 plasmid, and Justine Collier (University of Lausanne, Switzerland) for technical assistance.

References

[1] E. Betzig, "Imaging intracellular fluorescent proteins at nanometer resolution," *Science*, vol. 313, pp. 1642–1645, 2006.

[2] M. Rust, M. Bates, and X. Zhuang, "Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM)," *Nat. Methods*, vol. 3, pp. 793–795, 2006.

[3] S. J. Holden, S. Uphoff, and A. N. Kapanidis, "DAOSTORM: an algorithm for high- density super-resolution microscopy," *Nat. Methods*, vol. 8, no. 4, pp. 279–280, Apr. 2011.

[4] S. Wolter, M. Schttpelz, M. Tscherepanow, S. Van De Linde, M. Heilemann, and M. Sauer, "Real-time computation of subdiffraction-resolution fluorescence images," *J. Microsc.* vol. 237, no. 1, pp. 12–22, 2010.

[5] A. Kechkar, D. Nair, M. Heilemann, D. Choquet, and J.-B. Sibarita, "Real-time analysis and visualization for single-molecule based super-resolution microscopy," *PLoS ONE*, vol. 8, no. 4, p. e62918, 2013.

[6] A. Edelstein, N. Amodaj, K. Hoover, R. Vale, and N. Stuurman, "Computer control of microscopes using μManager," in *Curr. Protoc. Mol. Biol* , 14.20.1–14.20.17, 2010.

[7] J. S. Biteen, E. D. Goley, L. Shapiro, and W. E. Moerner, "Three-dimensional super-resolution imaging of the midplane protein FtsZ in live caulobacter crescentus cells using astigmatism," *ChemPhysChem*, vol. 13, no. 4, pp. 1007–1012, Mar. 2012.

[8] S.-H. Shim, C. Xia, G. Zhong, H. P. Babcock, J. C. Vaughan, B. Huang, X. Wang, C. Xu, G.-Q. Bi, and X. Zhuang, "Super-resolution fluorescence imaging of organelles in live cells with photoswitchable membrane probes," *Proc. Natl. Acad. Sci. USA.*, vol. 109, no. 35, pp. 13 978–13 983, 2012.

Fast Maximum Likelihood High-density Low-SNR Super-resolution Localization Microscopy

Kyungsang Kim[1], Junhong Min[1], Lina Carlini[2], Michael Unser[3], Suliana Manley[2], Daejong Jeon[1] and Jong Chul Ye[1]
Bio-Imaging & Signal Processing Lab[1], Laboratory of Experimental Biophysics[2] and Biomedical Imaging Group[3]
KAIST, Republic of Korea[1] and EPFL, Switzerland[2,3]
Email:[kssigari, minimok]@kaist.ac.kr, [lina.carlini, michael.unser, suliana.manley]@epfl.ch and jong.ye@kaist.ac.kr

***Abstract*—Localization microscopy such as STORM/PALM achieves the super-resolution by sparsely activating photo-switchable probes. However, to make the activation sparse enough to obtain reconstruction images using conventional algorithms, only small set of probes need to be activated simultaneously, which limits the temporal resolution. Hence, to improve temporal resolution up to a level of live cell imaging, high-density imaging algorithms that can resolve several overlapping PSFs are required. In this paper, we propose a maximum likelihood algorithm under Poisson noise model for the high-density low-SNR STORM/PALM imaging. Using a sparsity promoting prior with concave-convex procedure (CCCP) optimization algorithm, we achieved high performance reconstructions with fast reconstruction speed of 5 second per frame under high density low SNR imaging conditions. Experimental results using simulated and real live-cell imaging data demonstrate that proposed algorithm is more robust than previous methods in terms of both localization accuracy and molecular recall rate.**

I. Introduction

For the past decades, several innovative methods for surpassing the diffraction limit in far-field optical microscopy have been proposed. It is now well known that their significant resolution improvement was originated from exploiting the optical non-linearity. For example, STED and SSIM can achieve the super-resolution by exploiting the non-linearity of high power illumination, whereas the stochastic optical reconstruction microscopy (STORM) [1] and photo-activated localization microscopy (FALM) [2] exploit non-linearity of photoswitchable fluorescence dyes. Specifically, STORM/FALM rely on sparse fluorophore activations such that fluorophores are sparsely activated in both spatial and temporal domain. When the point spread functions (PSFs) of the activated fluorophore are usually not overlapped, these fluorophores can be localized individually based on the least-squares [1], [2] or the maximum-likelihood [3] PSF fitting. To achieve sparse activation, an accumulation rate of localized fluorophores is, however, limited; so that typically several thousands frames are required to reconstruct a single super-resolution image. In other words, its temporal resolution is on the order of minutes, which allows only limited live-cell imaging.

In order to improve the temporal resolution, one of the possible approaches is high-density imaging. However, in the high-density imaging, many fluorophores are activated at the same time so that there are many overlapping PSFs at each snapshot. There have been several approaches to resolve the overlapping PSFs. For example, DAOSTORM algorithm [4] iteratively fits overlapping spots in a greedy manner. CSSTORM (compressed sensing STORM) [5] and DeconSTROM [6] solve this problem as a sparse recovery among which the latter approach has been demonstrated to be more efficient for high-density imaging in terms of localization accuracy as well as molecular recall rates. For example, in CSSTORM, Gaussian noise model with sparsity constraint is assumed, which is solved by linear programming. As linear programming is computationally expensive, it adopts the local approach in which a reconstructed image is divided into several small-sized blocks processed individually, which potentially degrades the localization accuracy. In DeconSTORM, they use a modified Lucy-Richardson deconvolution in order to utilize Poisson statistics and temporal correlation of activated fluorophores.

In this paper, we present a new localization algorithm for high-density imaging by using a maximum likelihood estimation with a sparsity constraint, which is extremely fast compared to the existing approaches due to perfectly parallelizable structure. Using both simulation and real experiment, we confirmed that the proposed algorithm is especially robust in high-density live-cell imaging at low SNR by low emitted photons from activated fluorophores and high background level.

II. CCCP Framework using Generalized Huber Penalty

A. Problem Formulation

Let $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{r} \in \mathbb{R}$ and $\mathbf{y} \in \mathbb{R}^m$ denote the unknown fluorophore distribution, background fluorescence signals, and detector measurements, respectively; and $A = [a_{ij}]_{i,j=1}^{m,n}$ denotes the probability matrix that a emission photon from a voxel is detected at a detector position. Then, the negative loglikelihood function from Poisson intensity measurement is given by:

$$L(\mathbf{x}) \quad = \quad \mathbf{1}^T(A\mathbf{x}+\mathbf{r}) - \mathbf{y}^T \log(A\mathbf{x}+\mathbf{r}) \qquad (1)$$

where $\mathbf{1}$ denotes a vector with elements of ones with an appropriate size and $\log(A\mathbf{x}+\mathbf{r})$ is treated as element by element operation. Then, our superresolution imaging problem can be formulated as the following minimization problem:

$$\min_{\mathbf{x}} J(\mathbf{x}) \qquad \text{where} \quad J(\mathbf{x}) = L(\mathbf{x}) + pen(\mathbf{x}) \ , \qquad (2)$$

where the penalty function $pen(\mathbf{x})$ imposes a penalty to guide the reconstruction. Note that the optimization problem is not trivial since 1) the gradient of $L(\mathbf{x})$ is non-Lipschitz, and 2) each element of $\mathbf{x}$ should be nonnegative. Another technical difficulties in minimizing $L(\mathbf{x})$ in Eq. (1) is the existence of non-separable term in the likelihood, *i.e.* $\log(A\mathbf{x})$. Quadratic approximation [7] or Anscombe transform [8] was used to make this separable. Recently, for the case of Poisson image deconvolution using total variation (TV) or frame based analysis/synthesis penalty, Figueiredo and Bioucas-Dia [9] proposed so-called PIDAL algorithm using alternating direction of method of multiplier (ADMM) without approximating the Poisson loglikelihood. However, these ADMM algorithm requires huge additional memory to store the Lagrangian parameters to deal with non-separability of loglikelihood and the non-negativeness of $\mathbf{x}$.

To overcome these issues, this paper proposes a new optimization algorithm using the concave-convex procedure [10], which does not need any approximation of the cost function, or to store additional Lagrangian parameters. CCCP is a special case of majorized-minimzation algorithm, which utilizes the Legendre-Fenchel transform as a majorization function.

Legendre-Fenchel Transform of the Penalty: More specifically, as a sparsity inducing penalty, consider the following:

$$\|\mathbf{x}\|_{\mu,p} = \sum_{j=1}^{n} h_{\mu,p}(x_j), \quad 0 < p \leq 1. \tag{3}$$

where the generalized p-Huber function $h_{\mu,p}(t)$ is defined as

$$h_{\mu,p}(t) = \begin{cases} |t|^2/2\mu, & \text{if } |t| < \mu^{1/(2-p)} \\ |t|^p/p - \delta & \text{if } |t| \geq \mu^{1/(2-p)} \end{cases} \tag{4}$$

and where $\delta = (1/p - 1/2)\mu^{p/(2-p)}$ to make the function continuous and differentiable [11]. Note that for $p < 1$ the prior is non-convex. For the generalized p-Huber function in Eq. (4), it is easy to show that $|t|^2/\mu - h_{\mu,p}(t)$ is strictly convex [12]. Therefore, the Legendre-Fenchel transform tells us that there exist $g_{\mu,p}$ such that

$$h_{\mu,p}(t) = \min_{s}\{|s-t|^2/\mu + g_{\mu,p}(s)\} \ . \tag{5}$$

Chartrand [13], [11] showed that $g_{\mu,p}(s)$ is convex when $p = 1$, but in general it is not convex. However, even when $g_{\mu,p}(s)$ is non-convex, $|s|^2/\mu + g_{\mu,p}(s)$ is convex and Eq. (5) becomes a convex minimization problem with respect to s that has a closed form expression for the minimizer given as

$$\begin{aligned} \text{shrink}_p(t,\mu) &:= \arg\min_{s}\{|s-t|^2/\mu + g_{\mu,p}(s)\} \\ &= \max\{0, |t| - \mu|t|^{p-1}\}t/|t| \ . \end{aligned} \tag{6}$$

Here, $p \in [0,1]$ in which $p = 0$ is similar to hard thresholding and $p = 1$ is the same as soft thresholding [11].

Legendre-Fenchel Transform of the Negative Loglikelihood: Note that the negative loglikelihood term for the Poisson noise in Eq. (1) is convex. However, to deal with the existence of non-separable term in the likelihood, we utilize the CCCP with the help of a concave coordinate transform. More specifically, using an appropriate coordinate transform and application of Legendre-Fenchel transform, we can show that

$$\begin{aligned} L(\mathbf{x}) &= \min_{\mathbf{v}} L_c(\mathbf{x}, \mathbf{c}) \\ L_c(\mathbf{x}, \mathbf{c}) &= \sum_{i=1}^{m}\left(\sum_{j=1}^{n} a_{ij}x_j + c_{ij}\log\frac{c_{ij}}{a_{ij}x_j} - y_i\log(y_i)\right) \\ &\quad + \sum_{i=1}^{m}\left(r_i + c_i\log\frac{c_i}{r_i}\right) \end{aligned} \tag{7}$$

and $c_i + \sum_{j=1}^{m} c_{ij} = y_i$.

B. Optimization Framework

Now, we have the following minimization problem:

$$\min_{\mathbf{x},\mathbf{c},\mathbf{w}} L_c(\mathbf{x},\mathbf{c}) + \lambda\sum_{j=1}^{n}\left(\frac{1}{\mu}\|x_j - w_j\|^2 + g_{\mu,p}(w_j)\right) \ , \tag{8}$$

1) Minimization with respect to $\mathbf{w}$*:* Using the shrinkage relationship in Eq. (6), the close form solution is given by

$$w_j^{(k+1)} = \text{shrink}_p(x_j^{(k)}, \mu).$$

2) Minimization with respect to $\mathbf{c}$*:* The minimization problem has been studied by Hsiao *et al* [14] using Lagrangian for the constraint $c_i + \sum_{j=1}^{n} c_{ij} = y_i$ and it has been shown that we have the following closed form solution for the constrained optimization problem [15]:

$$c_{ij}^{(k+1)} = \frac{y_i a_{ij} x_j^{(k)}}{\sum_{j'=1}^{n} a_{ij'}x_{j'}^{(k)} + r_i}, \quad c_i^{(k+1)} = \frac{y_i r_i}{\sum_{j'=1}^{n} a_{ij'}x_{j'}^{(k)} + r_i}, \tag{9}$$

3) Minimization with respect to $\mathbf{x}$*:* Finally, for given $\mathbf{c}^{(k+1)}$ and $\mathbf{w}^{(k+1)}$, we can obtain a closed form solution for the update of $\mathbf{x}^{(k+1)}$. More specifically, a fixed point equation of the gradient of the cost function with respect to x_j satisfies the following second order polynomial:

$$0 = \sum_i a_{ij} - \frac{\sum_i c_{ij}^{(k+1)}}{x_j} + \frac{\lambda}{\mu}(x_j - w_j^{(k+1)}) \ , \tag{10}$$

Define

$$d := \frac{\lambda}{\mu}, \quad b_j^{(k+1)} = \left(\sum_i a_{ij}\right) - dw_j^{(k+1)}.$$

Then, the closed form solution is given by

$$x_j^{(k+1)} = \frac{-b_j^{(k+1)} + \sqrt{(b_j^{(k+1)})^2 + 4dx_j^{EM(k+1)}\sum_{i=1}^{m} a_{ij}}}{2d} \tag{11}$$

where $x_{ns}^{EM(k+1)}$ is similar to an ML-EM update given by

$$x_j^{EM(k+1)} = \frac{x_j^{(k)}}{\sum_{i=1}^{m} a_{ij}}\sum_{i=1}^{m}\frac{a_{ij}y_i}{\sum_{j'=1}^{n} a_{ij'}x_{j'}^{(k)} + r_i}. \tag{12}$$

Note that the solution is always non-negative, satisfying the positivity constraint. Moreover, our update equation is a pixel-by-pixel update similar to ML-EM algorithm or Lucy-Richardson method.

4) Advantages of the Proposed Method: Compared to a PI-DAL type algorithm, the proposed method has a unique advantage well-matched to super-resolution localization microscopy. The additional memory requirement for a Lagrangian approach is eliminated. Indeed, the computational complexity and memory usage during the calculation of x_j is similar to the standard Lucy-Richardson deconvolution method. Hence, the algorithm lends itself to a fast GPU implementation thanks to the efficient memory utilization and pixel-by-pixel update. On a intel i7 920 (CPU) and a Tesla C1060 (GPU), our GPU implementation of the proposed method takes only 5 seconds in reconstructing a five over-sampled image of a 128×128 CCD image with 1500 iterations.

III. Experimental Results

We performed experiments using simulated data and real high-density live-imaging PALM data. We compared the following algorithms: the least-square Gaussian fitting[1], CSSTORM[5], FISTA[16] using l1 norm, and the proposed algorithm. In CSSTORM, FISTA, and the proposed algorithm, uniform background is assumed and estimated. In the least-squares method, an elliptical Gaussian PSF to local maxima of the image is fitted.

A. Simluation

In the simulated data, each nanoscale molecule provides a Gaussian PSF of 340nm full width half maximum (FWHM). Emitted photons of the molecules follow the log-normal distribution with mean of 500 and standard deviation of 100. In order to generate low-SNR data, 70 background photons are added to every CCD pixel of 100 nm. In addition, we introduced Poisson shot noise and Gaussian readout noise with unit variance. We generated a data set of a wide range of imaging densities, from $0.2\mu m^{-2}$ to $3.4\mu m^{-2}$. At each density level, flurophores are generated at random locations within 40×40 pixel image, and the total of 30 realizations were used. To quantify the error, each true molecular positions are matched to the closest localized fluorophore within 200nm radius. Then, we calculated the standard deviation of the localization errors and the molecular recall rates.

In all ranges of the density level, the proposed algorithm demonstrated better recall rates than the others. Specifically, the proposed algorithm can identify 10 times more fluorophore molecules than the least-squares method, and improve about 10-30% compared to CSSTROM and FISTA. Moreover, the proposed method is much accurate than CSSTROM & FISTA in terms of localization accuracy. While the least squares method have the smallest localization error, these errors are only for the corrected identified fluoresphores, whose number is significantly smaller compared to others. Therefore, this confirmed that the proposed algorithm is more effective in low-SNR & high-density imaging data than the conventional approaches.

B. Live-cell Super-resolution Imaging

U2OS cells were maintained in Dulbecco's Modified Eagles's Medium (DMEM) (Gibco) supplemented with 10% Fetal Bovine Serum (FBS) (Gibco) in an atmosphere containing 5% CO2 at 37 °C. Cells were cultured and maintained in T-25 flasks and grown to about 70% confluency (corresponding to 2 days) before they were passaged. Prior to staining, cells were washed once with PBS (Sigma). A 200 nM dilution of Mitotracker Red CMH2XRos (Invitrogen) was made in Leibovitz (Invitrogen) and labelled in an inner membrane of the mitochondria. Cells were incubated with the dye for 1520 min at 37 °C in a CO2 atmosphere.

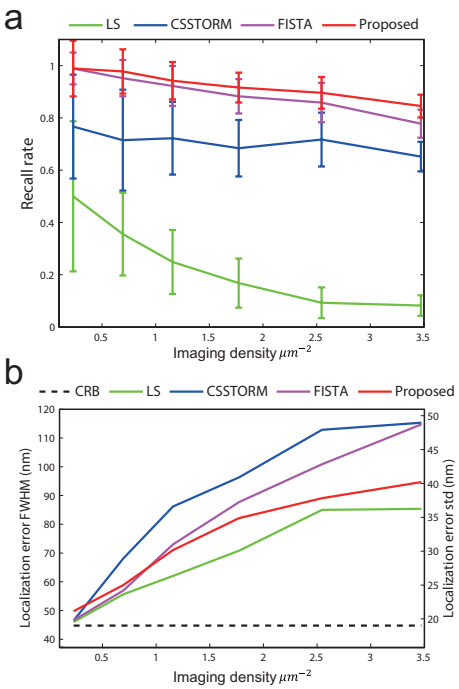


Fig. 1. Simulation analysis on low-SNR STORM/PALM data. (a) Localization error. (b) Identified molecular density. Cramer-Rao bound(CRB) is theoretical minimum accuracy of single molecule localization.

Imaging was performed on an inverted microscope (IX71, Olympus), equipped with an oil-immersion objective (UPlanSAPO 100 x, NA=1.40, Olympus). A 561 nm (Sapphire 561, Coherent) was used to excite Mitotracker Red CMH2XRos and fluorescence was directed onto an electron multiplying CCD camera (iXon+, Andor) with a resulting pixel size of 100 nm. The laser intensity was approximately 3 kW cm $^{-2}$ and an ET605/70 (Chroma) emission filter was used. 4000 frames were collected with a 20 ms exposure time per frame. Using the experimental data, we compared the reconstruction results using the three algorithms (Least-squares, FISTA, and the proposed). The proposed algorithm localized 30 % more molecules than FISTA and 8 times more than the least-squares. In figure 2(b-d), the proposed results show better internal matrix structure and have much clear boundaries of mitochondria than the others. Moreover, the size of the reconstructed mitochondria using FISTA (c) seems to be reduced compared to that of the proposed method (d). In order to observe the dynamic of mitochondria, we created time-lapse images (e,f). Every images in (e,f) were generated from the 1000 consecutive CCD frames for 20 sec and the time-gap

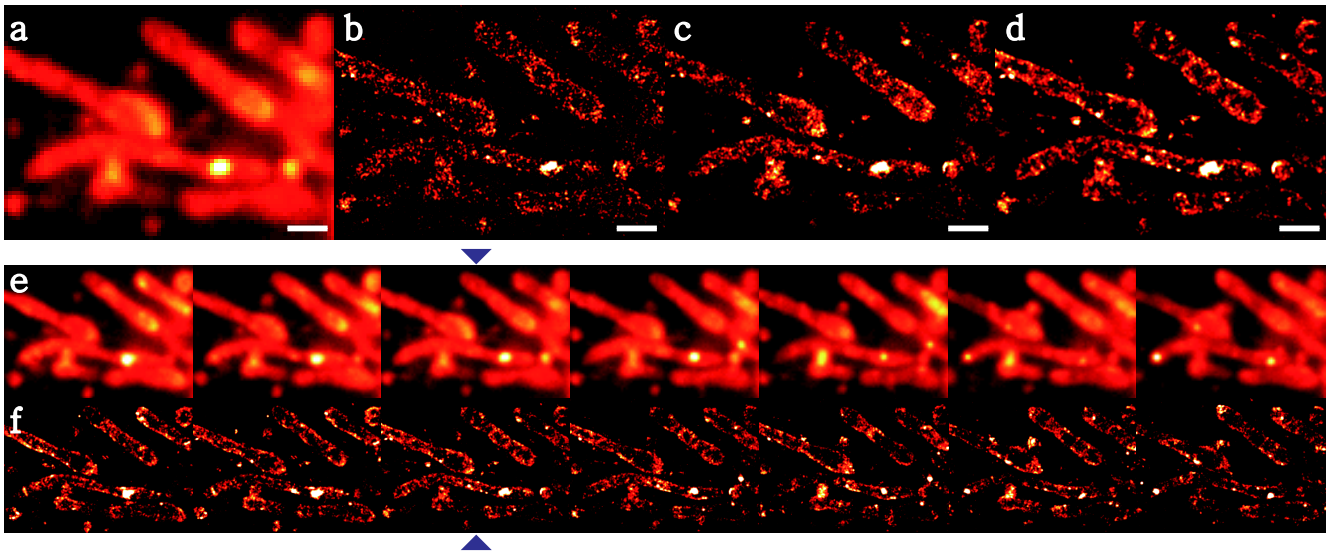


Fig. 2. Live-cell imaging of Mitochondria. Inner membrane of the Mitochondria was labbeled by Mitotracker. (a) Conventional image. (b)Least-square fitting. (c) FISTA decon. (d)Proposed. (e,f) Conventional and Proposed time-lapse images. Every image is generated from consecutive 1000 CCD frames(20sec) and time-gap is 10sec. (a-d) images correspond to blue maker in (e,f). Scale-bar in (a-d) is $1\mu m$.

between successive acquisitions was 10 sec. In the time-lapse images (e-f), we observed slow movements of mitochondria.

IV. Conclusion

We present a new localization algorithm for high-density super-resolution microscopy using the maximum-likelihood estimation of the Poisson noise model with sparsity promoting prior. Using concave-convex procedure, a highly parallelizable algorithm has been derived, which results in a fast GPU implementation with speed of 5 sec per frame. We demonstrated that our algorithm is much effective in low-SNR PALM data over wide range of imaging density in terms of recall rate and localization accuracy. Therefore, we expect that the proposed approach can significantly reduce the number of required CCD frames for super-resolution imaging, which can improve the temporal resolution significantly. Thus, our approach is appropriate for live-cell imaging to investigate biological interactions at the nanometer scale.

Acknowledgment

This research was supported by the National Research Foundation of Korea(NRF) grant funded in part by the Korean government (MEST) (No.2011-0030933) and (No.2012-0000173)

References

[1] M. J. Rust, M. Bates, and X. Zhuang, "Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM)," *Nature Methods*, vol. 3, no. 10, pp. 793–796, 2006.

[2] E. Betzig, G. H. Patterson, R. Sougrat, O. W. Lindwasser, S. Olenych, J.S. Bonifacino, M. W. Davidson, J. Lippincott-Schwartz, and H. F. Hess, "Imaging intracellular fluorescent proteins at nanometer resolution," *Science*, vol. 313, no. 5793, pp. 1642, 2006.

[3] Carlas S. Smith., Nikolai Joseph., Bernd Rieger., and Keith A. Lidke, "Fast, single-molecule localization that achieves theoretically minimum uncertainty," *Nature Methods*, vol. 7, no. 5, pp. 373–375, 2010.

[4] Seamus J Holden., Stephan Uphoff., and Achillefs N Kapanidis, "DAOSTORM: an algorithm for high-density super-resolution microscopy," *Nature Methods*, vol. 8, no. 4, pp. 279–280, 2011.

[5] L. Zhu, W. Zhang, D. Elnatan, and B. Huang, "Faster STORM using compressed sensing," *Nature Methods*, vol. 9, no. 7, pp. 721–723, 2012.

[6] Eran A Mukamel, Hazen Babcock, and Xiaowei Zhuang, "Statistical deconvolution for superresolution fluorescence microscopy," *Biophysical Journal*, vol. 102, no. 10, pp. 2391–2400, 2012.

[7] Z. Harmany, R. Marcia, and R. Willett, "This is SPIRAL-TAP: sparse Poisson intensity reconstruction algorithms - theory and practice," *IEEE Transactions on Image Processing*, , no. 99, pp. 1084–1096, 2010.

[8] F.X. Dupé, J.M. Fadili, and J.L. Starck, "A proximal iteration for deconvolving Poisson noisy images using sparse representations," *IEEE Transactions on Image Processing*, vol. 18, no. 2, pp. 310–321, 2009.

[9] M.A.T. Figueiredo and J.M. Bioucas-Dias, "Restoration of Poissonian images using alternating direction optimization," *IEEE Transactions on Image Processing*, vol. 19, no. 12, pp. 3133–3145, 2010.

[10] A.L. Yuille and A. Rangarajan, "The concave-convex procedure," *Neural Computation*, vol. 15, no. 4, pp. 915–936, 2003.

[11] R. Chartrand, "Nonconvex Splitting for Regularized Low-Rank Sparse Decomposition," *Los Alamos National Laboratory Report: LA-UR-11-11298*, 2012.

[12] S.P. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[13] R. Chartrand and V. Staneva, "Restricted isometry properties and nonconvex compressive sensing," *Inverse Problems*, vol. 24, pp. 035020.1–035020.14, 2008.

[14] I.T. Hsiao, A. Rangarajan, P. Khurd, and G. Gindi, "An accelerated convergent ordered subsets algorithm for emission tomography," *Physics in Medicine and Biology*, vol. 49, pp. 2145–2156, 2004.

[15] D.P. Bertsekas, "Constrained optimization and Lagrange multiplier methods," *Computer Science and Applied Mathematics*, vol. 1, 1982.

[16] A. Beck and M. Teboulle, "Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems," *Image Processing, IEEE Transactions on*, vol. 18, no. 11, pp. 2419–2434, 2009.

Analogies and differences in optical and mathematical systems and approaches

Bettina Heise
CDL MS-MACH
Johannes Kepler University
Linz, Austria
Email: bettina.heise@jku.at

Stefan E. Schausberger
CDL MS-MACH
Johannes Kepler University
Linz, Austria
Email: stefan.schausberger@jku.at

Martin Reinhardt
Institute Applied Analysis
TU Bergakademie Freiberg
Freiberg, Germany
Email: Martin.Reinhardt@math.tu-freiberg.de

David Stifter
CDL MS-MACH
Johannes Kepler University
Linz, Austria
Email: david.stifter@jku.at

***Abstract*—We review traditions and trends in optics and imaging recently arising by applying programmable optical devices or by sophisticated approaches for data evaluation and image reconstruction. Furthermore, a short overview is given about modeling of well-known classical optical elements, and vice versa, about optical realizations of classical mathematical transforms, as in particular Fourier, Hilbert, and Riesz transforms.**

I. Introduction

In the 18th/19th century the work of physicists and mathematicians was often closely connected. Scientists in that age were often acting concurrently in both fields: if we think about e.g. Augustin Fresnel explaining experimentally and theoretically the phenomena of light propagation and diffraction, or about Joseph Fourier, experimentally discovering mode decomposition of (mechanical) wave fields and delivering the basis for later theory about transforming signals and fields into the (temporal or spatial) frequency domain. In the nearer past both disciplines were developing rather independently in their own directions. In the field of optics important discoveries as the laser, wave-guides, novel microscopic or holographic techniques should be named as examples among others. In the field of mathematics the huge field of harmonic analysis, bringing up wavelets, frames etc., the several numerical approaches for solving differential equations and also the development of the functional theoretic background in analysis should be quoted as representatives here.

II. Optical devices and mathematical descriptions; Mathematical approaches and optical realizations

A. Analogies between optical and mathematical approaches

Due to the contemporary possibilities given on one hand by advanced digital optical devices, as spatial light modulators (SLM) or micro mirror arrays (MMA), deformable mirrors or phased arrays in combination with traditional optical elements, and on the other hand by the computational power of modern hard- and software architecture allowing sophisticated mathematical reconstruction algorithms, new fields of research and perspectives are opened. Computational or programmable optics are examples for this modern development and interdisciplinary entanglement of the different disciplines. They open a new branch of methods as for digital holography, lensless microscopy, or adaptive optics [1]–[3]; they comprehend several phase retrieval and reconstruction techniques [4], [5], adaptive wave front correction methods up to compressive sensing for optical applications [6], [7]. Whereas in the past optical imaging performance has often been hampered by scattering within materials, by turbulences within fluids, or speckles at rough surfaces, nowadays computational techniques and programmable optics deliver novel approaches as focusing through or within scattering materials, turbulence corrections or contrast enhancement by SLM-based techniques, [8]–[10].

Bringing now together optics and mathematics in such a way, touching points are noticed and furthermore, awareness is arising that in both fields similar approaches exist, only realization techniques or names may differ. This concerns for instance classical optical devices (as lenses, prism, cones,...) or classical imaging techniques (bright field, dark field, Schlieren or knife edge imaging technique, spiral phase quadrature imaging, or differential interference contrast (DIC) imaging), [11]–[14]. Primarily, these techniques are modifying contrast of the visualized specimen, but to a certain amount they are also quantitative with respect to phase or optical path length, which can be expressed and reconstructed mathematically under knowledge of their (complex-valued) point spread function (PSF) in the spatial domain or of their optical transfer function in the Fourier domain.

Beyond the well-known Fourier transform (FT) other classical mathematical transforms as the two-dimensional (2D) Hilbert transform (HT), also denoted as directional HT [15], with a kernel function H_{HT} defined in the Fourier do-

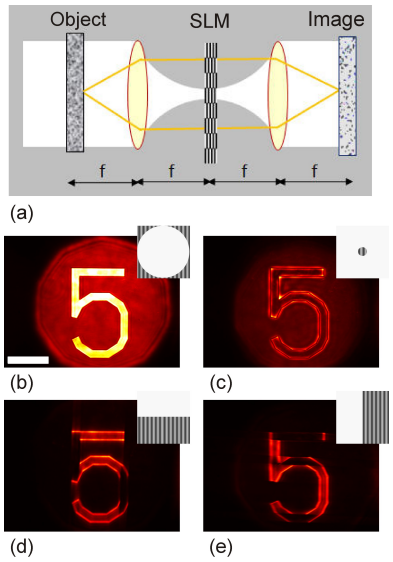


Fig. 1. Different contrast modifications emulated by means of the SLM device being addressed with different filter functions in a Fourier plane filtering unit (a). The emulated imaging type shows as representative contrast: (b) bright field (original scan), (c) dark field, (d) and (e) horizontal and vertical Schlieren/knife edge imaging contrast. (Striped regions encode a zero magnitude, continuous regions encode a constant unit magnitude, and a phase between [0, 2 π] according to the gray level. The white scale bar yields 100 μm.)

main (u, v) by

$$H_{HT_1}(u,v) = -\text{sgn(u)}\exp(il\pi/2), \quad (1)$$
$$H_{HT_2}(u,v) = -\text{sgn(v)}\exp(il\pi/2), \quad (2)$$

where $l = 1$ is chosen for the conventional HT, or as the 2D Riesz transform (RT), also denoted as complexified-valued Riesz transform [15] or radial Hilbert transform [16], with a kernel function H_{RT} defined in the Fourier domain by

$$H_{RT}(\hat{r}, \hat{\varphi}) = \exp(il\hat{\varphi}), \quad (3)$$

with $(\hat{r}, \hat{\varphi})$ denoting polar coordinates in the Fourier domain, and $l = 1$ is chosen for the conventional RT, find entrance in optical modeling, emulation, and settings. Optically these transforms can be realized by classical elements (lenses, apertures, spiral phase plates) or nowadays more and more by programmable SLM devices allowing flexible realizations. Vice versa, in the mathematical modeling of optical imaging techniques these transforms build the base for an (approximated) description of the PSF e.g. for Schlieren and DIC imaging, for pyramid and roof sensors (all with a PSF model based on the directional HT), [17], [18]. Also spiral phase/vortex filtering (with a PSF model based on the RT), and their fractional expressions as fractional half-plane and spiral phase filters (corresponding to a fractional HT resp. RT with $0 < l < 1$ in eq. (2) and (3)) can be modeled in such a way, as shown in Fig.1 and Fig.2, [19]–[23].

Here we can connect now optics with classical functional analysis. The PSF of a pyramid sensor [17] given in the spatial

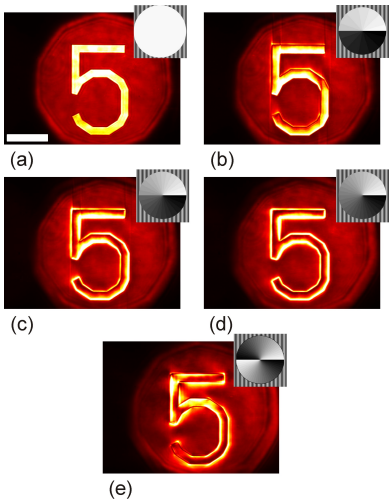


Fig. 2. Contrast modifications emulated by means of the SLM device being addressed with fractional spiral phase filter functions (fractional Riesz transform) of fractional coefficient (a) l=0, (b) l=0.4, (c) l=0.8, (d) l=1.0, (e) l=2.

domain (x, y) by

$$h_{PY}((-1)^n x, (-1)^m y) = \frac{1}{4}\delta(x,y) \quad (4)$$
$$+ \quad \frac{(-1)^{m+n}}{4\pi^2} p.v.\frac{1}{xy}$$
$$+ \quad \frac{i}{4\pi}\left[(-1)^n p.v.(\frac{1}{x}\delta(y)) + (-1)^m p.v.(\delta(x)\frac{1}{y})\right],$$

with p.v. denoting a principal value and (n, m) are enumerators $(0, 1)$, resembles in its structure the 2D analytic signal, as introduced by Hahn [24]. Whereas the PSF of a spiral phase filter or so called vortex filter [22] can be described by the (2D) Riesz kernel

$$h_{SP}(r, \varphi) = \frac{i}{2\pi r^2}\exp(il\varphi), \quad (5)$$

with (r, φ) denoting polar coordinates in spatial domain, and $l = 1$ is chosen in the conventional case. Furthermore, it should be noted that Riesz transform has been introduced by [25] in the field of optics under the name spiral phase quadrature transform. This filter tends rather to a monogenic signal approach, as introduced by Felsberg, [26]. Knowing now in principle the PSF of these imaging modalities, we can emulate the special imaging types by addressing SLMs in a corresponding way with amplitude or phase transfer functions in optical Fourier domain. So we can flexibly change the contrast corresponding to the envisaged imaging technique [27] and can go towards a quantitative reconstruction based on the emulated PSF in future.

B. Optical Fourier plane filtering and wavelet-like filters

In applied mathematics and signal analysis orthogonal, isotropic or anisotropic wavelet-based decomposition approaches play an important role for image processing, naming applications as image denoising, edge enhancement, or

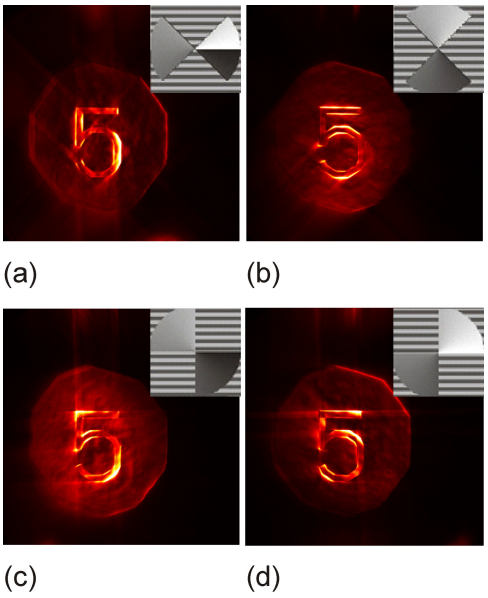


Fig. 3. The original image has been filtered optically in Fourier domain by different monogenic wavelet-like filters, the applied filter kernels are sketched in the inset. (Striped regions encode a zero magnitude, continuous regions encode a constant unit magnitude and a phase between [0, 2 π] according to the gray level.)

compression methods among other. In particular, analytic or monogenic wavelet approaches have found entrance in image processing delivering additional phase and orientation information or may be used for scale-based demodulation [28]–[32].

On the other hand we can also ask whether and to which extent a wavelet-like filtering can be performed analogously in an optical way. Here the classical principle for optical Fourier plane filtering finds its modern application anew. In combination with programmable optics as SLMs or MMAs we also can emulate to a certain amount the (compact and positive) support and transfer function of suitable wavelets (curvelets, shearlets) in Fourier domain. And for their analytic and monogenic wavelet complements, also in this case, optical realizations of Hilbert and Riesz transform build up the basis for the filtering approaches. Here, the methods are usable for isotropic or anisotropic contrast improvement in imaging [33], as shown in Fig.3, for orientation emphasizing, or for salient point detection.

C. Restrictions and differences between optical and mathematical approaches

However, we also should keep in mind the restrictions and differences between optical realizations and mathematical approaches. For instance SLMs or MMAs as pixelated and discrete arrays exhibit only a finite resolution; therefore, the available spatial frequency range is restricted for filtering. Furthermore, in optics without introducing additional sensors or working with interferometric imaging setups, we can only measure intensities at a conventional camera applied as detector. So the separated information given in the amplitude and phase spectrum - as easily obtained by Fourier transform in mathematics - is lost in their optical counterparts. At least for phase reconstruction an additional phase retrieval step using a multiple recording of the modified image would be required.

Coherence aspects in optical Fourier plane filtering provides an additional discussion point. Coherence may be regarded as an imaging feature closely related to the considered scale. Furthermore, it must be distinguished between temporal coherence and spatial coherence. Operating with broadband light sources for illumination, these sources exhibit a smaller temporal coherence length than conventional narrow band laser sources used in coherent imaging. Therefore, the phase filter applied on the SLM mask is exactly matching only for the central wavelength. This mismatch may result in a slight blurring of the image features such as edges. Spatial coherence can be maintained by coupling the illumination beam into a single mode fiber.

However, scattering within turbid materials severly restricts the fixed phase relationship within the electro-magnetic wave field required e.g. for the Fourier plane (phase-only) filtering (or correspondingly within a convolution kernel of a defined support). This demands again methods for wavefront correction to cope with scattering materials for future successful implementations.

III. Conclusion

In summary, the close connection between the modeling of well-known optical devices or elements and classical mathematical approaches or transforms has been demonstrated. Furthermore, by linear filtering in optics we can realize similar effects as with classical filtering in signal or image processing. The explanation of the obtained effects in optics and in mathematics is partly similar, but due to the complex-valued nature of the light also different mechanism, as e.g. interference or diffraction has to be considered.

Acknowledgment

We want to thank Swanhild Bernstein at TU Bergakademie Freiberg, Iuliia Shatokhina at JKU Linz for the interesting discussions, as well as Monika Ritsch-Marte and her team for the support in the field of SLM techniques. The financial support by the Federal Ministry of Economy, Family and Youth, the National Foundation for Research, Technology and Development is gratefully acknowledged.

References

[1] M. K. Kim, *Digital Holographic Microscopy,* Springer Series in Optical Sciences, 162, (2011).

[2] A. Greenbaum, W. Luo, T-W. Su, Z. Grcs, L. Xue, S. O. Isikman, A. F. Coskun, O. Mudanyali, A. Ozcan, *Imaging without lenses: achievements and remaining challenges of wide-field on-chip microscopy,* Nature Methods 9(9), 889-895 (2012).

[3] A. Rooda, *Adaptive optics for studying visual function: A comprehensive review,* J. Vision 11(5), 1-21 (2011).

[4] J. R. Fienup, *Phase retrieval algorithms: a personal tour,* Appl. Opt. 52, 45-56 (2013).

[5] M. Kanka, R. Riesenberg, H. J. Kreuzer, *Reconstruction of high-resolution holographic microscopic images,* Opt. Lett. 34, 1162-1164 (2009).

[6] T. Cismar, M. Mazilu, K. Dholakia, *In situ wavefront correction and its application to micromanipulation,* Nature Photonics 4, 388-394 (2010).
[7] R. M. Willett, R. F. Marcia, J. M. Nichols, *Compressed sensing for practical optical imaging systems: a tutorial,* Optical Engineering 50(7), 072601-1-13 (2011).
[8] S. Popoff, G. Lerosey, M. Fink, A. C. Boccara, S. Gigan, *Image Transmission Through an Opaque Material,* arXiv:1005.0532v2, (2010).
[9] Z. Hajjarian, M. Kavehrad, J. Fadlullah, *Spatially multiplexed multi-input multi-output optical imaging system in a turbid, turbulent atmosphere,* Appl. Opt. 49(9), 1528-1538 (2010).
[10] S. E. Schausberger, B. Heise, C. Maurer, S. Bernet, M. Ritsch-Marte, D. Stifter, *Flexible contrast for low-coherence interference microscopy by Fourier-plane filtering with a SLM,* Opt. Lett. 35, 4154-4156 (2010).
[11] S. K. Tiwari, S. R. Mishra, S. P. Ram, H. S. Rawat, *Generation of a Bessel beam of variable spot size,* Appl. Opt. 51(17), 3718-3725 (2012).
[12] P. K. Panigrahi, K. Muralidhar, *Schlieren and Shadowgraph Methods in Heat and Mass Transfer,* (Springer Briefs in Applied Sciences and Technology/Springer Briefs in Thermal Engineering and Applied Science), (2012).
[13] S. Fürhapter, A. Jesacher, S. Bernet, M. Ritsch-Marte, *Spiral phase contrast imaging in microscopy,* Opt. Express 13, 689-694 (2005).
[14] C. Preza, *Rotational-diversity phase estimation from differential-interference-contrast microscopy images,* J. Opt. Soc. Am. A 17(3), 415-424 (2000).
[15] M. Unser, M. Sage, D. Van de Ville, *Multiresolution monogenic signal analysis using the Riesz-Laplace wavelet transform,* IEEE Trans. Image Process. 18, 2402 (2009).
[16] J. A. Davis, D. E. McNamara, D. M. Cottrell, *Image processing with the radial Hilbert transform: theory and experiments*, Opt. Lett. 25, 99 (2001).
[17] V. Korkiakoski, C. Verinaud, M. Louarn, R. Conan, *Comparison between a model-based and a conventional pyramid sensor reconstructor,* Appl. Opt. 46(24), 6176-6184 (2007).
[18] M. R. Arnison, C. J. Cogswell, N. I. Smith, P. W. Fekete, K. G. Larkin, *Using the Hilbert transform for 3D visualization of differential interference contrast microscope images,* J. Microscopy, 99 (1), 79-84 (2000).
[19] H. M. Ozaktas, Z. Zalevsky, M. A. Kutay, *The fractional Fourier transform,* Wiley (2001).
[20] A. W. Lohmann, D. Mendlovic, Z. Zalevsky, *The fractional Hilbert transform*, Opt. Lett. 21(4), 281-283 (1996).
[21] Q. Xie, D. Zhao, *Generation of dark hollow beams by using a fractional radial Hilbert transform system,* Optics Communications 275, 394-398 (2007).
[22] S. Fürhapter, A. Jesacher, C. Maurer, S. Bernet, M. Ritsch-Marte, *Spiral Phase Microscopy,* Advances in Imaging and Electron Physics 146, 1-56 (2007).
[23] G. Situ, G. Pedrini, W. Osten, *Spiral phase filtering and orientation-selective edge detection/ enhancement,* J. Opt. Soc. Am. A 26, 1788-1797 (2009).
[24] S. L. Hahn, *Multi-dimensional complex signals with single-orthant spectra,* Proc. IEEE 80, 1287-1300 (1992).
[25] K. G. Larkin, D. J. Bone, M. Oldfield, *Natural demodulation of two-dimensional fringe patterns. I. General background of the spiral phase quadrature transform,* J. Opt. Soc. Am. A, 18, 1862-1870 (2001).
[26] M. Felsberg, G. Sommer, *The monogenic signal,* IEEE Trans. Signal Proc. 49, 3136-3144 (2001).
[27] B. Heise, S. E. Schausberger, D. Stifter, *Coherence Probe Microscopy: Contrast Modification and Image enhancement,* Imaging & Microscopy 2, 29-32 (2012).
[28] M. Storath, *Directional Multiscale Amplitude and Phase Decomposition by the Monogenic Curvelet Transform,* SIAM J. Imaging Sciences 4, 57-78 (2011).
[29] M. Unser, D. Sage, D. V. D. Ville, *Multiresolution monogenic signal analysis using the Riesz-Laplace wavelet transform,* IEEE Trans. Img. Proc. 18, 2402-2418 (2009).
[30] C. S. Seelamantula, N. Pavillon, C. Depeursinge, M. Unser, *Exact complex wave reconstruction in digital holography,* J. Opt. Soc. Am. A 28, 983-992 (2011).
[31] C. S. Seelamantula, N. Pavillon, C. Depeursinge, M. Unser, *Local demodulation of holograms using the Riesz transform with application to microscopy,* J. Opt. Soc. Am. A 29, 2118-2129 (2012).
[32] S. E. Schausberger, B. Heise, S. Bernstein, D. Stifter, *Full-field optical coherence microscopy with Riesz transform-based demodulation for dynamic imaging,* Opt. Lett. 37, 4937-4939 (2012).
[33] B. Heise, S. E. Schausberger, C. Maurer, M. Ritsch-Marte, S. Bernet, D. Stifter, *Enhancing of structures in coherence probe microscopy imaging,* Proc. SPIE, 8335, 83350G (2012).

The Nyquist theorem for cellular sheaves

Michael Robinson
Department of Mathematics and Statistics
American University
4400 Massachusetts Ave NW
Washington, DC 20016
Email: michaelr@american.edu

***Abstract*—We develop a unified sampling theory based on sheaves and show that the Shannon-Nyquist theorem is a cohomological consequence of an exact sequence of sheaves. Our theory indicates that there are additional cohomological obstructions for higher-dimensional sampling problems. Using these obstructions, we also present conditions for perfect reconstruction of piecewise linear functions on graphs, a collection of non-bandlimited functions on topologically nontrivial domains.**

I. Introduction

The Shannon-Nyquist sampling theorem states that sampling a signal at twice its bandwidth is sufficient to reconstruct the signal. Its wide applicability leads to the question of whether there exist similar conditions for reconstructing other data from samples in more general settings. This article shows that perfect reconstruction for sampling of local algebraic data on simplicial complexes can be addressed through the machinery exact sequences of cellular sheaves. As a demonstration of our technique, we recover the Nyquist theorem and generalize it to perfect reconstruction of piecewise linear signals on graphs. Piecewise linear functions are not bandlimited, since their derivatives are not continuous.

A. Historical context

Sampling theory has a long and storied history, about which a number of recent survey articles [1], [2], [3], [4] have been written. Since sampling plays an important role in applications, substantial effort has been expended on practical algorithms. Our approach is topologically-motivated, like the somewhat different approach of [5], [6], so it is less constrained by specific timing constraints. Relaxed timing constraints are an important feature of bandpass [7] and multirate [8] algorithms. We focus on signals with local control, of which splines [9] are an excellent example.

Sheaf theory has not been used in applications until fairly recently. The catalyst for new applications was the technical tool of *cellular sheaves*, developed in [10]. Since that time, an applied sheaf theory literature has emerged, for instance [11], [12], [13], [14], [15].

Our sheaf-theoretic approach allows sufficient generality to treat sampling on non-Euclidean spaces. Others have studied sampling on non-Euclidean spaces, for instance general Hilbert spaces [16], Riemann surfaces [17], symmetric spaces [18], the hyperbolic plane [19], combinatorial graphs [20], and quantum graphs [21], [22]. We show that sheaves provide unified sufficiency conditions for perfect reconstruction on abstract simplicial complexes, which encompass all of the above cases.

A large class of local signals are those with *finite rate of innovation* [23], [24]. Our ambiguity sheaf is a generalization of the Strang-Fix conditions as identified in [25]. With our approach, one can additionally consider reconstruction using richer samples than simply convolutions with a function.

II. Cellular sheaves

A. What is a sheaf?

A sheaf is a mathematical object that stores locally-defined data over a space. In order to formalize this concept, we need a concept of space that is convenient for computations. The most efficient such definition is that of a simplicial complex.

Definition 1. An *abstract simplicial complex* X on a set A is a collection of ordered subsets of A that is closed under the operation of taking subsets. We call each element of X a *face*. A face with $k+1$ elements is called a k-face, though we usually call a 0-face a *vertex* and a 1-face an *edge*. The *face category* has as objects the elements of X and as morphisms inclusions of one element of X into another.

Although sheaves have been extensively studied over topological spaces (see [26] or the appendix of [27] for a modern, standard treatment), the resulting definition is ill-suited for application to sampling. Instead, we follow a substantially more combinatorial approach introduced in the 1985 thesis of Shepard [10].

Definition 2. A *sheaf* F on an abstract simplicial complex X is a covariant functor from the face category of X to the category of vector spaces. Explicitly,

- for each element a of X, $F(a)$ is a vector space, called the *stalk at* a,
- for each inclusion of two faces $a \to b$ of X, $F(a \to b)$ is a linear function from $F(a) \to F(b)$ called a *restriction*, and
- for every composition of inclusions $a \to b \to c$, $F(b \to c) \circ F(a \to b) = F(a \to b \to c)$.

Definition 3. Suppose F is a sheaf on an abstract simplicial complex X and that $\mathcal{U}$ is a collection of faces of X. An assignment s which assigns an element of $F(u)$ to each face

$u \in \mathcal{U}$ is called a *section* supported on $\mathcal{U}$ when for each inclusion $a \to b$ (in X) of objects in $\mathcal{U}$, $F(a \to b)s(a) = s(b)$. A *global section* is a section supported on X. If r and s are sections supported on $\mathcal{U} \subset \mathcal{V}$, respectively, in which $r(a) = s(a)$ for each $a \in \mathcal{U}$ we say that s *extends* r. The collection of sections supported on a given set forms a vector space.

Example 4. Consider $Y \subseteq X$ a subset of the vertices of an abstract simplicial complex. The functor S which assigns a vector space V to vertices in Y and the trivial vector space to every other face is called a *V-sampling sheaf supported on Y*. To every inclusion between faces of different dimension, S will assign the zero function. For a finite abstract simplicial complex X, the space of global sections of a V-sampling sheaf supported on Y is isomorphic to $\bigoplus_{y \in Y} V$.

Recall that an abstract simplicial complex X consists of *ordered* sets. For a a k-face and b a $k+1$-face, define

$$[b : a] = \begin{cases} +1 & \text{if the order of elements in } a \text{ and } b \text{ agrees,} \\ -1 & \text{if it disagrees, or} \\ 0 & \text{if } a \text{ is not a face of } b. \end{cases}$$

Example 5. Suppose G is a graph in which each vertex has finite degree (evidently G can be realized as an abstract simplicial complex). Let PL be the sheaf constructed on G that assigns $PL(v) = \mathbb{R}^{1+\deg v}$ to each edge v of degree $\deg v$ and $PL(e) = \mathbb{R}^2$ to each edge e. The stalks of PL specify the value of the function (denoted y below) at each face and the slopes of the function on the edges (denoted $m_1, ..., m_k$ below). To each inclusion of a degree k vertex v into an edge e, let PL assign the linear function $(y, m_1, ..., m_e, ..., m_k) \mapsto (y + [e : v]\frac{1}{2}m_e, m_e)$. The global sections of this sheaf are *piecewise linear functions* on G.

Definition 6. A *sheaf morphism* is a natural transformation between sheaves. Explicitly, a morphism $f : F \to G$ of sheaves on an abstract simplicial complex X assigns a linear map $f_a : F(a) \to G(a)$ to each face a so that for every inclusion $a \to b$ in the face category of X, $f_b \circ F(a \to b) = G(a \to b) \circ f_a$.

B. Sheaf cohomology

Much of the theory of sheaves is concerned with computing spaces of sections and identifying obstructions to extending sections. The machinery of cohomology systematizes the computation of the space of global sections for a sheaf.

Define the following formal *cochain* vector spaces $C^k(X; F) = \bigoplus_{a \text{ a } k\text{-face of } X} F(a)$. The *coboundary map* $d^k : C^k(X; F) \to C^{k+1}(X; F)$ takes an assignment s from the k faces to an assignment $d^k s$ whose value at a $k+1$ face b is

$$(d^k s)(b) = \sum_{a \text{ a } k\text{-face of} X} [b : a] F(a \to b) s(a).$$

It can be shown that $d^k \circ d^{k-1} = 0$, so that the image of d^{k-1} is a subspace of the kernel of d^k.

Definition 7. The k-th *sheaf cohomology* of F on an abstract simplicial complex X is

$$H^k(X; F) = \ker d^k / \text{image } d^{k-1}.$$

Observe that $H^0(X; F) = \ker d^0$ consists precisely of those assignments s which are global sections. Cohomology is also a functor: sheaf morphisms induce linear functions between cohomologies. This indicates that cohomology preserves and reflects the underlying relationships between sheaves.

III. The Nyquist criterion for sheaves

Suppose that F is a sheaf on an abstract simplicial complex X, and that S is a V-sampling sheaf on X supported on a closed subcomplex Y. A *sampling* of F is a morphism $s : F \to S$ that is surjective on every stalk. Given a sampling, we can construct the *ambiguity sheaf* A in which the stalk $A(a)$ for a face $a \in X$ is given by the kernel of the map $F(a) \to S(a)$. If $a \to b$ is an inclusion of faces in X, then $A(a \to b)$ is $F(a \to b)$ restricted to $A(a)$. This implies that

$$0 \to A \hookrightarrow F \xrightarrow{\quad s \quad} S \to 0$$

is an exact sequence, which induces the long exact sequence (via the Snake lemma)

$$0 \to H^0(X; A) \to H^0(X; F) \to H^0(X; S) \to H^1(X; A) \to$$

An immediate consequence is therefore

Corollary 8. *(Sheaf-theoretic Nyquist theorem) The global sections of F are identical with the global sections of S if and only if $H^k(X; A) = 0$ for $k = 0$ and 1.*

The cohomology space $H^0(X; A)$ characterizes the *ambiguity* in the sampling, while $H^1(X; A)$ characterizes its *redundancy*. Optimal sampling therefore consists of identifying minimal closed subcomplexes Y so the resulting ambiguity sheaf A has $H^0(X; A) = H^1(X; A) = 0$.

Let us place bounds on the cohomologies of the ambiguity sheaf. For a closed subcomplex Y of X, let F^Y be the sheaf whose stalks are the stalks of F on Y and zero elsewhere, and whose restrictions are either those of F on Y or zero as appropriate. There is a surjective sheaf morphism $F \to F^Y$ and an induced ambiguity sheaf F_Y which can be constructed in exactly the same way as A before. Thus, the dimension of each stalk of F^Y is at least as large as that of any sampling sheaf, and the dimension of stalks of F_Y are therefore as small as or smaller than that of any ambiguity sheaf.

Proposition 9. *(Oversampling theorem) If X^k is the closed subcomplex generated by the k-faces of X, then $H^k(X^{k+1}; F_{X^k}) = 0$.*

Proof: By direct computation, the k-cochains of F_{X^k} are

$$\begin{aligned} C^k(X^{k+1}; F_{X^k}) &= C^k(X^{k+1}; F)/C^k(X^k; F) \\ &= \bigoplus_{a \text{ a } k\text{-face of} X} F(a) / \bigoplus_{a \text{ a } k\text{-face of} X} F(a) \\ &= 0. \end{aligned}$$

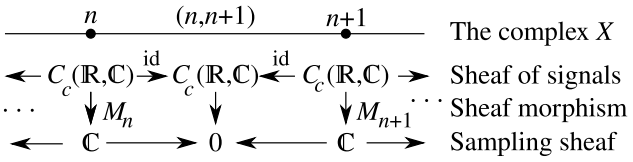


Fig. 1. The sheaves used in proving the traditional Nyquist theorem

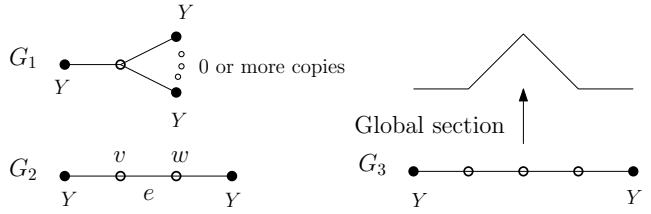


Fig. 2. Graphs G_1, and G_2 (left) and G_3 (right) for Lemma 12. Filled vertices represent elements of Y, empty ones are in the complement of Y.

■

As an immediate consequence, $H^0(X; F_Y) = 0$ when Y is the set of vertices of X.

Theorem 10. *(Sampling obstruction theorem) Suppose that Y is a closed subcomplex of X and $s : F \to S$ is a sampling of sheaves on X supported on Y. If $H^0(X, F_Y) \neq 0$, then the induced map $H^0(X; F) \to H^0(X; S)$ is not injective.*

Succinctly, $H^0(X, F_Y)$ is an obstruction to the recovery of global sections of F from its samples.

Proof: We begin by constructing the ambiguity sheaf A as before so that

$$0 \to A \to F \xrightarrow{s} S \to 0$$

is a short exact sequence. Observe that $S \to F^Y$ can be chosen to be injective, because the stalks of S have dimension not more than the dimension of F (and hence F^Y also). Thus the induced map $H^0(X; S) \to H^0(X; F^Y)$ is also injective. Therefore, by a diagram chase on

$$\begin{array}{ccccc} 0 \to H^0(X; A) & \longrightarrow & H^0(X; F) & \xrightarrow{s} & H^0(X; S) \\ & & \downarrow \cong & & \downarrow \\ 0 \to H^0(X; F_Y) & \longrightarrow & H^0(X; F) & \longrightarrow & H^0(X; F^Y) \end{array}$$

we infer that there is a surjection $H^0(X; A) \to H^0(X; F_Y)$. By hypothesis, this means that $H^0(X; A) \neq 0$, so in particular $H^0(X; F) \to H^0(X; S)$ cannot be injective. ■

IV. Applications

A. Bandlimited signals on the real line

In this section, we prove the traditional form of the Nyquist theorem by showing that bandlimiting is a sufficient condition for $H^0(X; A) = 0$. We begin by specifying the following 1-dimensional simplicial complex X. Let $X^0 = \mathbb{Z}$ and $X^1 = \{(n, n+1)\}$. We construct the sheaf C of signals (see Figure 1) so that for every simplex, the stalk of C is $C_c(\mathbb{R}, \mathbb{C})$, the set of compactly supported complex-valued continuous functions, and each restriction is the identity. Observe that the space of global sections of C is therefore just $C_c(\mathbb{R}, \mathbb{C})$.

Construct the sampling sheaf S whose stalk on each vertex is $\mathbb{C}$ and each edge stalk is zero. We construct a sampling morphism by the zero map on each edge, and by the inverse Fourier transform below on vertex $\{n\}$

$$M_n(f) = \int_{-\infty}^{\infty} f(\omega) e^{-2\pi i n \omega} d\omega.$$

Then the ambiguity sheaf A has stalks $C_c(\mathbb{R}, \mathbb{C})$ on each edge, and $\{f \in C_c(\mathbb{R}, \mathbb{C}) : M_n(f) = 0\}$ on each vertex $\{n\}$.

Theorem 11. *(Traditional Nyquist theorem) Suppose we replace $C_c(\mathbb{R}, \mathbb{C})$ with the set of continuous functions supported on $[-B, B]$. Then if $B \leq 1/2$, the resulting ambiguity sheaf A has $H^0(X; A) = 0$. Therefore, each such function can be recovered uniquely from its samples on $\mathbb{Z}$.*

Proof: The elements of $H^0(X; A)$ are given by the compactly supported continuous functions f on $[-B, B]$ for which

$$\int_{-B}^{B} f(\omega) e^{-2\pi i n \omega} d\omega = 0$$

for all n. Observe that if $B \leq 1/2$, this is precisely the statement that the Fourier series coefficients of f all vanish; hence f must vanish. This means that the only global section of A is the zero function. (Ambiguities can arise if $B > 1/2$, because the set of functions $\{e^{-2\pi i n \omega}\}_{n \in \mathbb{Z}}$ is then *not* complete.) ■

B. Beyond Nyquist: Piecewise linear functions on graphs

The sheaf-theoretic Nyquist theorem can treat nontrivial base space topologies as well as samples of different dimensions. Consider the example of the sheaf of piecewise linear functions PL on a graph, introduced in Section II-A and the sampling morphism $s : PL \to PL^Y$ where Y is a subset of the vertices of X. Excluding one or two vertices from Y does not prevent reconstruction in this case, because the samples include information about slopes along adjacent edges.

Lemma 12. *Consider PL_Y, the subsheaf of PL whose sections vanish on a vertex set Y and the graphs G_1, G_2, and G_3 as shown in Figure 2. There are no nontrivial sections of PL_Y on G_1 and G_2, but there are nontrivial sections of PL_Y on G_3.*

Proof: If a section of PL vanishes at a vertex x with degree n, this means that the value of the section there is an $(n+1)$-dimensional zero vector. The value of the section on every edge adjacent to x is then the 2-dimensional zero vector. Since the dimensions in each stalk of PL represent the value of the piecewise linear function and its slopes, linear extrapolation to the center vertex in G_1 implies that its value is zero too.

A similar idea applies in the case of G_2. The stalk at v has dimension 3. Any section at v that extends to the left must actually lie in the subspace spanned by $(0, 0, 1)$ (coordinates

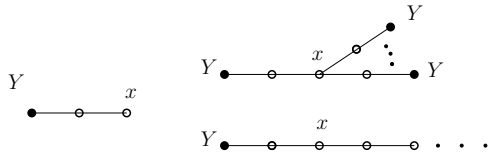


Fig. 3. The three families of subgraphs that arise when med $(Y) > 1$. Filled vertices represent elements of Y, empty ones are in the complement of Y.

represent the value, left slope, right slope respectively). In the same way, any section at w that extends to the right must lie in the subspace spanned by $(0, 1, 0)$. Any global section must extend to e, which must therefore have zero slope and zero value.

Finally G_3 has nontrivial global sections, spanned by the one shown in Figure 2. ∎

Definition 13. On a graph G, define the *edge distance* between two vertices v, w to be

$$\text{ed}\ (v,w) = \begin{cases} \min_p\{\#\text{ edges in } p \text{ such that } p \text{ is a} \\ \text{PL-continuous path from } v \to w\} \\ \infty \text{ if no such path exists} \end{cases}$$

From this, the maximal distance to a vertex set Y is

$$\text{med}\ (Y) = \max_{x \in X^0}\{\min_{y \in Y} \text{ed}\ (x,y)\}.$$

Proposition 14. *(Unambiguous sampling) Consider the sheaf PL on a graph X and $Y \subseteq X^0$. Then $H^0(X; F_Y) = 0$ if and only if med $(Y) \leq 1$.*

Proof: ($\Leftarrow$) Suppose that $x \in X^0 \backslash Y$ is a vertex not in Y. Then there exists a path with one edge connecting it to Y. Whence we are in the case of G_1 of Lemma 12, so any section at x must vanish.

($\Rightarrow$) By contradiction. Assume med $(Y) > 1$. Without loss of generality, consider $x \in X^0 \backslash Y$, whose distance to Y is exactly 2. Then one of the subgraphs shown in Figure 3 must be present in X. But case G_3 of Lemma 12 makes it clear that the most constrained of these (the middle panel of Figure 3) has nontrivial sections at x, merely looking at sections over the subgraph. ∎

Proposition 15. *(Non-redundant sampling) Consider the case of $s : PL \to PL^Y$. If $Y = X^0$, then $H^1(X; A) \neq 0$. If Y is such that med $(Y) \leq 1$ and $|X^0 \backslash Y| + \sum_{y \notin Y} \deg y = 2|X^1|$, then $H^1(X; A) = 0$.*

Proof: The stalk of A over each edge is $\mathbb{R}^2$, and the stalk over a vertex in Y is trival. However, the stalk over a vertex of degree n not in Y is $\mathbb{R}^{n+1}$. Observe that if $H^0(X; A) = 0$, then $H^1(X; A) = C^1(X; A)/C^0(X; A)$. Using the degree sum formula in graph theory, we compute that $H^1(X; A)$ has dimension $2|X^1| - \sum_{y \notin Y}(\deg y + 1)$. ∎

ACKNOWLEDGMENT

This work was partly supported under Federal Contract No. FA9550-09-1-0643.

REFERENCES

[1] J. Benedetto and W. Heller, "Irregular sampling and the theory of frames: I," *Note di Mathematica*, vol. 10, no. 1, pp. 103–125, 1990.
[2] H. Feichtinger and K. Gröchenig, "Theory and practice of irregular sampling," *Wavelets: mathematics and applications*, pp. 305–363, 1994.
[3] M. Unser, "Sampling–50 years after Shannon," *Proceedings of the IEEE*, vol. 88, no. 4, pp. 569–587, 2000.
[4] S. Smale and D. Zhou, "Shannon sampling and function reconstruction from point values," *Bulletin of the American Mathematical Society*, vol. 41, no. 3, pp. 279–306, 2004.
[5] P. Niyogi, S. Smale, and S. Weinberger, "Finding the homology of submanifolds with high confidence from random samples," in *Twentieth Anniversary Volume*, R. Pollack, J. Pach, and J. E. Goodman, Eds. Springer New York, 2009, pp. 1–23.
[6] F. Chazal, D. Cohen-Steiner, and A. Lieutier, "A sampling theory for compact sets in Euclidean space," *Discrete Comput. Geom.*, vol. 41, pp. 461–479, 2009.
[7] R. Vaughan, N. Scott, and D. White, "The theory of bandpass sampling," *Signal Processing, IEEE Transactions on*, vol. 39, no. 9, pp. 1973–1984, 1991.
[8] M. Unser and J. Zerubia, "A generalized sampling theory without band-limiting constraints," *Circuits and Systems II: Analog and Digital Signal Processing, IEEE Transactions on*, vol. 45, no. 8, pp. 959–969, 1998.
[9] M. Unser, "Splines: A perfect fit for signal and image processing," *Signal Processing Magazine, IEEE*, vol. 16, no. 6, pp. 22–38, 1999.
[10] A. Shepard, "A cellular description of the derived category of a stratified space," Ph.D. dissertation, Brown University, 1985.
[11] R. Ghrist and Y. Hiraoka, "Applications of sheaf cohomology and exact sequences to network coding," *preprint*, 2011.
[12] J. Lilius, "Sheaf semantics for Petri nets," Helsinki University of Technology, Digital Systems Laboratory, Tech. Rep., 1993.
[13] J. Curry, R. Ghrist, and M. Robinson, "Euler calculus and its applications to signals and sensing," in *Proceedings of Symposia in Applied Mathematics: Advances in Applied and Computational Topology*, A. Zomorodian, Ed., 2012.
[14] M. Robinson, "Inverse problems in geometric graphs using internal measurements, `arxiv:1008.2933`," 2010.
[15] ——, "Asynchronous logic circuits and sheaf obstructions," *Electronic Notes in Theoretical Computer Science*, pp. 159–177, 2012.
[16] I. Pesenson, "Sampling of band-limited vectors," *Journal of Fourier Analysis and Applications*, vol. 7, no. 1, pp. 93–100, 2001.
[17] A. Schuster and D. Varolin, "Interpolation and sampling for generalized Bergman spaces on finite Riemann surfaces," *Revista Matemática Iberoamericana*, vol. 24, no. 2, pp. 499–530, 2008.
[18] M. Ebata, M. Eguchi, S. Koizumi, and K. Kumahara, "Analogues of sampling theorems for some homogeneous spaces," *Hiroshima Math. J.*, vol. 36, pp. 125–140, 2006.
[19] H. Feichtinger and I. Pesenson, "A reconstruction method for band-limited signals on the hyperbolic plane," *Sampl. Theory Signal Image Process.*, vol. 4, no. 2, pp. 107–119, 2005.
[20] I. Pesenson and M. Pesenson, "Sampling, filtering and sparse approximations on combinatorial graphs," *Journal of Fourier Analysis and Applications*, vol. 16, no. 6, pp. 921–942, 2010.
[21] I. Pesenson, "Band limited functions on quantum graphs," *Proceedings of the American Mathematical Society*, vol. 133, no. 12, pp. 3647–3656, 2005.
[22] ——, "Analysis of band-limited functions on quantum graphs," *Applied and Computational Harmonic Analysis*, vol. 21, no. 2, pp. 230–244, 2006.
[23] K. Gröchening, "Reconstruction algorithms in irregular sampling," *Mathematics of Computation*, vol. 59, no. 199, pp. 181–194, July 1992.
[24] M. Vetterli, P. Marziliano, and T. Blu, "Sampling signals with finite rate of innovation," *Signal Processing, IEEE Transactions on*, vol. 50, no. 6, pp. 1417–1428, 2002.
[25] P. Dragotti, M. Vetterli, and T. Blue, "Sampling moments and reconstructing signals of finite rate of innovation: Shannon meets Strang-Fix," *IEEE Trans. Sig. Proc.*, vol. 55, no. 5, May 2007.
[26] G. Bredon, *Sheaf theory*. Springer, 1997.
[27] J. H. Hubbard, *Teichmüller Theory, volume 1*. Matrix Editions, 2006.

Frames of eigenspaces and localization of signal components

Monika Dörfler[(1)] and José Luis Romero[(2)]
Faculty of Mathematics, University of Vienna,
Nordbergstrasse 15,A-1090 Wien, Austria
Email: (1) Monika.Doerfler@univie.ac.at, (2) Jose.Luis.Romero@univie.ac.at.

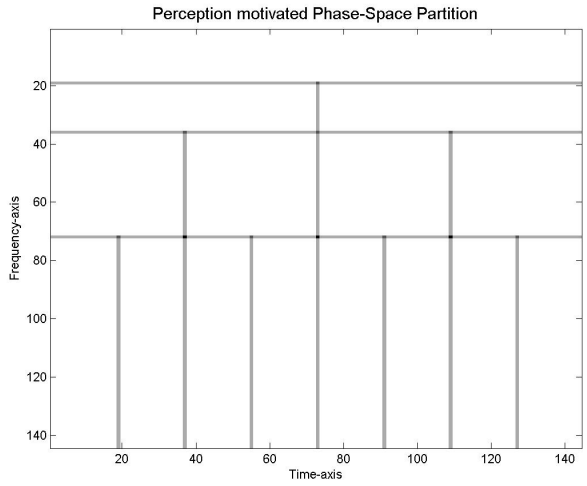


Fig. 1. Time-frequency partition with varying time-frequency bands

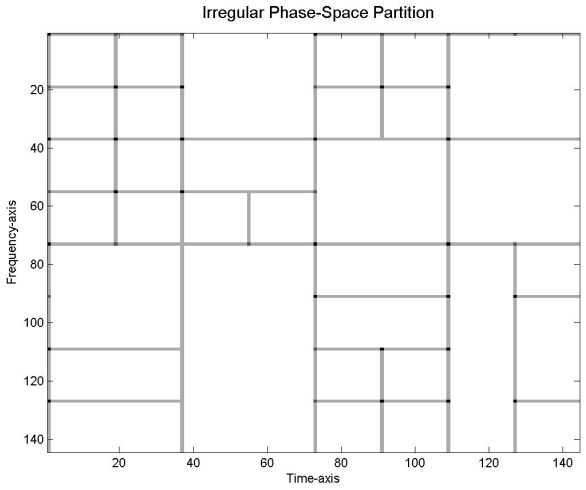


Fig. 2. Fully irregular time-frequency partition

***Abstract*—We present a construction of frames adapted to a given time-frequency cover and study certain computational aspects of it. These frames are based on a family of orthogonal projections that can be used to localize signals in the time-frequency plane. We compare the effect of the corresponding orthogonal projections to the traditional time-frequency masking.**

I. Introduction

When representing a signal in a time-frequency dictionary, the atoms are usually chosen as time-frequency shifts of a window along a lattice (Gabor frame). The choice of the lattice together with the characteristics (shape, width) of the basic window or family of windows determines the ability of the representation to localize certain signal components and, furthermore, the possibility to separate them. Various approaches have been taken to circumvent the restrictions possibly imposed by a rigid application of lattice structure (reassignment, adaptive frames) [1], [2], [6], [3], [17], giving time-frequency partitions consisting of frequency (resp. time) strips of varying widths (see figure 1)

In [10] we have presented a construction of frames whose spectrogram follows a prescribed time-frequency pattern. This pattern may be quite irregular and in particular does not need to be a Cartesian product of a time and a frequency partition (see figure 2).

This construction is achieved by selecting from each tile of the cover an orthonormal set of functions that maximizes its joint spectrogram within the tile. These functions are eigenfunctions of time-frequency localization operators (see below), whose concentration is no more restricted to be located at lattice points. By definition, the eigenfunctions corresponding to high eigenvalues of the localization operators, are maximally localized within a (weighted) subfamily of the time-frequency shifted atoms; thus, they provide potentially better localization in a certain time-frequency region than the time-frequency atoms themselves.

Since the frames introduced in [10] are constructed by choosing a finite number of eigenfunctions from each localization operator corresponding to a partition of the time-frequency plane, they produce a resolution of the identity by orthogonal projections. This means that replacing the usual time-frequency masking operators by certain orthogonal projections does not lead to loss of information, provided that the projection is chosen judiciously.

In this article we consider certain computational aspects of the construction of frames adapted to time-frequency covers and compare the effect of the corresponding orthogonal projections to the traditional time-frequency masking.

II. Time-frequency localization

A. Localization operators

The short-time Fourier transform (STFT) of a distribution $f \in \mathcal{S}'(\mathbb{R}^d)$ is a function defined on $\mathbb{R}^d \times \mathbb{R}^d$ defined, by means of an adequate smooth and fast-decaying window function $\varphi \in \mathcal{S}(\mathbb{R}^d)$, as

$$\mathcal{V}_\varphi f(z) = \int_{\mathbb{R}^d} f(t)\overline{\varphi(t-x)}e^{-2\pi i \xi t} dt, \quad z = (x,\xi) \in \mathbb{R}^d \times \mathbb{R}^d.$$

The number $\mathcal{V}_\varphi f(x,\xi)$ represents the influence of the frequency ξ near x. The distribution f can be re-synthesized from its time-frequency content by,

$$f(t) = \int_{\mathbb{R}^d \times \mathbb{R}^d} \mathcal{V}_\varphi f(x,\xi)\varphi(t-x)e^{2\pi i \xi t} dx d\xi. \qquad (1)$$

Given a compact set $\Omega \subseteq \mathbb{R}^{2d}$ in the time-frequency plane, the *time-frequency localization operator* L_Ω is defined by masking the coefficients in (1), cf. [5], i.e.

$$L_\Omega f(t) = \int_\Omega \mathcal{V}_\varphi f(x,\xi)\varphi(t-x)e^{2\pi i \xi t} dx d\xi. \qquad (2)$$

L_Ω is self-adjoint and trace-class, so we can consider its spectral decomposition

$$L_\Omega f = \sum_{k=1}^{\infty} \lambda_k^\Omega \langle f, \phi_k^\Omega \rangle \phi_k^\Omega.$$

The first eigenfunction, ϕ_1^Ω, is optimally concentrated inside Ω in the following sense,

$$\int_\Omega \left|\mathcal{V}_\varphi \phi_1^\Omega(z)\right|^2 dz = \max_{\|f\|_2=1} \int_\Omega |\mathcal{V}_\varphi f(z)|^2 dz.$$

More generally, the first N eigenfunctions of H_Ω form an orthonormal set in $L^2(\mathbb{R}^d)$ that maximizes the quantity

$$\sum_{j=1}^{N} \int_\Omega |\mathcal{V}_\varphi \phi_j^\Omega(z)|^2 dz,$$

among all orthonormal sets of N functions in $L^2(\mathbb{R}^d)$. In this sense, their time-frequency profile is optimally adapted to Ω.

B. Time-Frequency areas of interest

The shape of the time-frequency areas one may be interested to localize in, will usually depend on the application and the characteristics of the underlying class of signals. Typically, one may consider rectangles of different eccentricities in order to be able to focus on signal components showing a more transient or more harmonic characteristic. Examples are depicted in Figure 3. In some applications, one may be interested in more exotic shapes, such as triangular, cf. Figure 4, for example to account for the spectral roll-off in instrumental sounds, or chirped components, cf. Figure 5, which are also omnipresent in both speech and music signals.

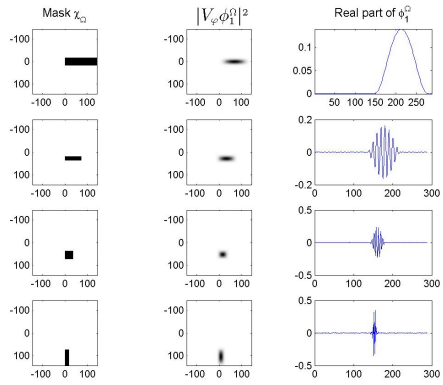


Fig. 3. Four different rectangular masks in time-frequency domain and the first eigenfunctions of the corresponding localization operators. Middle plots show the absolute value squared of the STFT and right plots show the real part.

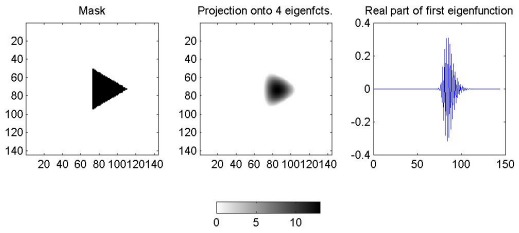


Fig. 4. Triangular-shaped mask, absolute value squared of the STFT of projection of random noise onto most localized resulting eigenfunctions, real part of most concentrated eigenfunction.

III. Frames of eigenfunctions

We now present the main result on the construction of frames adapted to a cover and then explore certain computational aspects of it. The proof of the following theorem can be found in [10], together with an extended discussion on its quantitative aspects (see also [8], [15], [9], [16]).

Theorem 1: Let $\{\Omega_\gamma : \gamma \in \Gamma\}$ be a cover of $\mathbb{R}^{2d}$ such that

$$B_r(\gamma) \subseteq \Omega_\gamma \subseteq B_R(\gamma), \quad \text{with } \Gamma \text{ a lattice and } R \geq r > 0.$$

Then, there exists a constant $C > 0$ such that for every choice of N_γ, $C\,|\Omega_\gamma| \leq N_\gamma \leq N < \infty$, the family of functions

$$\left\{\phi_k^{\Omega_\gamma} : \gamma \in \Gamma, 1 \leq k \leq N_\gamma\right\}$$

is a frame of $L^2(\mathbb{R}^d)$.

A. Computing the eigenfunctions in each tile

In practice we work with a discrete realization of L_Ω given by

$$H_{\mathbf{m},\mathbf{\Lambda}} f = \sum_{\lambda \in \Lambda} m(\lambda)\langle f, \pi(\lambda)g\rangle \pi(\lambda)g, \qquad (3)$$

where

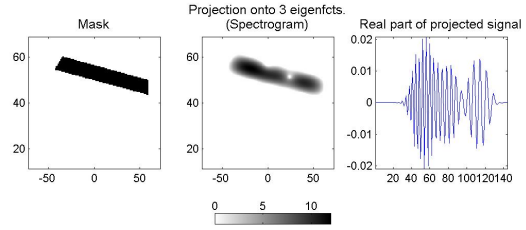


Fig. 5. Chirp-shaped mask, absolute value squared of the STFT of projection of random noise onto most localized resulting eigenfunctions, real part of the projection.

- $\Lambda \subseteq \mathbb{R}^{2d}$ is a lattice,
- $\{\pi(\lambda)g = e^{2\pi i \lambda_2 \cdot} g(\cdot - \lambda_1) : \lambda = (\lambda_1, \lambda_2) \in \Lambda\}$ is a tight Gabor frame of $L^2(\mathbb{R}^d)$.
- $\mathbf{m} = (\mathbf{m}_\lambda)_{\lambda \in \mathbf{\Lambda}}$ is a bounded sequence of complex numbers.

The operator $H_{\mathbf{m},\mathbf{\Lambda}}$ is called a Gabor multiplier with mask $\mathbf{m}$. If we let $m(\lambda) := 1$, if $\lambda \in \Omega$ and 0 otherwise, then $H_{\mathbf{m},\mathbf{\Lambda}}$ is a discretization of the operator L_Ω in (2).

Given an operator $H_{\mathbf{m},\mathbf{\Lambda}}$ defined in (3), mapping $L^2(\mathbb{R}^d)$ into itself, we denote $K = \#\,\mathrm{supp}(\mathbf{m})$, assume that K is finite, and write $H_{\mathbf{m}}$ as a composition of the operator $G_{\sqrt{\mathbf{m}}} : f \mapsto [\sqrt{m(\lambda)}\langle f, \pi(\lambda)g\rangle]_{\lambda \in \Lambda \cap \mathrm{supp}(\mathbf{m})}$, mapping $L^2(\mathbb{R}^d)$ into $\mathbb{C}^K$ and its adjoint $G^*_{\sqrt{\mathbf{m}}}$.
Both $G_{\sqrt{\mathbf{m}}}$ and $G^*_{\sqrt{\mathbf{m}}}$ are finite-rank operators and can be written in their singular value decomposition:

$$G_{\sqrt{\mathbf{m}}} = \sum_{j=1}^{K} s_j \langle \cdot, v_j \rangle_{L^2} u_j, \tag{4}$$

$$G^*_{\sqrt{\mathbf{m}}} = \sum_{j=1}^{K} s_j \langle \cdot, u_j \rangle_{\mathbb{C}^K} v_j. \tag{5}$$

Then, applying $G^*_{\sqrt{\mathbf{m}}}$ to u_k yields $G^*_{\sqrt{\mathbf{m}}} \cdot u_k = s_k \cdot v_k$ and thus the eigenfunctions v_j of $H_{\mathbf{m},\mathbf{\Lambda}}$ may be obtained from the eigenfunctions of the Gramian operator $\Gamma_m := G_{\sqrt{\mathbf{m}}} \cdot G^*_{\sqrt{\mathbf{m}}}$ by

$$v_j = \frac{1}{s_j} \cdot G^*_{\sqrt{\mathbf{m}}} \cdot u_j,\ j = 1, \ldots, K. \tag{6}$$

In typical applications, where $H_{\mathbf{m},\mathbf{\Lambda}}$ is a matrix whose size depends on the signal length, the size of the corresponding Gramian matrix is $K \times K$ with K being the size of the support of the mask (or, in the case of $0/1$-masks, the support of Ω) which is usually small enough for the computation of the spectral decomposition to be a feasible task. Furthermore, in (6) only the eigenfunctions corresponding to relevant eigenvalues s_j^2 need to be computed.

B. Computing the whole frame

Section III-A deals with the computation of the relevant eigenfunctions for each individual tile of the cover. To compute the whole frame we use the following observation based on the so-called covariance of the Short-Time Fourier transform.

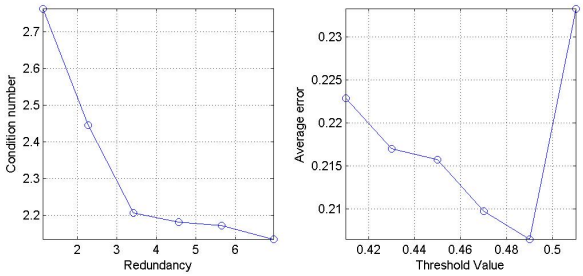


Fig. 6. *Evaluation of the procedure to obtain frames adapted to a given time-frequency partition.*

Lemma 1: If $\Omega' = \Omega + z_0$, for some $z_0 \in \mathbb{R}^{2d}$. Then the eigenfunctions of L_Ω and $L_{\Omega'}$ are related by

$$\phi_k^{\Omega'} = \pi(z_0)\phi_k^{\Omega}, \quad k \geq 1.$$

where $\pi(x,w)f(t) = e^{2\pi i w t} f(t-x)$. Hence, if the cover $\{\Omega_\gamma : \gamma \in \Gamma\}$ in Theorem 1 consists of translates of N basic tiles $\Omega^1, \ldots, \Omega^N$,

$$\Omega_\gamma = \Omega^{k_\gamma} + z_\gamma, \quad 1 \leq k_\gamma \leq N, z_\gamma \in \mathbb{R}^{2d},$$

then only N sets of eigenfunctions need to be computed.

C. The number of eigenfunctions and the resulting frame quality

In order to test the performance of the procedure described in Theorem 1 and Lemma 1 for the generation of a new frame, we generated random partitions of the time-frequency plane, consisting of three different rectangular shapes, thus in the spirit of the example shown in Figure 2. Then, the eigenvalues of the resulting spectral decomposition were thresholded by 6 different values between 0.51 and 0.41 and the corresponding eigenfunctions were used to generate time-frequency frames with redundancies between 1.15 and 7. The condition numbers of the resulting frames are shown in Figure 6, as well as the average error, when the corresponding frame operators are applied to (1000 realizations of) random noise. Interestingly, while the condition number of the resulting frame improves for increased redundancy, the optimal approximation of the identity seems to be obtained for a threshold very close to 0.5. This agrees with the observation that the number of eigenvalues above 0.5 is given by the volume of the localization area [14], [11], [7], [12]. This effect can be circumvented by renormalizing the eigenfunctions to its corresponding eigenvalue (see [10]).

IV. Frames of eigenspaces

Theorem 1 can be interpreted in the following way. For each $\gamma \in \Gamma$ let V_γ be the subspace spanned by the first N_γ eigenfunction $\{\phi_1^\Omega, \dots, \phi_{N_\gamma}^\Omega\}$ and let P_γ be the corresponding orthogonal projection. Then,

$$\|f\|_2^2 \approx \sum_{\gamma\in\Gamma} \|P_\gamma f\|_2^2, \quad f \in L^2(\mathbb{R}^d).$$

This means that $\{V_\gamma : \gamma \in \Gamma\}$ is a *fusion frame* in the sense of [4]. In certain situations, using the projection P_γ may be preferable to masking the coefficients with a multiplier like the one in (3).

A. Cutting with reduced spilling

Denoising by time-frequency masking is a ubiquitous method in signal restoration, cf. [18]. However, in dependence on the time-frequency concentration of the window used to obtain the time-frequency representation used, this method leads to significant spilling of energy outside the region of relevant signal components. Applying projection onto significant eigenfunctions of a time-frequency multiplier instead of applying the multiplier itself, can ameliorate this bias. An example for this is shown in Figure 7. Here, a Hann window h was chosen as a reference signal, while the analysis window is still a Gaussian window. The signal h was disturbed by additive white noise, with a signal to noise ratio (SNR) of $3.5dB$ to obtain the noisy signal h_n. Then, the original signal was recovered by either applying a Gabor multiplier derived from a $0/1$-mask on the estimated region, with an underlying Gabor frame of redundancy 16, and, on the other hand, the projection onto the eigenfunctions corresponding to eigenvalues close to 1. The average achieved SNR (over a 1000 noise-realizations) was $12.5dB$ for the projection approach and $11.3dB$ for the plain Gabor multiplier.

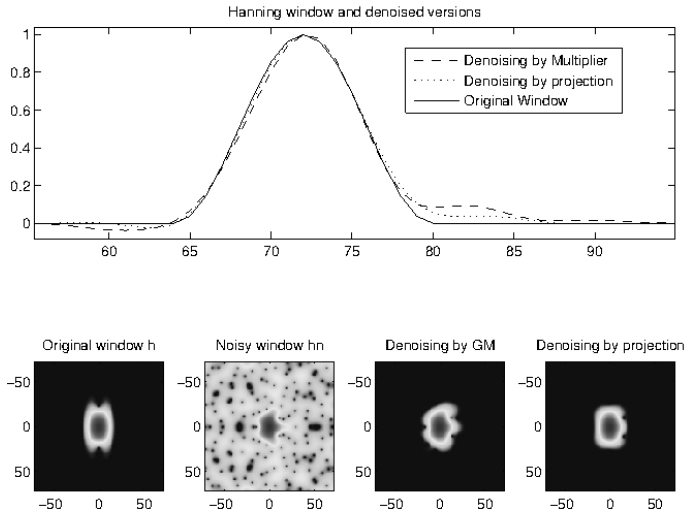


Fig. 7. *Denoising a disturbed signal by either time-frequency masking (GM) or projection onto relevant eigenfunctions. The upper plot shows the time-domain signals. The lower plots show the db-values of the spectrograms of the original window h, its noisy version h_n and the two denoised signals.*

V. Conclusion and Perspectives

In this article we have presented a new method to obtain frames adapted to a given partition of the time-frequency plane and addressed certain computational aspects of it. We also showed that using projections onto the space spanned by the first eigenfunctions corresponding to the Gabor multiplier of a certain localization region can yield better results than applying the Gabor multiplier itself. These are preliminary results that must be evaluated more extensively and in particular given a proof of concept by means of application to real-life data.

References

[1] A. Aldroubi, C. Cabrelli, and U. Molter. Wavelets on irregular grids with arbitrary dilation matrices, and frame atoms for $L^2(R^d)$. *Appl. Comput. Harmon. Anal.*, Special Issue on Frames II.:119–140, 2004.

[2] F. Auger and P. Flandrin. Improving the readability of time-frequency and time-scale representations by the reassignment method *IEEE Trans. Sig. Proc.*, 43(5):1068 –1089, May 1995.

[3] D. Jones and R. G. Baraniuk. A simple scheme for adapting time-frequency representations. *IEEE Trans. Signal Process.*, 42(12):3530–3535, Dec 1994.

[4] P. G. Casazza and G. Kutyniok. Frames of subspaces. In *Wavelets, Frames and Operator Theory*, volume 345 of *Contemp. Math.*, pages 87–113. Amer. Math. Soc., Providence, RI, 2004.

[5] I. Daubechies. Time-frequency localization operators: a geometric phase space approach. *IEEE Trans. Inform. Theory*, 34(4):605–612, July 1988.

[6] I. Daubechies, J. Lu, and H.-T. Wu. Synchrosqueezed wavelet transforms: an empirical mode decomposition-like tool. *Appl. Comput. Harmon. Anal.*, 30(2):243–261, 2011.

[7] F. DeMari, H. G. Feichtinger, and K. Nowak. Uniform eigenvalue estimates for time-frequency localization operators. *J. London Math. Soc.*, 65(3):720–732, 2002.

[8] M. Dörfler, H. G. Feichtinger, and K. Gröchenig. Time-frequency partitions for the Gelfand triple (S_0, L^2, S_0'). *Math. Scand.*, 98(1):81–96, 2006.

[9] M. Dörfler and K. Gröchenig. Time-frequency partitions and characterizations of modulations spaces with localization operators. *J. Funct. Anal.*, 260(7):1903 – 1924, 2011.

[10] M. Dörfler and J. L. Romero. Frames adapted to a phase-space cover. *Preprint.*, arXiv:1207.5383.

[11] H. G. Feichtinger and K. Nowak. A Szegö-type theorem for Gabor-Toeplitz localization operators. *Michigan Math. J.*, 49(1):13–21, 2001.

[12] H. G. Feichtinger and K. Nowak. A first survey of Gabor multipliers. In H. G. Feichtinger and T. Strohmer, editors, *Advances in Gabor Analysis*, Appl. Numer. Harmon. Anal., pages 99–128. Birkhäuser, 2003.

[13] F. Jaillet and B. Torrésani. Time–frequency jigsaw puzzle: adaptive multiwindow and multilayered Gabor expansions. *Int. J. Wavelets Multiresolut. Inf. Process.*, 2:293–316, 2007.

[14] J. Ramanathan and P. Topiwala. Time-frequency localization and the spectrogram. *Appl. Comput. Harmon. Anal.*, 1(2):209–215, 1994.

[15] J. L. Romero. Surgery of spline-type and molecular frames. *J. Fourier Anal. Appl.*, 17:135 – 174, 2011.

[16] J. L. Romero. Characterization of coorbit spaces with phase-space covers. *J. Funct. Anal.*, 262(1):59–93, 2012.

[17] P. J. Wolfe, D. Rudoy, and B. Prabahan. Superposition frames for adaptive time-frequency analysis and fast reconstruction. *IEEE Trans. Signal Process.*, 58:2581–2596, 2010.

[18] P. Wolfe and S. Godsill. Efficient alternatives to the Ephraim and Malah suppression rule for audio signal enhancement. *EURASIP J. Appl. Signal Process.*, 2003(10):1043–1051, 2003.

Monika Dörfler was supported by the WWTF project Audio-Miner (MA09-024). José Luis Romero gratefully acknowledges support from the Austrian Science Fund (FWF):[P22746-N13].

A Lie group approach to diffusive wavelets

Swanhild Bernstein
Institute of Applied Analysis
TU Bergakademie Freiberg
D-09596 Freiberg, Germany
Email: swanhild.bernstein@math.tu-freiberg.de

***Abstract*—The aim of this paper is to give an overview of diffusive wavelets on compact Lie groups, homogenous spaces and the Heisenberg group. This approach is based on Lie groups and representation theory and generalizes well-known constructions of wavelets on the sphere. We give also examples for the construction of diffusive wavelets.**

I. Introduction

The task of analyzing data, reconstructing functions from measurements or to save data in an handable way occurs in a lot of applications. Problems in geophysics, astronomy and in material sciences involve groups and homogeneous spaces such as the group SO(3) or the spheres S^2 or S^3. The group theoretic approach to wavelets as coherent states fails and it has been an open problem for along time to construct wavelets on the sphere. The breakthrough was the construction of spherical wavelets on S^2 based on convolution-type integrals by W. Freeden and co-workers [14]. An alternate successful approach was made by J.-P. Antoine and P. Vandergheynst [1], [2] by lifting up rotations and dilations on the sphere into the Lorentz group. The aim of this paper is to demonstrate that both approaches can be generalized to continuous diffusive wavelets. Diffusive wavelets can be build not only on compact groups and homogeneous spaces but also on stratified groups. The most well-known example here is the Heisenberg group.

Classical wavelet theory is based on the group generated by translations and dilations. The key idea of diffusive wavelets is to generate a dilation from a diffusive semigroup and to substitute translation by action of a compact group. A related approach based on spectral calculus of the Laplacian on closed manifolds was proposed by D. Geller and A. Mayeli [15] and related work by I. Pesenson and D. Geller [17], [18].

The construction of diffusive wavelets is based on convolution-type operators. These ideas were used in [6] to construct wavelets on the sphere S^3. Discrete wavelet transforms of that type were used by R. R. Coifman, M. Maggioni and others [11], [10], where the heat evolution was combined with an orthogonalization procedure to model a multi-resolution analysis in $L^2(S^3)$.

These ideas can be combined with other group structures to get wavelets invariant under finite reflection groups [3]. A similar construction is possible for the torus in [5].

An application material sciences and specifically to the crystallographic Radon-transform [8], [7], [9] we need wavelets on S^3, $SO(3)$ and $S^2 \times S^2$ [4].

This approach was generalized by a representation theory based approach where the heat flow was replaced by a more general approximate convolution identity in [13] and [12].

The aim of this paper is to give a general approach to diffusive wavelets on compact groups, homogeneous spaces and stratified groups which is based on the theory of Lie groups. Several examples explain the construction of diffusive wavelets for specific situations.

II. Diffusive wavelets

A. Preliminaries on compact Lie groups

Let $\mathscr{G}$ be a compact Lie group. A unitary representation of $\mathscr{G}$ is a continuous group homomorphism $\pi\colon \mathscr{G} \to U(d_\pi)$ of $\mathscr{G}$ into the group of unitary matrices of a certain dimension d_π. Such a representation is irreducible if $\pi(g)M = M\pi(g)$ for all $g \in \mathscr{G}$ and some $M \in \mathbb{C}^{d_\pi \times d_\pi}$ implies $M = cId$, where Id is the identity matrix.

Theorem 1 (Peter-Weyl)**.** *Let $\widehat{\mathscr{G}}$ be the set of all equivalence classes of irreducible representations of the compact Lie group $\mathscr{G}$, choose one unitary representation $\pi_\alpha(g)$ from each class, and let the dimension of the representation $\pi_\alpha(g)$ be d_α, and its matrix elements be π^α_{ij}, $1 \le i,j \le d_\alpha$, and $\mathcal{H}_\alpha = \operatorname{span}(\pi^\alpha_{ij})_{i,j}^{d_\alpha}$. Then*

$$L^2(\mathscr{G}) = \oplus_\alpha \mathcal{H}_{\pi_\alpha} = \oplus_{\pi\in\widehat{\mathscr{G}}} \mathcal{H}_\pi$$

and any function $f \in L^2(\mathscr{G})$ has a unique decomposition into

$$f(g) = \sum_\alpha \sum_{i,j}^{d_\alpha} c^\alpha_{ij} \pi^\alpha_{ij},$$

with Fourier coefficents c^α_{ij}.
The orthogonal projection $L^2(\mathscr{G}) \to \mathcal{H}_\alpha$ is given by

$$f_\alpha(g) = \sum_{i,j}^{d_\alpha} c^\alpha_{ij}\pi^\alpha_{ij} = d_\alpha(\phi * \chi_{pi_\alpha}),$$

where $\chi_{\pi_\alpha}(g) = \operatorname{trace}\pi_\alpha(g)$ is the character of the representation.

The Fourier coefficient $\hat{f}(\pi)$ can be calculated as

$$\hat{f}(\pi) = \int_{\mathscr{G}} f(g)\pi^*(g)\,dg.$$

and the inversion formula (the Fourier expansion) is then given by

$$f(g) = \sum_{\pi \in \hat{\mathscr{G}}} d_\pi \operatorname{trace}(\pi(g)\hat{f}(\pi)).$$

The Laplace-Beltrami operator $\Delta_{\mathscr{G}}$ on $\mathscr{G}$ is bi-invariant. Therefore, all of its eigenspaces are also bi-invariant subspaces of $L^2(\mathscr{G})$. As $\mathcal{H}_\pi$ are minimal bi-invariant subspaces, each of them has to be an eigenspace of Δ_G with respect to an eigenvalue $-\lambda_\pi^2$. Hence,

$$\Delta_{\mathscr{G}}\phi = -\sum_{\pi \in \hat{\mathscr{G}}} d_\pi \lambda_\pi^2 \operatorname{trace}(\pi(g)\hat{\phi}(\pi))$$

and the solution to the heat equation

$$(\partial_t - \Delta_{\mathscr{G}})u = 0,\ u(0,\cdot) = \phi,$$

is given as convolution with the heat kernel $p_t(g)$ as $u(t,\cdot) = \phi * p_t$, where

$$\hat{p}_t(\pi) = e^{-t\lambda_\pi^2} I \text{ and } p_t(g) = \sum_{\hat{\mathscr{G}}} d_\pi e^{-t\lambda_\pi^2}\chi_\pi(g).$$

In particular $\phi * p_t \to \phi$ for all $\phi \in L^p(\mathscr{G})$, $1 \le p < \infty$.

B. Wavelets on compact groups

Definition 1 (Diffusive approximate identity)**.** *Let $\hat{\mathscr{G}}_+ \subset \hat{\mathscr{G}}$ be cofinite. A family $t \to p_t$ from $C^1(\mathbb{R}_+; L^1(\mathscr{G}))$ will be called diffusive approximate identity with respect to $\hat{\mathscr{G}}_+$ if it satisfies*

- $||\hat{p}_t(\pi)|| \le C$ *uniform in* $\pi \in \hat{\mathscr{G}}_+$ *and* $t \in \mathbb{R}_+$;
- $\lim_{t\to 0} \hat{p}_t(\pi) = I$ *for all* $\pi \in \hat{\mathscr{G}}_+$;
- $\lim_{t\to\infty} \hat{p}_t(\pi) = 0$ *for all* $\pi \in \hat{\mathscr{G}}_+$;
- $-\partial_t \hat{p}_t(\pi)$ *is a positive matrix for all* $t \in \mathbb{R}_+$ *and* $\lim_{t\to 0} \hat{p}_t(\pi) = I$ *for all* $\pi \in \hat{\mathscr{G}}_+$.

For $f \in L^2(\mathscr{G})$ the projection onto $L^2_0(\mathscr{G})$ is

$$f\,|_{\widehat{\mathscr{G}}_+} = \sum_{\pi \in \widehat{\mathscr{G}}_+} f * \chi_\pi.$$

Definition 2 (Diffusive wavelets on a compact Lie group)**.** *Let p_t be a diffusive approximate identity and $\alpha(\rho) > 0$ a given weight function.*
A family $\psi_\rho \in L^2_0(\mathscr{G}) = \bigoplus_{\pi\in\hat{\mathscr{G}}_+} H_\pi$ is called diffusive wavelet family*, if it satisfies the admissibility condition*

$$p_t|_{\hat{\mathscr{G}}_+} = \int_t^\infty \check{\psi}_\rho * \psi_\rho\, \alpha(\rho)\, d\rho,$$

where $\check{\psi}_\rho(g) = \overline{\psi_\rho(g^{-1})}$.

Applying Fourier transform to the admissibility condition yields:

$$\hat{p}_t(\pi) = \int_t^\infty \hat{\psi}_\rho(\pi)\hat{\psi}_\rho^*(\pi)\alpha(\rho).$$

Differentiation with respect to t results in

$$-\partial_t \hat{p}_t(\pi) = \hat{\psi}_\rho(\pi)\hat{\psi}_\rho^*(\pi)\alpha(\rho).$$

If $\widehat{\psi}_\pi(\pi)$ are the Fourier coefficients than a multiplication with a unitary matrix $\eta_\rho(\pi)$ does not change the last equality.

C. Wavelets based on the heat kernel

Let p_t be the heat kernel e_t^{heat} on the group $\mathscr{G}$. We know that

$$\lim_{t\to\infty} \hat{e}_t^{heat}(\pi) = 0$$

for all nontrivial representations of $\mathscr{G}$. Since the character of the trivial representation π_0 is $\chi_{\pi_0} \equiv 1$ the corresponding invariant subspace in $L^2(\mathscr{G})$ is the space of constant functions and hence the eigenvalue vanishes, which implies $\hat{e}_t^{heat}(\pi) = Id$ and contradicts the definition of the diffusive approximate identity. Therefore we choose

$$\widehat{\mathscr{G}_+} = \widehat{\mathscr{G}} \backslash \{\pi_0\}.$$

That means $L^2_0(\mathscr{G})$ contains all square integrable functions with vanishing mean. The admissibility condition reads now as

$$\partial_\rho \hat{e}_\rho^{heat}(\pi) = \lambda_\pi^2 e^{-\rho\lambda_\pi^2} Id = \hat{\psi}_\rho(\pi)\hat{\psi}_\rho^*(\pi)\alpha(\rho).$$

Due to the freedom in choosing a unitary matrix $\eta_\rho(\pi)$ we get

$$\hat{\psi}_\rho(\pi) = \tfrac{1}{\sqrt{\alpha(\rho)}} \lambda_\pi e^{-\lambda_\pi \frac{\rho}{2}}\, Id$$

and the wavelet has the form

$$\psi_\rho = \tfrac{1}{\sqrt{\alpha(\rho)}} \sum_{\pi\in\widehat{G}_+} d_\pi \lambda_\pi e^{-\lambda_\pi \frac{\rho}{2}} \operatorname{trace}(\eta_\rho(\pi)\pi(g)).$$

Definition 3 (Wavelet transform)**.** *Let $\mathscr{G}$ be a compact group, $\alpha(\rho) > 0$ a weight function on $\mathscr{G}$ and $\psi_\rho \in L^2_0(\mathscr{G})$ a diffusive wavelet family. The wavelet transform $\mathscr{W} : L^2_0(\mathscr{G}) \to L^2(\mathbb{R}_+ \times \mathscr{G},\, \alpha(\rho)\, d\rho \otimes dg)$ is defined as*

$$(\mathscr{W} f)(\rho,\, g) := (f * \check{\psi}_\rho)(g)$$

Theorem 2. *The wavelet transform $\mathscr{W} : L^2_0(\mathscr{G}) \to L^2(\mathbb{R}_+ \times \mathscr{G}, \alpha(\rho)\, d\rho \otimes dg)$ is a unitary operator and the wavelet transform is invertible on its range by*

$$\int_{\mathbb{R}_+}\int_{\mathscr{G}} (\mathscr{W} f)(\rho,\, h)\, \psi_\rho(h^{-1}g)\, dh\, \alpha(\rho)\, d\rho$$
$$= \int_{\to 0}^{\infty} (\mathscr{W} f)(\rho, \cdot) * \psi_\rho\, \alpha(\rho)\, d\rho = f(g), \quad \forall f \in L^2_0(\mathscr{G}).$$

D. Wavelets on homogeneous spaces

We have two options to construct wavelets on homogeneous spaces:
The naive way: We apply the wavelet transform to the lifted function $\tilde{f}(g) = f(g \cdot x_0)$ with base-point $x_0 \in \mathscr{X} = \mathscr{G}/\mathscr{H}$ for some $f \in L^2(\mathscr{X})$. This defines a function on $\mathbb{R}_+ \times \mathscr{G}$ via

$$(\mathscr{W}\tilde{f})(\rho, g) = \int_{\mathscr{G}} \tilde{f}(h)\check{\psi}_\rho(h^{-1}g)\, dh$$
$$= \int_{\mathscr{G}} f(h \cdot x_0)\check{\psi}_\rho(h^{-1}g)\, dh$$

But we would prefer to have a transform living on $\mathbb{R}_+ \times \mathscr{X}$ instead of $\mathbb{R}_+ \times \mathscr{G}$.

For that we introduce the following zonal product

$$f \hat{\bullet} \psi(x) = \int_{\mathscr{G}} \overline{f(g \cdot x_0)}\psi(g \cdot x)\, dg \in L^1(\mathscr{G}).$$

Definition 4. *Let $\mathscr{X} = \mathscr{G}/\mathscr{H}$ be a homogeneous space and p_t be a diffusive approximate identity and $\alpha(\rho) > 0$ be a given weight function. A family $\psi_\rho \in L^2(\mathscr{X})$ is called a* diffusive wavelet family *if the admissibility condition*

$$p_t^{\mathscr{X}}(x)\big|_{\hat{\mathscr{G}}_+} = \int_t^\infty \psi_\rho \hat{\bullet} \psi_\rho(x)\, \alpha(\rho)\, d\rho$$

is satisfied.
We associate to this family the wavelet transform

$$(\mathscr{W}_{\mathscr{X}} f)(\rho, g) = f \bullet \psi_\rho(g) = \int_{\mathscr{X}} f(x)\overline{\psi(g^{-1}\cdot x)}\, dx.$$

Theorem 3. *The wavelet transform $\mathscr{W}_{\mathscr{X}} : L_0^2(\mathscr{X}) \to L^2(\mathbb{R}_+ \times \mathscr{G},\, \alpha(\rho)\, d\rho \otimes dg)$ is invertible on its range by*

$$\tilde{f} = \int_{\to 0}^{\infty} (\mathscr{W}_{\mathscr{X}} f)(\rho, \cdot) * \tilde{\psi}_\rho\, \alpha(\rho)\, d\rho \quad \textit{for all} \quad f \in L_0^2(X).$$

E. The non-compact case

We only mention the key points for this case. The spectrum of the Laplacian of non-compact groups becomes continuous. Consequently, the expansion in eigenfunctions of the Laplacian becomes a direct integral

$$f(g) = \int_{\mathbb{R}}^{\oplus} \hat{f}(\lambda)\pi_\lambda(g)\, d\mu(\lambda).$$

The critical question here is to have an appropriate Fourier transform. That means, does there exist a measure $d\mu$ on $\widehat{\mathscr{G}}$, such that the integral

$$\int_{\widehat{G}} \hat{f}(\lambda)\pi_\lambda\, d\mu(\lambda), \quad \text{where } \hat{f}(\lambda) := \int_{\mathscr{G}} \pi_\lambda^*(g)\, f(g)\, dg$$

is well-defined for some function space on $\mathscr{G}$. If such measure exists it is called Plancherel measure. In this case the construction of diffusive wavelets works similar to the compact case. In general a Plancherel measure does not exist for locally compact groups. But since the Plancherel measure exists for nilpotent Lie groups, one can extend diffusive wavelets to nilpotent Lie groups.

III. Wavelet packets

Definition 5. *Let $\{\rho_j,\, j \in \mathbb{Z}\}$ be a strictly decreasing sequence of real numbers such that*

$$\lim_{j\to\infty} \rho_j = 0 \textit{ and } \lim_{j\to-\infty} \rho_j = \infty.$$

Let $\{\Psi_\rho,\, \rho > 0\}$ be a family of diffusive wavelets. A wavelet packet *is defined by*

$$\hat{\Psi}_j^P(\pi) = \left(\int_{\rho_{j+1}}^{\rho_j} (\hat{\Psi}_\rho)^2 \alpha(\rho)\, d\rho\right)^{\frac{1}{2}},$$

and in spatial domain

$$\Psi_j^P = \sum_{\pi\in\hat{G}} d_\pi \lambda_\pi \left(\int_{\rho_{j+1}}^{\rho_j} e^{-\rho\lambda_\pi^2}\, d\rho\right)^{\frac{1}{2}} \operatorname{trace}(\eta(\pi)\pi(g)).$$

The wavelet transform is now given by

$$(\mathscr{W}^P f)(j, g) := (f * \check{\Psi}_j^P)(g).$$

Theorem 4. *The wavelet transform $\mathscr{W}^P$ is an isometry $L^2(\mathscr{G}) \to L^2(\mathbb{Z}\times\mathscr{G})^2$.*

Theorem 5. *The wavelet transform $\mathscr{W}^P$ is invertible on its range by*

$$f(g) = \sum_{j\in\mathbb{Z}} (\mathscr{W}^P f)(j, \cdot) * \Psi_j^P(\cdot))(g).$$

A common strategy is to build up a multiresolution analysis corresponding to Ψ^P.

IV. Examples

1) The torus $\mathbb{T}^n$: Let $\mathbb{T}^n$ denote the n-dimensional torus which can be identified with

$$\mathbb{T}^n = \mathbb{R}^n/(2\pi\mathbb{Z})^n.$$

We will identify n-fold periodic functions on $\mathbb{R}^n$ with their projection on $\mathbb{T}^n$. The corresponding projection will be called periodization and is defined by

$$\mathbb{P}f(x) = \sum_{\omega\in 2\pi\mathbb{Z}^n} f(x+\omega).$$

In particular, the periodization of the heat kernel on $\mathbb{R}^n$ give the heat kernel on $\mathbb{T}^n$. We have

$$e_t^{heat,\mathbb{R}^n}(x) = \frac{1}{2(\pi t)^n} e^{\frac{||x||^2}{4t}}$$

Let $m \in \mathbb{Z}^n$. For $f \in L^2(\mathbb{T}^n)$ we have

$$f(x) = \sum_{m\in\mathbb{Z}^n} \hat{f}(m)\, e^{i\sum_{j=1}^n m_j x_j},$$
$$\hat{f}(m) = \frac{1}{(2\pi)^n} \int_{[0,2\pi]^n} f(x) e^{-i\sum_{j=1}^n m_j x_j}\, dx.$$

The Fourier coefficients of the heat kernel $e_t^{heat,\mathbb{T}^n}$ can be given explicitly

$$\hat{e}_t^{heat,\mathbb{T}^n} = \frac{1}{(2\pi)^n} \int_{[0,2\pi]^n} \sum_{\omega\in 2\pi\mathbb{Z}^n} e_t^{heat,\mathbb{R}^n}(x+\omega) e^{-i\sum_{j=1}^n m_j x_j} dx$$
$$= \frac{1}{(2\pi)^n} \int_{\mathbb{R}^n} \frac{1}{2(\pi t)^n} e^{\frac{||x||^2}{4t}} e^{-i\sum_{j=1}^n m_j x_j}\, dx = \frac{1}{2\pi^n} e^{-\sum_{j=1}^n m_j^2 t}.$$

Let $\{\psi_\rho\}$ be a subfamily of $L^2(\mathbb{T}^n)$. the wavelet we are looking for has the Fourier series expansion

$$\hat{\psi}_\rho(x) = \sum_{m\in\mathbb{Z}^n} \frac{1}{\sqrt{2\pi^n}} \sum_{j=1}^n m_j^2 e^{-\sum_{j=1}^n m_j^2 \rho} e^{i\sum_{j=1}^n m_j x_j}.$$

2) The unit sphere S^n: The unit sphere is a homogeneous space $S^n = SO(n+1)/SO(n)$. An orthonormal system in $L^2(S^n)$ is given by the *spherical harmonics* $\{Y_k^i, k = 0, 1, \dots, i = 1, \dots d_k(n)\}$, where $d_k(n) = (2k+n-1)\frac{(k+n-2)!}{k!(n-1)!}$. We denote by C_k^λ the Gegenbauer polynomials of order $\lambda = \frac{n-1}{2}$. The eigenvalues of the Laplace-Beltrami operator on S^n are $-\lambda_k^2 = -k(k+n-2)$ and the heat kernel is given by

$$e_t^{heat,S^n}(x) = \sum_{k=0}^{\infty} d_k(n) e^{-\lambda_k^2 t} \frac{C_k^\lambda(x_0 \cdot x)}{C_k^\lambda(1)} = \sum_{k=0}^{\infty} \frac{2k+n-1}{n-1} e^{-k(k+n-2)t} C_k^\lambda(x_0 \cdot x),$$

where x_0 is base point. Let $\alpha(\rho) > 0$ be a weight function on S^n. Then (radial) diffusive wavelets are given by

$$\Psi_\rho(x) = \frac{1}{\sqrt{\alpha(\rho)}} \sum_{k=0}^{\infty} \frac{(2k+n-1)\lambda_k}{n-1} e^{-\lambda_k^2 \rho/2} C_k^\lambda(x_0 \cdot x),$$

where $\lambda_k^2 = k(k+n-2)$. This construction is based on the Gauß-Weierstraß kernel. A similar construction can be done with the Abel-Poisson kernel, where $\lambda_k^2 = k$.

3) The compact group SO(3): For SO(3) all irreducible representations are unitary equivalent to one of the irreducible components of the quasi-regular representation in $L^2(S^2)$. In $L^2(S^2)$ the translation invariant subspaces are spanned by the spherical harmonics of the same degree of homogenity. We have $d(2) = 2k+1$ and the eigenvalues of the Laplace-Beltrami operator are $-\lambda_k^2 = -k(k+1)$. The eigenfunctions are the so-called Wigner polynomials. Hence the heat kernel on $SO(3)$ is

$$e_t^{SO(3)}(g) = \frac{1}{4\pi} \sum_{k=0}^{\infty} (2k+1) e^{-k(k+1)} C_{2k}\left(\sin\left(\frac{\gamma(g)}{2}\right)\right),$$

where $\gamma(g)$ denotes the angle of g [16]

$$\gamma(g) = \arccos\left(\frac{\operatorname{trace}(g)-1}{2}\right).$$

By our construction a family of wavelets on $SO(3)$ corresponding to the heat kernel is given by

$$\Psi_\rho(g) = \frac{1}{\sqrt{\alpha(g)}} \frac{1}{4\pi} \sum_{k=0}^{\infty} (2k+1)\sqrt{k(k+1)} e^{\frac{k(k+1)}{2}\rho} C_{2k}^1\left(\sin\left(\frac{\gamma(g)}{2}\right)\right).$$

4) The Heisenberg group: The construction of diffusive wavelets is not restricted to compact groups and homogeneous spaces. As long as we have some Plancharel formula we can construct diffusive wavelets. Therefore we can construct diffusive wavelets on the Heisenberg group. Since the Heisenberg group is noncompact we cannot use the Peter-Weyl theorem. But fortunately similar results can be obtained from the Stone-von-Neumann theorem. Due to the existence of a Plancherel measure the Fourier transform can be developed in a similar way, where the sum over irreducible representations becomes an integral since the spectrum of the Laplacian is continuous. While the Laplacian involves a complete basis of the Lie algebra, the sub-Laplacian involves only those operators which corresponds to vector fields belonging to the sub-Riemannian structure. Therefore we consider the heat equation

$$(\Delta_{sub} - \partial_r) u((x,y,t),r) = 0$$

with fundamental solution

$$p_r(x,y,t) = \int (2\pi)^{n/2} \sum_{k=0}^{\infty} \sum_{|\alpha|=k} e^{-((2|\alpha|+n)|\lambda|)r} \phi_k^\lambda(x,y,t)\, d\mu(\lambda),$$

where $\phi_k^\lambda(x,y,t)$ are the radial-symmetric eigenfunctions of Δ_{sub}. For the three dimensional Heisenberg group H^1 we obtain the diffusive wavelets

$$\Psi_\rho(x,y,t) = -\sum_{k=0}^{\infty} \left(\frac{1}{k!} \frac{1}{(it-(2k+1)\frac{\rho}{2})} \left(1 + \frac{\frac{1}{2}|x+iy|^2}{it-(2k+1)\frac{\rho}{2}}\right)^k + \frac{(-1)^k}{k!} \frac{1}{(-it-(2k+1)\frac{\rho}{2})} \left(1 + \frac{\frac{1}{2}|x+iy|^2}{it-(2k+1)\frac{\rho}{2}}\right)^k \right) e^{-\frac{1}{4}|x+iy|^2}.$$

REFERENCES

[1] J.-P. Antoine and P. Vandergheynst, *Wavelets on the n-sphere and other manifolds*, J. Math. Physics, **39**, 3987–4008, 1998,

[2] J.-P. Antoine, L. Demanet, L. Jacques and P. Vandergheynst, *Wavelets on the Sphere: Implementation and Approximations,* Appl. Comput. Harmon. Anal., **13**, No.3, 177-200, 2002,

[3] G. Bernardes, S. Bernstein, P. Cerejeiras, U. Kähler, *Wavelets invariant under finite reflection groups,* Math. Meth. Appl. Sci., Special Issue: Complex-Analytic Methods, 33(4), 473–484, 2010,

[4] S. Bernstein, S. Ebert, *Wavelets on S^3 and $SO(3)$ – their construction, relation to each other and Radon transform of wavelets on $SO(3)$*. Math. Methods Appl. Sci., 33(16):1895–1909, 2010,

[5] S. Bernstein, S. Ebert, R.S. Krausshar, *On the diffusion equation and diffusion wavelets on flat cylinders and the n-torus,* Math. Meth. Appl. Sci. 34, No. 4, 428–441, 2011,

[6] S. Bernstein, S. Ebert, *Kernel based Wavelets on S^3*, Journal of Concrete and Applicable Mathematics, 8(1), 110–124, 2010,

[7] S. Bernstein, S. Ebert, I. Pesenson, *Generalized Splines for Radon Transform on Compact Lie Groups with Applications to Crystallography,* J. Fourier Anal. Appl., doi 10.1007/s00041-012-9241-6, 2012,

[8] S. Bernstein, R. Hielscher, H. Schaeben, *The generalized totally geodesic Radon transform and its application to texture analysis,* Math. Meth. Appl. Sci. 32, 379-394, 2009,

[9] S. Bernstein, I. Pesenson, *Crystallographic and geodesic Radon transforms on $SO(3)$: motivation, generalization, discretization*, accepted,

[10] J. Bremer, R. Coifman, M. Maggioni, A. Szlam, Diffusion wavelet packets, *Appl. Comput. Harmon. Anal.* 21(1): 95–112, 2006,

[11] R.R. Coifman, M. Maggioni, Diffusion wavelets, *Appl. Comput. Harmon. Anal.*, **21**, 53-94, 2006,

[12] S. Ebert, *Wavelets and Lie groups and homogeneous spaces.* PhD thesis, TU Bergakademie Freiberg, 2011.

[13] S. Ebert, J. Wirth, *Diffusive wavelets on groups and homogeneous spaces.* Proc. Roy. Soc. Edinburgh 141A:497-520, 2011,

[14] W. Freeden, T. Gervens and M. Schreiner, *Constructive Approximation on the Sphere with Applications to Geomathematics,* Oxford Sciences Publ., Clarendon Press, Oxford, 1998,

[15] D. Geller, A. Mayeli, *Nearly Tight Frames and Space-Frequency Analysis on Compact Manifolds*, Math. Z., **263**, 235 - 264, 2009,

[16] R. Hielscher. Die Radontransformation auf der Drehgruppe – Inversion und Anwendung in der Texturanalyse. PhD thesis, TU Bergakademie Freiberg, 2007,

[17] D. Geller, I. Pesenson, *Band-limited localized Parseval frames and Besov spaces on compact homogeneous manifolds*, J. Geom. Anal. 21(2), 334-371, 2011,

[18] I. Pesenson, D. Geller, Cubature formulas and discrete Fourier transform on compact manifolds, in Farkas, H.E., *From Fourier anaysis and number theory to Radon transform and geometry*, Developments in Mathematics 28, Springer, Berlin, 431-453, 2013.

Shannon Sampling and Parseval Frames on Compact Manifolds

Isaac Z. Pesenson
Temple University and CCP
Philadelphia, USA
Email: pesenson@temple.edu

I. Introduction

The problem of representation and analysis of functions defined on manifolds (signals, images, and data in general) is ubiquities in many fields ranging from statistics and cosmology to neuroscience and biology. It is very common to consider input signals as points in a high-dimensional measurement space, however, meaningful structures lay on a manifold embedded in this space.

In the last decades, the importance of these applications triggered the development of various generalized wavelet bases suitable for the unit spheres S^2 and S^3 and the rotation group of $\mathbf{R}^3$. The goal of the present study is to describe a general approach to bandlimited localized Parseval frames in a space $L_2(\mathbf{M})$, where $\mathbf{M}$ is a compact homogeneous Riemannian manifold.

One can think of a Riemannian manifold as of a surface in a Euclidean space. A homogeneous manifold is a surface with "many" symmetries like the sphere $x_1^2 + ... + x_d^2 = 1$ in Euclidean space $\mathbf{R}^d$.

Our construction of frames in a function space $L_2(\mathbf{M})$ heavily depends on proper notions of bandlimitedness and Shannon-type sampling on a manifold $\mathbf{M}$. The crucial role in this development is played by positive cubature formulas (Theorem 1.3) and by the product property (Theorem 1.2), which were proved in [1] and [10].

The notion of bandlimideness on a compact manifold $\mathbf{M}$ is introduced in terms of eigenfunctions of a certain second-order differential elliptic operator on $\mathbf{M}$. The most important fact for our construction of frames is that in a space of ω-bandlimited functions the regular $L_2(\mathbf{M})$ norm can be descretized. This result in the case of compact manifolds (and even non-compact manifolds of bounded geometry) was first discovered and explored in many ways in our papers [**?**]-[**?**]. In the classical cases of straight line $\mathbf{R}$ and circle $\mathbf{S}$ the corresponding results are known as Plancherel-Polya and Marcinkiewicz-Zygmund inequalities. Our generalization of Plancherel-Polya and Marcinkiewicz-Zygmund inequalities implies that ω-bandlimited functions on manifolds are completely determined by their vales on discrete sets of points "uniformly" distributed over $\mathbf{M}$ with a spacing comparable to $1/\sqrt{\omega}$ and can be completely reconstructed in a stable way from their values on such sets. The last statement is an extension of the Shannon sampling theorem to the case of Riemannian manifolds.

Our article is a summary of some results for Riemannian manifolds that were obtained in [1]-[12]. To the best of our knowledge these are the pioneering papers which contain the most general results about frames, Shannon sampling, and cubature formulas on compact and non-compact Riemannian manifolds. In particular, the paper [1] gives an "end point" construction of tight localized frames on homogeneous compact manifolds. The paper [11] is the first systematic development of localized frames on compact domains in Euclidean spaces.

A. Compact homogeneous manifolds

A homogeneous compact manifold $\mathbf{M}$ is a C^∞-compact manifold on which a compact Lie group G acts transitively. In this case $\mathbf{M}$ is necessary of the form G/H, where H is a closed subgroup of G. The notation $L_2(\mathbf{M})$, is used for the usual Hilbert spaces, where dx is an invariant measure.

If $\mathbf{g}$ is the Lie algebra of a compact Lie group G then it is a direct sum $\mathbf{g} = \mathbf{a} + [\mathbf{g}, \mathbf{g}]$, where $\mathbf{a}$ is the center of $\mathbf{g}$, and $[\mathbf{g}, \mathbf{g}]$ is a semi-simple algebra. Let Q be a positive-definite quadratic form on $\mathbf{g}$ which, on $[\mathbf{g}, \mathbf{g}]$, is opposite to the Killing form. Let $X_1, ..., X_d$ be a basis of $\mathbf{g}$, which is orthonormal with respect to Q. Since the form Q is $Ad(G)$-invariant, the operator

$$-X_1^2 - X_2^2 - \ ... - X_d^2, \ d = dim\ G$$

is a bi-invariant operator on G, which is known as the Casimir operator. This implies in particular that the corresponding operator on $L_2(\mathbf{M})$,

$$L = -D_1^2 - D_2^2 - ... - D_d^2, \quad D_j = D_{X_j}, \ d = dim\ G, \quad (1)$$

commutes with all operators $D_j = D_{X_j}$. Operator L, which is usually called the Laplace operator, is the image of the Casimir operator under differential of quazi-regular representation in $L_2(\mathbf{M})$. Note that if $\mathbf{M} = G/H$ is a compact symmetric space then the number $d = dim\, G$ of operators in the formula (1) can be strictly bigger than the dimension $n = dim\, \mathbf{M}$. For example on a two-dimensional sphere $\mathbf{S}^2$ the Laplace-Beltrami operator $L_{\mathbf{S}^2}$ is written as $L_{\mathbf{S}^2} = D_1^2 + D_2^2 + D_3^2$, where $D_i, i = 1, 2, 3$, generates a rotation in $\mathbf{R}^3$ around coordinate axis x_i: $D_i = x_j\partial_k - x_k\partial_j$, where $j, k \neq i$.

It is important to realize that in general, the operator L is not necessarily the Laplace-Beltrami operator of the natural

invariant metric on $\mathbf{M}$. But it coincides with such operator at least in the following cases:

1) If the manifold $\mathbf{M}$ is itself a compact Lie group G then L is exactly the Laplace-Beltrami operator of an invariant metric on G. In particular it happens if $\mathbf{M}$ is an n-dimensional torus, and L is the sum of squares of partial derivatives;
2) If $\mathbf{M} = G/H$ is a compact symmetric space of rank one, then the operator L is proportional to the Laplace-Beltrami operator of an invariant metric on G/H. This follows from the fact that, in the rank one case, every second-order operator which commutes with all isometries $x \to g \cdot x, \quad x \in \mathbf{M}, \quad g \in G$, is proportional to the Laplace-Beltrami operator. The important examples of such manifolds are spheres and projective spaces.

Since manifold $\mathbf{M}$ is compact and L is a second-order differential elliptic self-adjoint positive definite operator $L_2(\mathbf{M})$ it has a discrete spectrum $0 = \lambda_0 < \lambda_1 \leq \lambda_2 \leq .....$ which goes to infinity and there exists a complete family $\{u_j\}$ of orthonormal eigenfunctions which form a basis in $L_2(\mathbf{M})$.

Definition 1.1: The span of eigenfunctions u_j

$$Lu_j = \lambda_j u_j$$

with $\lambda_j \leq \omega, \quad \omega > 0$, is denoted as $\mathbf{E}_\omega(L)$ and is called the space of bandlimited functions on $\mathbf{M}$ of bandwidth ω.

According to the Weyl's asymptotic formula one has

$$dim\, \mathbf{E}_\omega(L) \sim C\, Vol(\mathbf{M})\omega^{n/2}, \tag{2}$$

where $n = dim\, \mathbf{M}$ and C is an absolute constant.

Let $B(x, r)$ be a metric ball on a compact Riemannian manifold $\mathbf{M}$ whose center is x and radius is r. The following lemma can be found in [2], [5].

Lemma 1.1: There exists a natural number $N_{\mathbf{M}}$, such that for any sufficiently small $\rho > 0$, there exists a set of points $\{x_k\}$ such that:

1) the balls $B(x_k, \rho/4)$ are disjoint,
2) the balls $B(x_k, \rho/2)$ form a cover of $\mathbf{M}$,
3) the multiplicity of the cover by balls $B(x_k, \rho)$ is not greater than $N_{\mathbf{M}}$.

Definition 1.2: Any set of points $\mathbf{M}_\rho = \{x_k\}$ which is described in Lemma 1.1 will be called a metric ρ-lattice.

The following theorems are of primary importance.

Theorem 1.2: (Product property [1], [10]) If $\mathbf{M} = G/H$ is a compact homogeneous manifold and L is the same as above, then for any f and g belonging to $\mathbf{E}_\omega(L)$, their product fg belongs to $\mathbf{E}_{4d\omega}(L)$, where d is the dimension of the group G.

Remark 1: At this moment it is not known if the constant $4d$ can be lowered in general situation. However, it is easy to verify that in the case of two-point homogeneous manifolds (which include spheres and projective spaces) a stronger result holds: if $f,\ g \in \mathbf{E}_\omega(L)$ then $fg \in \mathbf{E}_{2\omega}(L)$.

Theorem 1.3: (Cubature formula [1], [10]) There exists a positive constant $c = c(\mathbf{M})$, such that if $\rho = c\omega^{-1/2}$, then for any ρ-lattice $\mathbf{M}_\rho$, there exist strictly positive coefficients $\alpha_{x_k} > 0,\ x_k \in \mathbf{M}_\rho,$ for which the following equality holds for all functions in $\mathbf{E}_\omega(\mathbf{M})$:

$$\int_{\mathbf{M}} f dx = \sum_{x_k \in \mathbf{M}_\rho} \alpha_{x_k} f(x_k). \tag{3}$$

Moreover, there exists constants $c_1,\ c_2$, such that the following inequalities hold:

$$c_1 \rho^n \leq \alpha_{x_k} \leq c_2 \rho^n,\ n = dim\ \mathbf{M}. \tag{4}$$

II. Hilbert frames

Since eigenfunctions have perfect localization properties in the spectral domain they cannot be localized on the manifold.

It is the goal of our development to construct "better bases" in corresponding $L_2(\mathbf{M})$ spaces which will have rather strong localization on a manifold and in the spectral domain.

In fact, the "kind of basis" which we are going to construct is known today as a frame.

A set of vectors $\{\psi_v\}$ in a Hilbert space $\mathcal{H}$ is called a frame if there exist constants $A, B > 0$ such that for all $f \in \mathcal{H}$

$$A\|f\|_2^2 \leq \sum_v |\langle f, \psi_v \rangle|^2 \leq B\|f\|_2^2. \tag{5}$$

The largest A and smallest B are called lower and upper frame bounds.

The set of scalars $\{\langle f, \psi_v \rangle\}$ represents a set of measurements of a signal f. To synthesize signal f from this set of measurements one has to find another (dual) frame $\{\Psi_v\}$ and then a reconstruction formula is

$$f = \sum_v \langle f, \psi_v \rangle\, \Psi_v. \tag{6}$$

Dual frame is not unique in general. Moreover it is difficult to find a dual frame. If in particular $A = B = 1$ the frame is said to be tight or Parseval.

The main feature of Parseval frames is that decomposing and synthesizing a signal or image from known data are tasks carried out with the same set of functions. In other words in (6) one can have $\Psi_v = \psi_\nu$.

Parseval frames are similar in many respects to orthonormal wavelet bases. For example, if in addition all vectors ψ_v are unit vectors, then the frame is an orthonormal basis. However, the important differences between frames and, say, orthonormal bases is their redundancy that helps reduce the effect of noise in data.

Frames in Hilbert spaces of functions whose members have simultaneous localization in space and frequency arise naturally in wavelet analysis on Euclidean spaces when continuous wavelet transforms are discretized. Such frames have been constructed, studied, and employed extensively in both theoretical and applied problems.

III. Bandlimited localized Parseval frames on compact homogeneous manifolds

According to spectral theorem if F is a Schwartz function on the line, then there is a well defined operator $F(L)$ in the

space $L_2(\mathbf{M})$ such that for any $f \in L_2(\mathbf{M})$ one has

$$(F(L)f)(x) = \int_{\mathbf{M}} \mathcal{K}^F(x,y) f(y) dy, \tag{7}$$

where dy is the invariant normalized measure on $\mathbf{M}$ and

$$\mathcal{K}^F(x,y) = \sum_{j=0}^{\infty} F(\lambda_j) u_j(x) \overline{u_j}(y). \tag{8}$$

We will be especially interested in operators of the form $F(t^2 L)$, where F is a Schwartz function and $t > 0$. The corresponding kernel will be denoted as $\mathcal{K}_t^F(x,y)$ and

$$\mathcal{K}_t^F(x,y) = \sum_{j=0}^{\infty} F(t^2\lambda_j) u_j(x) \overline{u_j}(y). \tag{9}$$

Note, that variable t here is a kind of scaling parameter.

Localization properties of the kernel $\mathcal{K}_t^F(x,y)$ are given in the following statement.

Lemma 3.1: If L is an elliptic self-adjoint second order differential operators on compact manifolds, then the following holds

1) If F is any Schwartz function on $\mathbf{R}$, then

$$\mathcal{K}_t^F(x,x) \sim c\, t^{-d}, \quad t \to 0. \tag{10}$$

2) If, in addition, $F \in C_c^{\infty}(\mathbf{R})$ is even, then on $\mathbf{M} \times \mathbf{M} \setminus \Delta$, where $\Delta = \{(x,x)\}$, $x \in \mathbf{M}$, $\mathcal{K}_t^F(x,y)$ vanishes to infinite order as t goes to zero.

Let $g \in C^{\infty}(\mathbf{R}_+)$ be a monotonic function such that $supp\ g \subset [0,\ 2^2]$, and $g(s) = 1$ for $s \in [0,\ 1]$, $0 \leq g(s) \leq 1$, $s > 0$. Setting $G(s) = g(s) - g(2^2 s)$ implies that $0 \leq G(s) \leq 1$, $s \in supp\ G \subset [2^{-2},\ 2^2]$. Clearly, $supp\ G(2^{-2j}s) \subset [2^{2j-2}, 2^{2j+2}]$, $j \geq 1$. For the functions $F_0(s) = \sqrt{g(s)}$, $F_j(s) = \sqrt{G(2^{-2j}s)}$, $j \geq 1$, one has $\sum_{j\geq 0} F_j^2(s) = 1$, $s \geq 0$. Using the spectral theorem for L one can define bounded self-adjoint operators $F_j(L)$ as

$$F_j(L)f(x) = \int_{\mathbf{M}} \mathcal{K}_{2^{-j}}^F(x,y) f(y) dy,$$

where

$$\mathcal{K}_{2^{-j}}^F(x,y) = \sum_{\lambda_m \in [2^{2j-2}, 2^{2j+2}]} F(2^{-2j}\lambda_m) u_m(x) \overline{u_m(y)}. \tag{11}$$

The same spectral theorem implies $\sum_{j\geq 0} F_j^2(L)f = f$, $f \in L_2(\mathbf{M})$, and taking inner product with f gives

$$\|f\|^2 = \sum_{j\geq 0} \left\langle F_j^2(L)f, f \right\rangle = \sum_{j\geq 0} \|F_j(L)f\|^2. \tag{12}$$

Moreover, since the function $F_j(s)$ has its support in $[2^{2j-2},\ 2^{2j+2}]$ the functions $F_j(L)f$ are bandlimited to $[2^{2j-2},\ 2^{2j+2}]$.

Consider the sequence $\omega_j = 2^{2j+2}, j = 0, 1, ...$ By (12) the equality $\|f\|^2 = \sum_{j\geq 0} \|F_j(L)f\|^2$ holds, were every function $F_j(L)f$ is bandlimited to $[2^{2j-2},\ 2^{2j+2}]$. Since for every $F_j(L)f \in \mathbf{E}_{2^{2j+2}}(L)$ one can use Theorem 1.2 to conclude that

$$|F_j(L)f|^2 \in \mathbf{E}_{4d2^{2j+2}}(L),$$

where $d = dim\ G$, $\mathbf{M} = G/H$. This shows that for every $f \in L_2(\mathbf{M})$ we have the following decomposition

$$\sum_{j\geq \infty} \|F_j(L)f\|_2^2 = \|f\|_2^2, \quad |F_j(L)f|^2 \in \mathbf{E}_{4d2^{2j+2}}(L). \tag{13}$$

According to Theorem 1.3 there exists a constant $a > 0$ such that for all integers j if

$$\rho_j = a d^{-1/2} 2^{-j} \sim 2^{-j},\ d = dim\ G,\ \mathbf{M} = G/H, \tag{14}$$

then for any ρ_j-lattice $\mathbf{M}_{\rho_j}$ one can find coefficients $\mu_{j,k}$ with $\mu_{j,k} \sim \rho_j^n, n = dim \mathbf{M}$, for which the following exact cubature formula holds

$$\|F_j(L)f\|_2^2 = \sum_{k=1}^{K_j} \mu_{j,k} \left|F_j(L)f(x_{j,k})\right|^2, \tag{15}$$

where $x_{j,k} \in \mathbf{M}_{\rho_j}$, $k = 1, \ldots, K_j = card\,(\mathbf{M}_{\rho_j})$. Using the kernel $\mathcal{K}_{2^{-j}}^F$ of the operator $F_j(L)$ we define the functions

$$\Theta_{j,k}(y) = \sqrt{\mu_{j,k}}\, \overline{\mathcal{K}_{2^{-j}}^F}(x_{j,k}, y) =$$

$$\sqrt{\mu_{j,k}} \sum_{\lambda_m \in [2^{2j-2}, 2^{2j+2}]} \overline{F}(2^{-2j}\lambda_m) \overline{u}_m(x_{j,k}) u_m(y). \tag{16}$$

We find that for every $f \in L_2(\mathbf{M})$ the following equality holds $\|f\|_2^2 = \sum_{j,k} |\langle f, \Theta_{j,k}\rangle|^2$.

Theorem 3.2: (Kernel localization [1]) If $\mathbf{M}$ is compact then the functions $\Theta_{j,k}$ are localized around the points $x_{j,k}$ in the sense that for any $N > 0$ there exists a $C(N) > 0$ such that

$$|\Theta_{j,k}(x)| \leq C(N) \frac{2^{dj}}{\max(1,\ 2^j d(x, x_{j,k}))^N}, \tag{17}$$

for all natural j.

Theorem 3.3: (Bandlimited localized Parseval localized frames on homogeneous manifolds) For any compact homogeneous manifold $\mathbf{M}$ the set of functions $\{\Theta_{j,k}\}$, constructed in (16) forms a Parseval frame in the Hilbert space $L_2(\mathbf{M})$. In particular the following reconstruction formula holds true

$$f = \sum_{j\geq 0} \sum_{k=1}^{K_j} \langle f, \Theta_{j,k}\rangle \Theta_{j,k}, \tag{18}$$

with convergence in $L_2(\mathbf{M})$. Every $\Theta_{j,k}$ is bandlimited to $[2^{2j-2}, 2^{2j+2}]$ and its localization on manifold is given by (17).

The condition (14) imposes a specific rate of sampling in (15). It is interesting to note that this rate is essentially optimal. Indeed, on one hand the Weyl's asymptotic formula (2) gives the dimension of the space $\mathbf{E}_{\omega}(L)$. On the other hand, the condition (14) and the definition of a ρ-lattice imply that the number of points in an "optimal" lattice $\mathbf{M}_{\rho_j}$ for $\rho_j \sim 2^{-j}$ can be approximately estimated as

$$card\, \mathbf{M}_{\rho_j} \sim c \frac{Vol(\mathbf{M})}{2^{-jn/2}} = c Vol(\mathbf{M}) 2^{jn/2}, \quad n = dim\, \mathbf{M},$$

which is in agreement with the Weyl's formula (2) with $\omega \sim 2^j$.

IV. Shannon sampling of bandlimited functions

We consider an even $F \in C_c^\infty(\mathbf{R})$ which equals 1 on $[-1,1]$, and which is supported in $[-\Omega,\ \Omega]$, $\Omega > 1$. Let $\mathcal{K}^F_{\Omega^{-1/2}}(x,y)$ be the kernel of $F(\Omega^{-1}L)$ defined by (9). If $0 < \omega \le \Omega$ then since $F(\Omega^{-1}\lambda_k) = 1$ whenever $\lambda_k \le \omega$, we have that according to (7) - (9) for every $f \in \mathbf{E}_\omega(L)$ the following reproducing formula holds

$$f(x) = \left[F(\Omega^{-1}L)f\right](x) = \int_{\mathbf{M}} \mathcal{K}^F_{\Omega^{-1/2}}(x,y)f(y)dy \quad (19)$$

where dy is the normalized invariant measure. Clearly, for a fixed $x \in \mathbf{M}$ the kernel $\mathcal{K}^F_{\Omega^{-1/2}}(x,y)$ as a function in y belongs to $\mathbf{E}_\Omega(L)$. Thus, for $f \in \mathbf{E}_\omega(L)$, $\omega < \Omega$, the Product property (Theorem 1.2) implies that the product $\mathcal{K}^F_{\Omega^{-1/2}}(x,y)f(y)$ belongs to $\mathbf{E}_{4d\Omega}(L)$, where $d = dim\ G$. Now an application of the Cubature formula (Theorem 1.3) implies the following theorem.

Theorem 4.1: For every compact homogeneous manifold $\mathbf{M} = G/H$ there exists a constant $c = c(\mathbf{M})$ such that for any $\Omega > 0$ and any lattice $\mathbf{M}_\rho = \{x_k\}_{k=1}^{m_\Omega}$ with $\rho = c\Omega^{-1/2}$ one can find positive weights μ_k

$$\mu_k \asymp \Omega^{-n/2},\ \ n = dim\ \mathbf{M},$$

such that for any $f \in \mathbf{E}_\omega(L)$ with $\omega \le \Omega$ the following analog of the Shannon formula holds

$$f(x) = \sum_{k=1}^{m_\Omega} \mu_k f(x_k)\mathcal{K}^F_{\Omega^{-1/2}}(x,x_k), \quad f \in \mathbf{E}_\omega(L). \quad (20)$$

Remark 2: Note that our definition of a ρ-lattice and the Weyl's asymptotic formula (2) for eigenvalues of L imply that m_Ω is "essentially" the dimension of the space $\mathbf{E}_{4d\Omega}(L)$ with $d = dim\ G$. In other words there exists a constants $C_1(\mathbf{M}) > 0$, $C_2(\mathbf{M}) > 0$ (which are independent on Ω) such that the number m_Ω of sampling points satisfies the following inequalities

$$C_1(\mathbf{M})\Omega^{n/2} \le m_\Omega \le C_2(\mathbf{M})\Omega^{n/2}$$

$$C_1(\mathbf{M})\mathbf{E}_{4d\Omega}(L) \le m_\Omega \le C_2(\mathbf{M})\mathbf{E}_{4d\Omega}(L). \quad (21)$$

Remark 3: Lemma 3.1 shows that for large Ω functions $\mathcal{K}^F_{\Omega^{-1/2}}(x,x_k)$ in (20) are essentially localized around sampling points x_k.

V. A discrete formula for evaluating Fourier coefficients on manifolds.

As another application of the Product Property and the Cubature Formula, we prove an analog of the Shannon Sampling Theorem on compact homogeneous manifolds.

Theorems 1.2 and 1.3 imply the following theorem which shows that on a compact homogeneous manifold $\mathbf{M}$ there are finite sets of points which yield exact discrete formulas for computing Fourier coefficients of bandlimited functions.

Theorem 5.1: For every compact homogeneous manifold $\mathbf{M} = G/H$ there exists a constant $c = c(\mathbf{M})$ such that for any $\omega > 0$ and any lattice $\mathbf{M}_\rho = \{x_k\}_{k=1}^{r_\omega}$ with $\rho = c\omega^{-1/2}$ one can find positive weights μ_k comparable to $\omega^{-n/2}$, $n = dim\ \mathbf{M}$, such that Fourier coefficients $c_i(f)$ of any f in $\mathbf{E}_\omega(L)$ with respect to the basis $\{u_i\}_{i=1}^\infty$ can be computed by the following *exact* formula

$$c_i(f) = \int_{\mathbf{M}} f(x)\overline{u_i}(x)dx = \sum_{k=1}^{r_\omega} \mu_k f(x_k)\overline{u_i}(x_k),$$

with r_ω satisfying relations

$$C_1(\mathbf{M})\omega^{n/2} \le r_\omega \le C_2(\mathbf{M})\omega^{n/2}$$

$$C_1(\mathbf{M})\mathbf{E}_{4d\omega}(L) \le r_\omega \le C_2(\mathbf{M})\mathbf{E}_{4d\omega}(L), \quad (22)$$

where $C_1(\mathbf{M})$ and $C_2(\mathbf{M})$ are the same as in (21).

We obviously have the following *"discrete" representation formula of f in $\mathbf{E}_\omega(L)$ in terms of eigenfunctions u_i*

$$f = \sum_i \sum_{k=1}^{r_\omega} \mu_k f(x_k)\overline{u_i}(x_k)u_i. \quad (23)$$

Acknowledgment

The work was supported in part by the National Geospatial-Intelligence Agency University Research Initiative (NURI), grant HM1582-08-1-0019.

References

[1] D. Geller and I. Pesenson, *Band-limited localized Parseval frames and Besov spaces on compact homogeneous manifolds*, J. Geom. Anal. 21 (2011), no. 2, 334-37.

[2] I. Pesenson, *A sampling theorem on homogeneous manifolds*, Trans. Amer. Math. Soc. **352** (2000), no. 9, 4257–4269.

[3] I. Pesenson, *Poincare-type inequalities and reconstruction of Paley-Wiener functions on manifolds,* J. of Geometric Analysis , (4), 1, (2004), 101-121.

[4] I. Pesenson, *An approach to spectral problems on Riemannian manifolds,* Pacific J. of Math. Vol. 215(1), (2004), 183-199.

[5] I. Pesenson, *Poincare-type inequalities and reconstruction of Paley-Wiener functions on manifolds,* J. of Geometric Analysis , (4), 1, (2004), 101-121.

[6] I. Pesenson, *Deconvolution of band limited functions on symmetric spaces*, Houston J. of Math., 32, No. 1, (2006), 183-204.

[7] I. Pesenson, *Frames in Paley-Wiener spaces on Riemannian manifolds*, in Integral Geometry and Tomography, Contemp. Math., 405, AMS, (2006), 137-153.

[8] I. Pesenson, *Bernstein-Nikolski inequality and Riesz interpolation Formula on compact homogeneous manifolds,* J. Approx. Theory,150, (2008), no. 2, 175-198.

[9] I. Pesenson, *A Discrete Helgason-Fourier Transform for Sobolev and Besov functions on noncompact symmetric spaces*, Contemp.Math, 464, AMS, (2008), 231-249.

[10] I. Pesenson, D. Geller, *Cubature formulas and discrete fourier transform on compact manifolds* in "From Fourier Analysis and Number Theory to Radon Transforms and Geometry: In Memory of Leon Ehrenpreis" (Developments in Mathematics 28) by Hershel M. Farkas, Robert C. Gunning, Marvin I. Knopp and B. A. Taylor, Springer NY 2013.

[11] I. Pesenson, *Localized Bandlimited nearly tight frames and Besov spaces on domains in Euclidean spaces*, submitted, arXiv:1208.5165v1.

[12] I. Pesenson, *Paley-Wiener frames and Besov spaces on non-compact manifolds*, submitted.

Signal Analysis with Frame Theory and Persistent Homology

Holger Boche
TU München
Theresienstr. 90/IV (LTI)
80333 München, Germany
boche@tum.de

Mijail Guillemard
TU Berlin
Institut für Mathematik
10623 Berlin, Germany
guillemard@math.tu-berlin.de

Gitta Kutyniok
TU Berlin
Institut für Mathematik
10623 Berlin, Germany
kutyniok@math.tu-berlin.de

Friedrich Philipp
TU Berlin
Institut für Mathematik
10623 Berlin, Germany
philipp@math.tu-berlin.de

***Abstract*—A basic task in signal analysis is to characterize data in a meaningful way for analysis and classification purposes. Time-frequency transforms are powerful strategies for signal decomposition, and important recent generalizations have been achieved in the setting of frame theory. In parallel recent developments, tools from algebraic topology, traditionally developed in purely abstract settings, have provided new insights in applications to data analysis. In this report, we investigate some interactions of these tools, both theoretically and with numerical experiments, in order to characterize signals and their frame transforms. We explain basic concepts in persistent homology as an important new subfield of computational topology, as well as formulations of time-frequency analysis in frame theory. Our objective is to use persistent homology for constructing topological signatures of signals in the context of frame theory. The motivation is to design new classification and analysis methods by combining the strength of frame theory as a fundamental signal processing methodology, with persistent homology as a new tool in data analysis.**

I. Introduction

Modern developments in signal processing have triggered important interactions between pure and applied mathematics. A basic example is given by new advances in time-frequency analysis and its generalizations to frame theory [2, 8], but another recent and major development illustrating a rich interplay between abstract ideas and practical applications is persistent homology [1, 7], which in the last few years has become an important subfield of computational topology. These developments in persistent homology have been applied in different situations, and particular results relevant in our setting are recent results in sensor networks [10, 11]. This report is a natural continuation of our previous work [12] which introduced a strategy for integrating time-frequency analysis with persistent homology. Our contribution now is to further understand and improve these interactions by combining frame theory with the stability of persistent homology.

The outline of this report is as follows. We begin with a short overview of time-frequency analysis and frame theory, with a particular focus on voice transformations and how this setting is generalized in (continuous) frame theory by considering analysis operators $V : \mathcal{H} \to L^2(\mathcal{X})$. Here, $\mathcal{X}$ is a locally compact group for the case of voice transformations, and a locally compact Hausdorff space in frame theory.

We then shortly present elements of persistent homology as a new important branch in data analysis which, given a point cloud data $X = \{x_i\}_{i=1}^m$, (or more generally, a family of simplicial complexes $K_1 \subset K_2 \subset \cdots \subset K_r = \mathcal{X}$) constructs a diagram that encodes a topological features of X (resp. $\mathcal{X}$). We then proof a property combining the basic stability of persistent diagrams with frame theory, and illustrate this concept with computational experiments.

II. Time-frequency Analysis and Frame Theory

Given a Hilbert space $\mathcal{H}$ as, for instance, a functional space of signals $L^2(\mathbb{R})$, the basic strategy in time-frequency analysis is to segment a signal $f \in \mathcal{H}$ in smaller chunks $x_b = fg_b$, for g a window function, and $g_b(t) = g(t-b)$. This segmentation procedure is the basis of *Gabor analysis* and the short term Fourier transform (STFT), and it allows to locally analyze the frequency behavior of f and its evolution in time. A generalization of this method can be described using a locally compact group G acting in a Hilbert space $\mathcal{H}$ (see [8]). This action is an irreducible and square integrable group representation, $\pi : G \to U(\mathcal{H})$, defined as a group homomorphism between G and $U(\mathcal{H})$, the group of unitary operators in $\mathcal{H}$. The basic transformation that is constructed with π is the *analysis operator* or the *voice transform*:

$$V_\psi : \mathcal{H} \to L^2(G), \qquad V_\psi(f)(x) = \langle f, \pi(x)\psi \rangle,$$

which maps each $f \in \mathcal{H}$ to a square integrable function $V_\psi f$ that "unfolds" the content of f in the setting provided by G. We remark that a fundamental property of V is to be a quasi isometry, which allows to perform not only analysis but also synthesis procedures.

A. Continuous and Discrete Frames

Despite the major role of the voice transform and its group representation background, in some applications it is too restrictive to assume the existence of a group G that parametrizes the family of dictionary vectors $\{\pi(x)\psi\}_{x\in G}$. An important generalization of these procedures is frame theory which considers a family of vectors $\{\psi_x\}_{x\in\mathcal{X}}$ in a Hilbert space $\mathcal{H}$, where $\mathcal{X}$ is a locally compact Hausdorff space with a positive Radon measure μ (see [9]). When $\mathcal{X}$ is finite or discrete (e.g. $\mathcal{X} = \mathbb{N}$), we will consider a counting measure

μ, and the resulting concept will be a generalization of an orthogonal basis, and it provides powerful mechanisms for the analysis and synthesis of a signal $f \in \mathcal{H}$.

The main property required by a frame $\{\psi_x\}_{x\in\mathcal{X}} \subset \mathcal{H}$ is the stabilization of the analysis operator.

Definition 1. A set of vectors $\{\psi_x\}_{x\in\mathcal{X}} \subset \mathcal{H}$ in a Hilbert space $\mathcal{H}$ is a *frame*, if

$$A||f||^2 \leq ||Vf||^2 \leq B||f||^2, \quad \forall f \in \mathcal{H}$$

for $0 < A \leq B < \infty$, the *lower and upper frame bounds*, and $V : \mathcal{H} \to L^2(\mathcal{X}), (Vf)(x) = \langle f, \psi_x\rangle$ is the *analysis operator*.

Reducing the difference between A and B improves the stability of V, and for the case of $A = B$, or $A = B = 1$, the resulting frame is denominated *tight frame* and *Parseval frame*, respectively. The corresponding synthesis operator $V^* : L^2(\mathcal{X}) \to V$, with $V^*((a_x)_{x\in\mathcal{X}}) = \int_{\mathcal{X}} a_x\psi_x \, d\mu(x)$ is defined with an adequate positive Radon measure μ, when $\mathcal{X}$ is a locally compact Hausdorff space (see [9]). The maps V^* and V are combined in the frame operator

$$S = V^*V : \mathcal{H} \to \mathcal{H}, Sf = \int_{\mathcal{X}} \langle f, \psi_x\rangle\psi_x d\mu(x),$$

which plays an important role due to the fact that the operator norm of S can be bounded by A and B, namely:

$$A \leq ||S||_{op} \leq B. \tag{1}$$

III. Persistent Homology

In order to shortly introduce the basic concepts in persistent homology, we recall elementary ideas in simplicial homology. One of the simplest homology theories available is simplicial homology which translates topological data into an algebraic formulation. The fundamental objective is to compute qualitative properties of a topological space $\mathcal{X}$, as the number of n-dimensional holes $\mathcal{X}$ has. The basic object to analyze is a (finite) *abstract simplicial complex* K, defined as a nonempty family of subsets of a vertex set $V = \{v_i\}_{i=1}^m$ with $\{v\} \in K$ if $v \in V$, and if $\alpha \in K, \beta \subseteq \alpha$, then $\beta \in K$. We define *faces* (or *simplices*) to be the elements of K, and their corresponding *dimension* will be their cardinality minus one.

In order to compute the number of holes of a given simplicial complex K, we translate its topological or combinatorial properties in the language of linear algebra. There are three basic steps in this procedure: first, we construct a family of free groups C_n, the *group of n-chains* defined as the formal combinations of k-dimensional faces with coefficients in a given group (or rings and fields in more specific cases). Secondly, we construct the *boundary operators* ∂_n, defined as homomorphisms (or more specifically linear maps) between the group of k-chains C_k. The homomorphism maps a face $\sigma = [p_0, \cdots p_n] \in C_n$ into C_{n-1} by

$$\partial_n\sigma = \sum_{k=0}^{n}(-1)^k[p_0, \cdots, p_{k-1}, p_{k+1}, \cdots p_n].$$

Finally, in the third step, we construct the *homology groups* defined as the quotients $H_k := \ker(\partial_k)/\mathrm{im}(\partial_{k+1})$. The main property is now the computation of the *Betti numbers*, which represent the amount of k-dimensional holes, and it corresponds to the rank of the homology groups, $\beta_k = \mathrm{rank}(H_k)$.

The fundamental ideas of persistent homology have been introduced at the end of the last century (see [6]) where the estimation of topological properties of finite sets arises as an important problem in many applications. An important scenario is the analysis of a point cloud data $X = \{x_i\}_{i=1}^m$ which represents the challenging situation that no simplicial structure is given a priori. The strategy is to consider special type of simplicial complexes (e.g. Čech complexes, Vietoris Rips complexes) arising by considering the set $R_\epsilon(X)$, defined with X as the vertex set, and setting the vertices $\sigma = \{x_0, \ldots, x_k\}$ to span a k-simplex of $R_\epsilon(X)$ if $d(x_i, x_j) \leq \epsilon$ for all $x_i, x_j \in \sigma$. The fundamental remark is to notice that for a finite point cloud data $X = \{x_i\}_{i=1}^m$ there is only a finite number of simplicial complexes that fully characterize the family $\{R_\epsilon(X)\}_{\epsilon>0}$. Namely, there is only a finite number of non-homeomorphic simplicial complexes $K_1 \subset K_2 \subset \cdots \subset K_r$ (a so called *filtration*) that fully describe the topology of the sets $\{R_\epsilon(X)\}_{\epsilon>0}$. The power of persistent homology lies in efficient algorithms that compute homology information for the filtration $K_1 \subset K_2 \subset \cdots \subset K_r$.

Definition 2 (Persistent homology)**.** A *filtration* is the basic input of persistent homology, and it is defined for a topological space $\mathcal{X}$, as a family of non-homeomorphic simplicial complexes $K_1 \subset K_2 \subset \cdots \subset K_r = \mathcal{X}$. We define a *persistent homology group* (at the level n) of a filtration as the image of a group homomorphisms $f_n^{ij} : H_n(K_i) \to H_n(K_{i+j})$. The maps f_n^{ij} are induced from the continuous inclusions $K_i \subset K_j$ by the functorial properties of homology. The images of f_n^{ij} represent the homology classes born at i and still alive at $i + j$. The rank of theses images $\beta_n^{ij} = \mathrm{rank}(\mathrm{Im}f_n^{ij})$ is the *persistent Betti number* (at the homology level n). The *persistent diagram* $\mathrm{dgm}(\mathcal{X})$ (at the level n) of $\mathcal{X}$ is constructed by associating the value β_n^{ij} to the pairs (i, j), $1 \leq i \leq j \leq r$.

A. Stability Properties

We now present an important component in the persistent homology toolbox denominated the *stability of persistent diagrams* [5]. In order to explain this concept, we first introduce some preliminary notions.

Definition 3 (Homological critical values and tame functions)**.** Let $\mathcal{X}$ be a topological space, and $\alpha : \mathcal{X} \to \mathbb{R}$ a continuous function. A *homological critical value* (or HCV) is a number $a \in \mathbb{R}$ for which the map induced by α

$$H_n(\alpha^{-1}(]-\infty, a-\epsilon[)) \to H_n(\alpha^{-1}(]-\infty, a+\epsilon[))$$

is not an isomorphism for all $\epsilon > 0$. Remember that each $\alpha^{-1}(]-\infty, a[)$ is a *level sets* of α, and it plays a crucial role in Morse theory, as well as in our current setting. A *tame function* is now defined to be a function $\alpha : \mathcal{X} \to \mathbb{R}$ that has only a finite number of HCV.

Typical examples of tame functions are *Morse functions* on compact manifolds, and piecewise linear functions on finite simplicial complexes [5].

Definition 4. For a tame function $\alpha : \mathcal{X} \to \mathbb{R}$, we define its *persistent diagram* $\mathrm{dgm}(\alpha)$ as the persistent diagram of the filtration $K_1 \subset K_2 \subset \cdots \subset K_r = \mathcal{X}$ where we define $K_i = f^{-1}(]-\infty, a_i])$, and $a_1 < a_2 < \cdots < a_r$ are the critical values of α (see [4]).

Definition 5 (Bottleneck and Hausdorff distances)**.** For two nonempty sets $X, Y \subset \mathbb{R}^2$ the *Hausdorff distance* and *bottleneck distances* are defined as

$$d_H(X,Y) = \max\Big\{ \sup_{x\in X} \inf_{y\in Y} ||x-y||_\infty, \sup_{y\in Y} \inf_{x\in X} ||y-x||_\infty \Big\}$$
$$d_B(X,Y) = \inf_{\gamma} \sup_{x\in X} ||x-\gamma(x)||_\infty,$$

where we consider all possible bijections $\gamma : X \to Y$. Here, we use $||p-q||_\infty = \max\{|p_1-q_1|, |p_2-q_2|\}$ for $p, q \in \mathbb{R}^2$. We also remark the following inequality between these distances: $d_H(X,Y) \leq d_B(X,Y)$ (see [5]).

Theorem 1 (Stability of persistent diagrams [3, 4, 5])**.** Let $\mathcal{X}$ be a topological space with tame functions $\alpha, \beta : \mathcal{X} \to \mathbb{R}$. Then, the following stability property holds:

$$d_B(\mathrm{dgm}(\alpha), \mathrm{dgm}(\beta)) \leq ||\alpha - \beta||_\infty. \tag{2}$$

IV. Frames Analysis and Persistent Homology

Our objective is now to combine the core concepts in frame theory with persistent diagrams in order to combine the strength and features of these different analysis tools. Our theorem provides stability properties of persistent diagrams of frame transforms $|Vf|$, when considering a frame decomposition $Vf \in L^2(\mathcal{X})$. We assume the frame parametrization space $\mathcal{X}$ to have a counting measure, which is anyway the case when considering discrete structures for applications.

Theorem 2. Let $f, g \in \mathcal{H}$ and $|Vf|$, $|Vg|$ tame functions with $V : \mathcal{H} \to L^2(\mathcal{X})$ a frame analysis operator with upper bound B. We consider a discrete topological space $\mathcal{X}$ with a counting measure. Then, the following stability property holds:

$$d_B(\mathrm{dgm}(|Vf|), \mathrm{dgm}(|Vg|)) \leq \sqrt{B}||f-g||_{\mathcal{H}}.$$

Proof: This is a consequence of the inequality (1) (the bounding of the norm of the frame operator) and the stability of the persistent diagrams described in the inequality (2):

$$\begin{aligned} d_B(\mathrm{dgm}(|Vf|), \mathrm{dgm}(|Vg|)) &\\ &\leq ||\,|Vf| - |Vg|\,||_\infty \\ &\leq ||Vf - Vg||_2 \\ &\leq ||V||\,||f-g||_{\mathcal{H}} \\ &= \sqrt{||V^*V||}||f-g||_{\mathcal{H}} \\ &= \sqrt{||S||}\,||f-g||_{\mathcal{H}} \\ &\leq \sqrt{B}||f-g||_{\mathcal{H}}, \end{aligned}$$

where we use $||V||^2 = ||V^*V||$. ■

This proposition is an initial step towards the integration of frame theory and persistent stability. We remark that important developments have been achieved in generalizing the work in [5], and the inequality (2), by avoiding the restrictions imposed by the functional setting and expressing the stability in a purely algebraic language (see [1, 3, 4]). The usage of these more flexible and general stability properties is a natural future step in our program.

A. Experiments

We now experiment with acoustic signals the interaction between the components in our framework (frame transformations and persistent diagrams). A main objective is to study both the stability and the discriminative power of persistent diagrams in the setting of frame theory. We consider two signals f_0 and f_1 together with a process transforming f_0 into f_1 encoded with a family of signals $\{f_t\}_{0\leq t\leq 1}$ defined as:

$$f_t = (1-t)f_0 + tf_1, \quad t \in [0,1].$$

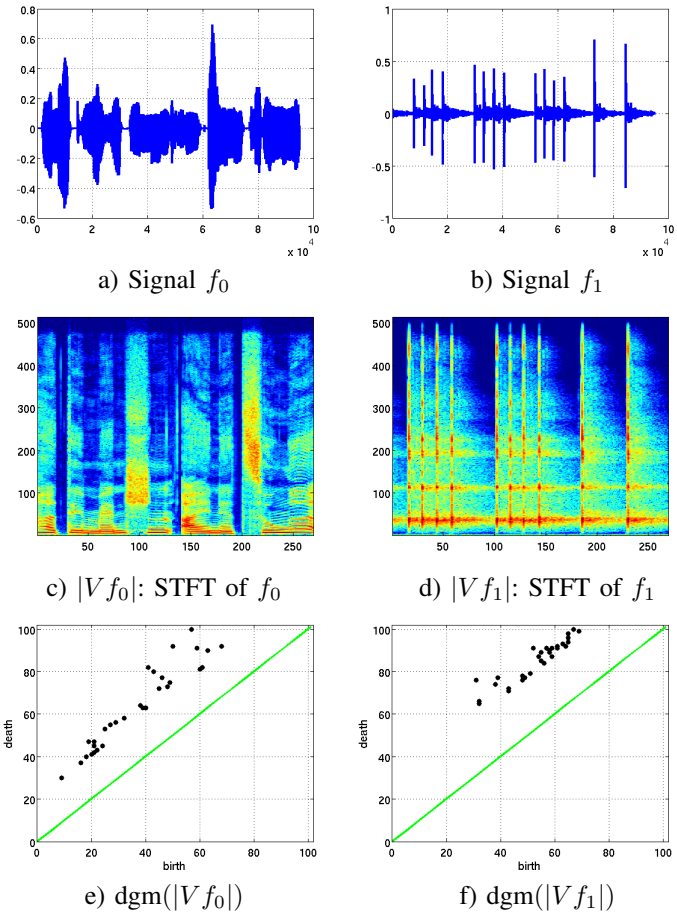


Fig. 1. Time-frequency plots and discriminative properties of persistence

In Fig. 1 (a) and Fig. 1 (b), the plots of f_0 and f_1 are shown, and they represent a female speech recording and a castanet signal respectively. In Fig. 1(c) and Fig. 1(d) the corresponding spectrograms (STFT) $|Vf_0|$ and $|Vf_1|$ are displayed, indicating different frequency characteristics. Here,

the horizontal axis refers to the time domain, and the vertical axis corresponds to the frequency domain. The case of the speech signal f_0 presents a mixture of harmonic and transitory effects originated by vocal and consonants features of a speech signal. In the case of the castanets signal f_1, a sequence of transients are displayed indicating the complex frequency behavior of the rapid series of clicks.

The spectrograms $|Vf_0|$ and $|Vf_1|$ are then fed to a persistent homology algorithm by considering its level sets as indicated in Definitions 3 and 4. We use a Morse-theory based algorithm that analyzes a quantized version of an input function, and feeds the resulting level sets to an efficient persistent homology implementation, see [13]. In our persistent diagrams of Fig. 1(e) and Fig. 1(f), we have selected only the 30 most prominent 1-dimensional homological structures, displayed by the 30 dots with the largest distance to the diagonal in the persistent diagram. We are therefore not considering topological unstable (noisy) components represented by dots, or 1-homology features, located in closer regions to the diagonal in Figures 1(e) and 1(f). These persistent diagrams are homological fingerprints characterizing the shape of the corresponding spectrograms. Notice that these homological structures are clearly identifying and discriminating these spectrograms using a limited set of homological components. This description represents a new type of topological characterization of time-frequency data.

As indicated in Theorem 2, the persistent diagram $\mathrm{dgm}(|Vf|)$ has the crucial property to be robust with respect to perturbations of the signal f. This important feature can be used to illustrate the discriminative power of persistent homology by studying the distances between persistent diagrams $\mathrm{dgm}(|Vf_0|)$ and $\mathrm{dgm}(|Vf_t|)$, for $t \in [0,1]$. In Fig. 2, we display the function $d(t) := d_H(\mathrm{dgm}(|Vf_0|), \mathrm{dgm}(|Vf_t|))$ using the Hausdorff distance, whose implementation is simpler and it does not interfere with the stabilty property, due to the inequality $d_H(X,Y) \leq d_B(X,Y)$ (see Definition 5). Notice that when the parameter t increases from 0 to 1, the Hausdorff distance between $\mathrm{dgm}(|Vf_0|)$ and $\mathrm{dgm}(|Vf_t|)$ increases, which indeed resonates with the discriminative properties of persistent homology in the setting of frame analysis.

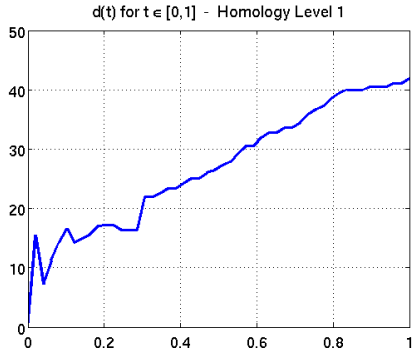


Fig. 2. $d(t) := d_H(\mathrm{dgm}(|Vf_0|), \mathrm{dgm}(|Vf_t|)),\ t \in [0,1]$

ACKNOWLEDGMENT

This work is supported by the DFG Research Center Matheon Project B26: Information Extracting Sensor Networks. G. Kutyniok acknowledges support by the Einstein Foundation Berlin, by Deutsche Forschungsgemeinschaft (DFG) Grant SPP- 1324 KU 1446/13 and DFG Grant KU 1446/14, by the DFG Collaborative Research Center TRR 109 “Discretization in Geometry and Dynamics”, and by the DFG Research Center Matheon “Mathematics for key technologies” in Berlin, Germany.

REFERENCES

[1] G. Carlsson. Topology and data. *Bull. Amer. Math. Soc.*, 46(2):255–308, 2009.

[2] P.G. Casazza and G. Kutyniok. *Finite frames: theory and applications*. Birkhäuser, 2012.

[3] F. Chazal, D. Cohen-Steiner, M. Glisse, L.J. Guibas, and S.Y. Oudot. Proximity of persistence modules and their diagrams. In *ACM Symposium on computational Geometry*, pages 237–246, 2009.

[4] F. Chazal, V. de Silva, M. Glisse, and S. Oudot. The structure and stability of persistence modules. *arXiv:1207.3674*, 2012.

[5] D. Cohen-Steiner, H. Edelsbrunner, and J. Harer. Stability of Persistence Diagrams. *Discrete Comput. Geom.*, 37(1):103–120, 2007.

[6] H. Edelsbrunner and J. Harer. Persistent homology - a survey. In *Surveys on discrete and computational geometry: twenty years later: AMS-IMS-SIAM Joint Summer Research Conference, June 18-22, 2006, Snowbird, Utah*, volume 453, page 257. American Mathematical Society, 2008.

[7] H. Edelsbrunner and J. L Harer. *Computational topology: an introduction*. American Mathematical Soc., 2010.

[8] H. G. Feichtinger and K. Gröchenig. Gabor wavelets and the Heisenberg group: Gabor expansions and short time Fourier transform from the group theoretical point of view. In *Wavelets: a tutorial in theory and applications*, volume 2 of *Wavelet Anal. Appl.*, pages 359–397. Academic Press, Boston, 1992.

[9] M. Fornasier and H. Rauhut. Continuous frames, function spaces, and the discretization problem. *J. of Fourier Anal and Appl.*, 11(3):245–287, 2005.

[10] R. Ghrist and V. de Silva. Coverage in sensor networks via persistent homology. *Algebr. Geom. Topol.*, 7:339–358, 2007.

[11] R. Ghrist and V. de Silva. Homological sensor networks. *Notices Amer. Math. Soc*, 54(1):10–17, 2007.

[12] M. Guillemard and A. Iske. Signal filtering and persistent homology: an illustrative example. In *Proc. Sampling Theory and Applications (SampTA’11)*, 2011.

[13] K. Mischaikow and V. Nanda. Morse theory for filtrations and efficient computation of persistent homology. *Discrete Comput. Geom.*, to appear.

Signal Adaptive Frame Theory

Stephen D. Casey
Department of Mathematics and Statistics
American University
Washington, DC 20016 U.S.A
Email: *scasey@american.edu*

***Abstract*—The projection method is an atomic signal decomposition designed for adaptive frequency band (AFB) and ultra-wide-band (UWB) systems. The method first windows the signal and then decomposes the signal into a basis via a continuous-time inner product operation, computing the basis coefficients in parallel. The windowing systems are key, and we develop systems that have variable partitioning length, variable roll-off and variable smoothness. These include systems developed to preserve orthogonality of any orthonormal systems between adjacent blocks, and *almost orthogonal* windowing systems that are more computable/constructible than the orthogonality preserving systems. The projection method is, in effect, an adaptive Gabor system for signal analysis. The natural language to express this structure is frame theory.**

I. Introduction

Adaptive frequency band (AFB) and ultra-wide-band (UWB) systems present challenges to current methods of signal processing. Despite extensive advances, wideband problems continue to hit barriers in sampling architectures and analog-to-digital conversion (ADC). ADC signal-to-noise and distortion ratio (the effective number of resolution bits) declines with sampling rate due to timing jitter, circuit imperfections, and electronic noise. ADC performance (speed and total integrated noise) can be improved to some extent, e.g., by cooling. However, the energy cost may be significant, and this presents a major hurdle for implementation in miniaturized devices. Digital circuitry has provided dramatically enhanced DSP operation speeds, but there has not been a corresponding dramatic energy capacity increase in batteries to operate these circuits. Moore's Law for chips is slowing down, and there is no Moore's Law for batteries or ADCs.

A growing number of applications face this challenge, such as miniature and hand-held devices for communications, robotics, and micro aerial vehicles (MAVs). Very wideband sensor bandwidths are desired for dynamic spectrum access and cognitive radio, radar, and ultra-wideband systems. Multi-channel and multi-sensor systems compound the issue, such as MIMO, array processing and beamforming, multi-spectral imaging, and vision systems. All of these rely on analog sensing and a digital interface, perhaps with feedback. This motivates mixed-signal circuit designs that tightly couple the analog and digital portions, and operate with parallel reduced bandwidth paths to relax ADC requirements. The goal of such wideband integrated circuit designs is to achieve good tradeoffs in dynamic range, bandwidth, and parallelization, while maintaining low energy consumption.

From a signal processing perspective, we can approach this problem by implementing an appropriate signal decomposition in the analog portion that provides parallel outputs for integrated digital conversion and processing. This naturally leads to an architecture with windowed time segmentation and parallel analog basis expansion. In this paper we view this from the sampling theory perspective, including segmentation and window design, achieving orthogonality between segments, basis expansion and choice of basis, signal filtering, and reconstruction. Definitions and computations for the paper follow those given in Benedetto [1].

II. Windowing

We first construct smooth bounded adaptive partitions of unity, or *BAPU Systems*. These are generalizations of bounded uniform partitions of unity (*BUPU Systems*) in that they allow for signal adaptive windowing. These systems give a flexible adaptive partition of unity of variable smoothness and are useful whenever a partition of unity is used, such as in compressed sensing. The construction elements for this system are B-splines. The second type of system we develop preserves orthogonality of any orthonormal (ON) system between adjacent blocks. The construction here uses any orthonormal basis for $L^2(\mathbb{R})$ and is created by solving a Hermite interpolation problem with constraints. These ON preserving window systems were the motivation for the methods in this paper. They allow us to create a method of time-frequency analysis for a wide class of signals. The third type of system we develop uses the concept of *almost orthogonality* developed by Cotlar, Knapp and Stein. It employs our B-spline techniques to create almost orthogonal windowing systems that are more computable/constructible than the orthogonality preserving systems.

The windowing systems for the partition of unity $\{\mathbb{B}_k(t)\}$ satisfy $\sum_k \mathbb{B}_k(t) \equiv 1$. The key difference between the partition of unity systems and (ON) systems is that the second preserves orthogonality. Preserving orthogonality requires that the windowing systems $\{\mathbb{W}_k(t)\}$ satisfy $\sum_k [\mathbb{W}_k(t)]^2 \equiv 1$. The almost orthogonal systems require that there exists a δ, $0 \leq \delta \leq 1/2$ such that for all k

$$1 - \delta \leq [\mathbb{A}_k(t)]^2 + [\mathbb{A}_{k+1}(t)]^2 \leq 1 + \delta$$

for $t \in [kT, (k+1)T]$.

A. Partition of Unity Systems

The theory of B-splines gives us the tools to create smooth partition of unity systems.

Definition 1 (Bounded Adaptive Partition of Unity): A *Bounded Adaptive Partition of Unity* is a set of functions $\{\mathbb{B}_k(t)\}$ such that

$$\begin{aligned}
&(i.) \quad \mathrm{supp}(\mathbb{B}_k(t)) \subseteq [kT - r, (k+1)T + r]\,,\\
&(ii.) \quad \mathbb{B}_k(t) \equiv 1 \text{ for } t \in [kT + r, (k+1)T - r]\,,\\
&(iii.) \quad \sum_k \mathbb{B}_k(t) \equiv 1\,,\\
&(iv.) \quad \{\widehat{\mathbb{B}_k{}^{\circ}}[n]\} \in l^1\,.
\end{aligned} \tag{1}$$

Conditions $(i.), (ii.)$ and $(iii.)$ make $\{\mathbb{B}_k(t)\}$ a bounded partition of unity. Condition $(iii.)$ means that these systems do not preserve orthogonality between blocks. We will generate our systems by translations and dilations of a given window $\mathbb{B}_I$, where $\mathrm{supp}(\mathbb{B}_I) = [(-T/2 - r), (T/2 + r)]$.

Our general window function $\mathbb{W}_I$ is k-times differentiable, has $\mathrm{supp}(\mathbb{B}_I) = [(-T/2 - r), (T/2 + r)]$ and has values

$$\mathbb{B}_I = \begin{cases} 0 & |t| \geq T/2 + r \\ 1 & |t| \leq T/2 - r \\ \rho(\pm t) & T/2 - r < |t| < T/2 + r \end{cases} \tag{2}$$

We solve for $\rho(t)$ by solving the Hermite interpolation problem

$$\begin{cases} (a.) & \rho(T/2 - r) = 1 \\ (b.) & \rho^{(n)}(T/2 - r) = 0\,,\ n = 1, 2, \ldots, k \\ (c.) & \rho^{(n)}(T/2 + r) = 0\,,\ n = 0, 1, 2, \ldots, k\,, \end{cases}$$

with the conditions that $\rho \in C^k$ and

$$[\rho(t)] + [\rho(-t)] = 1 \text{ for } t \in [T/2 - r, T/2 + r]\,. \tag{3}$$

We use B-splines as our cardinal functions. Let $0 < \alpha \ll \beta$ and consider $\chi_{[-\alpha,\alpha]}$. We want the n-fold convolution of $\chi_{[\alpha,\alpha]}$ to fit in the interval $[-\beta, \beta]$. Then we choose α so that $0 < n\alpha < \beta$ and let

$$\Psi(t) = \underbrace{\chi_{[-\alpha,\alpha]} * \chi_{[-\alpha,\alpha]} * \cdots * \chi_{[-\alpha,\alpha]}(t)}_{n-times}\,.$$

The β-periodic continuation of this function, $\Psi^{\circ}(t)$ has the Fourier series expansion

$$\sum_{k \neq 0} \frac{\alpha}{n\beta} \left[\frac{\sin(\pi k\alpha/n\beta)}{2\pi k\alpha/n\beta}\right]^n \exp(\pi i k t/\beta)\,.$$

The C^k solution for ρ is given by a theorem of Schoenberg (see [7], pp. 7-8). Schoenberg solved the Hermite interpolation problem

$$\begin{cases} (a.) & S^{(n)}(-1) = 0\,,\ n = 0, 1, 2, \ldots, k\,, \\ (b.) & S(1) = 1\,, \\ (b.) & S^{(n)}(1) = 0\,,\ n = 1, 2, \ldots, k\,. \end{cases}$$

An interpolant that minimizes the Chebyshev norm is called the *perfect spline*. The perfect spline $S(t)$ for Hermite problem above is given by the integral of the function

$$M(x) = (-1)^n \sum_{j=0}^{k} \frac{\Psi(t - t_j)}{\phi'(t_j)}\,,$$

where Ψ is the $(k+1)$ convolution of characteristic functions, the knot points are $t_j = -\cos(\frac{\pi j}{k})$ and $\phi(t) \prod_{j=0}^{k}(t - t_j)$. We then have that $\rho(t) = S \circ \ell(t)\,,$ where $\ell(t) = \frac{1}{r}t - \frac{2T}{2r}$. For this ρ, and for

$$\mathbb{B}_I = \begin{cases} 0 & |t| \geq T/2 + r \\ 1 & |t| \leq T/2 - r \\ \rho(\pm t) & T/2 - r < |t| < T/2 + r \end{cases}$$

we have that $\widehat{\mathbb{B}_I}(\omega)$ is given by the antiderivative of a linear combination of functions of the form $[\sin(\omega)/\omega]^{k+1}$, and therefore has decay $1/\omega^{k+2}$ in frequency.

B. Orthogonality Preserving Systems

Our first system of signal segmentation uses sine, cosine and linear functions. This was created because it is relatively easy to implement, cuts down on frequency error and preserves orthogonality. Consider a signal block of length $T + 2r$ centered at the origin. Let $0 < r \ll T$. Ideally, we would like to make r as small as possible. Define $\mathrm{Cap}(t)$ as follows.

$$\begin{cases} 0 & |t| \geq \frac{T}{2} + r\,, \\ 1 & |t| \leq \frac{T}{2} - r\,, \\ \sin(\pi/(4r)(t + (T/2 + r))) & \frac{-T}{2} - r < t < \frac{-T}{2} + r\,, \\ \cos(\pi/(4r)(t - (T/2 - r))) & \frac{T}{2} - r < t < \frac{T}{2} + r\,. \end{cases} \tag{4}$$

Given Cap, we form a tiling system $\{\mathrm{Cap}_k(t)\}$ such that $\mathrm{supp}(\mathrm{Cap}_k(t)) \subseteq [kT - r, (k+1)T + r]$ for all k. Note that the Cap window has several properties that make it a good window for our purposes. It has a partition property in that it windows the signal in $[\frac{-T}{2} - r, \frac{T}{2} + r]$ and is identically 1 on $[\frac{-T}{2} + r, \frac{T}{2} - r]$. It has a continuous roll-off at the endpoints. Finally, it has the property that for all $t \in \mathbb{R}$

$$[\mathrm{Cap}_k(t)]^2 + [\mathrm{Cap}_{k+1}(t)]^2 = 1\,.$$

This last condition is needed to preserve the orthogonality of basis elements between adjacent blocks. Additionally, it has $1/\omega^2$ decay in frequency space, and, when one time block is ramping down, the adjacent block is ramping up at exactly the same rate. If we had a signal f with an absolutely convergent Fourier series,

$$(f \cdot \mathrm{Cap})_k\widehat{\ }[n] = \sum_m f[n - m]\mathrm{Cap}\widehat{\ }[m] = \widehat{f} * \mathrm{Cap}\widehat{\ }[n]\,.$$

The Fourier transform of Cap is a linear combination of $\mathrm{sinc}(\omega)$ and $\sin(\omega)$ functions and has an asymptotic $1/\omega^2$ decay.

The theory of splines gives us the tools to generalize this system. The idea is to cut up the time domain into perfectly aligned segments so that there is no loss of information.

We also want the systems to be smooth, so as to provide control over decay in frequency, and adaptive, so as to adjust accordingly to changes in frequency band. Finally, we develop our systems so that the orthogonality of bases in adjacent and possible overlapping blocks is preserved.

Definition 2 (ON Window System): An *ON Window System* is a set of functions $\{\mathbb{W}_k(t)\}$ such that for all $k \in \mathbb{Z}$

$$\begin{array}{ll} (i.) & \text{supp}(\mathbb{W}_k(t)) \subseteq [kT - r, (k+1)T + r]\,, \\ (ii.) & \mathbb{W}_k(t) \equiv 1 \text{ for } t \in [kT + r, (k+1)T - r]\,, \\ (iii.) & \mathbb{W}_k \text{ is symmetric about its midpoint }\,, \\ (iv.) & [\mathbb{W}_k(t)]^2 + [\mathbb{W}_{k+1}(t)]^2 = 1\,, \\ (v.) & \{\widehat{\mathbb{W}_k}^{\circ}[n]\} \in l^1\,. \end{array} \tag{5}$$

Conditions $(i.)$ and $(ii.)$ are partition properties, in that they give an exact snapshot of the input function f on $[kT + r, (k+1)T - r]$ with smooth roll-off at the edges. Conditions $(iii.)$ and $(iv.)$ are needed to preserve orthogonality between adjacent blocks. Condition $(v.)$ is needed for the computation of Fourier coefficients. We generate our systems by translations and dilations of a given window $\mathbb{W}_I$, where $\text{supp}(\mathbb{W}_I) = [-T/2 - r, T/2 + r]$. Our next proposition shows the need for the condition $(v.)$. Let $I = T + 2r$ and let $\mathbb{PW}_\Omega$ denote the Paley-Wiener space for bandlimit Ω.

Proposition 1: Let $f \in \mathbb{PW}_\Omega$ and let $\{\mathbb{W}_k(t)\}$ be an ON Window System with generating window $\mathbb{W}_I$. Then

$$\frac{1}{I}\int_{-T/2-r}^{T/2-r} [f \cdot \mathbb{W}_I]^{\circ}(t) \exp(-2\pi i n t/[I])\, dt = \widehat{f} * \widehat{\mathbb{W}_I}[n]\,. \tag{6}$$

Our general window function $\mathbb{W}_I$ is k-times differentiable, has $\text{supp}(\mathbb{W}_I) = [-T/2 - r, T/2 + r]$ and has values

$$\mathbb{W}_I = \begin{cases} 0 & |t| \geq T/2 + r \\ 1 & |t| \leq T/2 - r \\ \rho(\pm t) & T/2 - r < |t| < T/2 + r \end{cases} \tag{7}$$

We solve for $\rho(t)$ by solving the Hermite interpolation problem

$$\begin{cases} (a.) & \rho(T/2 - r) = 1 \\ (b.) & \rho^{(n)}(T/2 - r) = 0\,,\ n = 1, 2, \ldots, k \\ (c.) & \rho^{(n)}(T/2 + r) = 0\,,\ n = 0, 1, 2, \ldots, k\,, \end{cases}$$

with the conditions that $\rho \in C^k$ and

$$[\rho(t)]^2 + [\rho(-t)]^2 = 1 \text{ for } t \in [\pm(\frac{T}{2} - r), \pm(\frac{T}{2} + r)]\,. \tag{8}$$

The constraint (8) directs us to get solutions expressed in terms of $\sin(t)$ and $\cos(t)$. Solving for ρ so that the window in C^1, we get that $\rho(t)$ equals

$$\begin{cases} \sqrt{\left[1 - \frac{1}{2}\left[1 - \sin(\frac{\pi}{2r}(\frac{T}{2} - t))\right]^2\right]} & \frac{T}{2} - r \leq t \leq \frac{T}{2} \\ \frac{1}{\sqrt{2}}\left[1 - \sin(\frac{\pi}{2r}(t - \frac{T}{2}))\right] & \frac{T}{2} \leq t \leq \frac{T}{2} + r\,. \end{cases} \tag{9}$$

With each degree of smoothness, we get an additional degree of decay in frequency.

C. Orthogonality Between Blocks

We designed the ON Window Systems $\{\mathbb{W}_k(t)\}$ so that they would preserve orthogonality of basis element of overlapping blocks. Because of the partition properties of these systems, we need only check orthogonality of adjacent overlapping blocks. The best way to think about the construction is to visualize how one would do the extension for a system of sines and cosines. We would extend the odd reflections about the left endpoint and the even reflections about the right. Let $\{\varphi_j(t)\}$ be an orthonormal basis for $L^2[-T/2, T/2]$. Define

$$\widetilde{\varphi_j}(t) \;=\; \begin{cases} 0 & |t| \geq T/2 + r \\ \varphi_j(t) & |t| \leq T/2 - r \\ -\varphi_j(-T - t) & -T/2 - r < t < -T/2 \\ \varphi_j(T - t) & T/2 < t < T/2 + r\,. \end{cases} \tag{10}$$

Theorem 1: $\{\Psi_{k,j}\} = \{\mathbb{W}_k \widetilde{\varphi_j}(t)\}$ is an ON basis for $L^2(\mathbb{R})$.

Proof : See [3]. □

D. Almost Orthogonal Systems

The Partition of Unity Systems do *not* preserve orthogonality between blocks. However, they are easier to compute in both time and frequency. Therefore, these systems can be used to approximate the Cap system with B-splines. We get windowing systems that nearly preserve orthogonality. Each added degree of smoothness in time adds to the degree of decay in frequency.

Cotlar, Knapp and Stein introduced *almost orthogonality* via operator inequalities. The concept allows us to create windowing systems that are more computable/constructible such as the Bounded Adaptive Partition of Unity Systems $\{\mathbb{B}_k(t)\}$ with the orthogonality preservation of the ON Window Systems $\{\mathbb{W}_k(t)\}$.

Definition 3 (Almost ON System): Let $0 < r \ll T$. An **Almost ON System** for adaptive and ultra-wide band sampling is a set of functions $\{\mathbb{A}_k(t)\}$ for which there exists δ, $0 \leq \delta < 1/2$, such that

$$\begin{array}{ll} (i.) & \text{supp}(\mathbb{A}_k(t)) \subseteq [kT - r, (k+1)T + r]\,, \\ (ii.) & \mathbb{A}_k(t) \equiv 1 \text{ for } t \in [kT + r, (k+1)T - r]\,, \\ (iii.) & \mathbb{A}_k((kT + T/2) - t) = \mathbb{A}_k(t - (kT + T/2))\,, \\ (iv.) & 1 - \delta \leq [\mathbb{A}_k(t))]^2 + [\mathbb{A}_{k+1}(t))]^2 \leq 1 + \delta\,, \\ (v.) & \{\widehat{\mathbb{A}_k}^{\circ}[n]\} \in l^1\,. \end{array}$$

Starting with $\text{Cap}(t)$, let $\Delta_{(T,r)} = \frac{T+2r}{m}$. By placing equidistant knot points $-T/2 - r = x_0, -T/2 - r + \Delta_{(T,r)} = x_1, \ldots, T/2 + r = x_m$, we can construct C^{m-1} polynomial splines S_{m+1} approximating $\text{Cap}(t)$ in $[(-T/2 - r), (T/2 + r)]$. A theorem of Curry and Schoenberg gives that the set of B-splines $\{B^{(m+1)}_{-(m+1)}, \ldots, B^{(m+1)}_k\}$ forms a basis for S_{m+1}. Therefore, $\text{Cap}(t) \approx \sum_{i=-(m+1)}^{k} a_i B_i^{(m+1)}$. Let

$$\delta = \left\| \sum_{i=-(m+1)}^{k} a_i B_i^{(m+1)} - \text{Cap}(t) \right\|_\infty .$$

Then, $\delta < 1/2$, with the largest value for the piecewise linear spline approximation. Moreover, $\delta \longrightarrow 0$ as m and k increase. Thus we get computable windowing systems that nearly preserve orthogonality. Each added degree of smoothness in time adds to the degree of decay in frequency.

III. Signal Expansions

Given characteristics of the class of input signals, the choice of basis functions used can be tailored to optimal representation of the signal or a desired characteristic in the signal.

Theorem 2 (The Projection Formula for ON Windowing): Let $\{\mathbb{W}_k(t)\}$ be an ON Window System, and let $\{\Psi_{k,j}\}$ be an orthonormal basis that preserves orthogonality between adjacent windows. Let $f \in \mathbb{PW}_\Omega$ and $N = N(T,\Omega)$ be such that $\langle f \cdot \mathbb{W}_k, \Psi_{k,n}\rangle = 0$ for all $n > N$ and all k. Then, $f(t) \approx f_{\mathcal{P}}(t)$, where

$$f_{\mathcal{P}}(t) = \sum_{k\in\mathbb{Z}}\left[\sum_{n=-N}^{N} \langle f \cdot \mathbb{W}_k, \Psi_{k,n}\rangle \Psi_{k,n}(t)\right]. \qquad (11)$$

This theorem gives a new method for A-D conversion. Unlike the Shannon method which examined the function at specific points, then used those individual points to recreate the curve, the projection method breaks the signal into time blocks and then approximates their respective periodic expansions with a Fourier series. This process allows the system to individually evaluate each piece and base its calculation on the needed bandwidth. The individual Fourier series are then summed, recreating a close approximation of the original signal. It is important to note that instead of fixing T, the method allows us to fix any of the three while allowing the other two to fluctuate. From the design point of view, the easiest and most practical parameter to fix is N. For situations in which the bandwidth does not need flexibility, it is possible to fix Ω and T by the equation $N = \lceil T \cdot \Omega \rceil$. However, if greater bandwidth Ω is need, choose shorter time blocks T.

The windowing systems above allow us to develop *Signal Adaptive Frame Theory*. The idea is as follows. If we work with an ON Windowing System $\{\mathbb{W}_k(t)\}$, let $\{\Psi_{k,j}\}$ be an orthonormal basis that preserves orthogonality between adjacent windows. Let $f \in \mathbb{PW}_\Omega$ and $N = N(T,\Omega)$ be such that $\langle f \cdot \mathbb{W}_k, \Psi_{k,n}\rangle = 0$ for all $n > N$ and all k. Then

$$f(t) = \sum_{k\in\mathbb{Z}}\left[\sum_{n\in\mathbb{Z}} \langle f \cdot \mathbb{W}_k, \Psi_{k,n}\rangle \Psi_{k,n}(t)\right]. \qquad (12)$$

This also gives

$$\|f\|^2 = \sum_{k\in\mathbb{Z}}\left[\sum_{n\in\mathbb{Z}} |\langle f \cdot \mathbb{W}_k, \Psi_{k,n}\rangle|^2\right]. \qquad (13)$$

Given that $\{\Psi_{k,j}\} = \{\mathbb{W}_k\widetilde{\varphi_j}(t)\}$ is an orthonormal basis for $L^2(\mathbb{R})$, we have a representation of a given function f in L^2. The set $\{\Psi_{k,j}\} = \{\mathbb{W}_k\widetilde{\varphi_j}(t)\}$ is an exact normalized tight frame for L^2. The restriction that these basis elements present is computability. They become increasing difficult to compute as the smoothness in time/decay in frequency increases.

A way around this is to connect the Bounded Adaptive Partition of Unity Systems $\{\mathbb{B}_k(t)\}$ to frame theory. The ideas behind this connection go back to the curvelet work of Candès and Donoho. The paper of Borup and Neilsen [2] gives a nice overview of this connection, and we will refer to that paper for the background from which we develop our approach. The set $\{\mathbb{B}_k(t)\}$ form an *admissible* cover, in that they form a partition of unity and have overlap with only their immediate neighbors.

For each window $\mathbb{B}_k(t)$, let $\phi_{n,k}(t)$ be the shifted $\exp[\pi i t T/n]$ centered in the window. Then define

$$\Phi_{k,n} = \mathbb{B}_k(t)\phi_{k,n}(t)\,.$$

Given and $f \in L^2$ we can write

$$f(t) \approx \sum_{k\in\mathbb{Z}}\left[\sum_{n\in\mathbb{Z}} \langle f \cdot \mathbb{B}_k, \Phi_{k,n}\rangle \Phi_{k,n}(t)\right]. \qquad (14)$$

For this system we can compute

$$A\|f\| \le \sum_{k\in\mathbb{Z}}\left[\sum_{n\in\mathbb{Z}} |\langle f \cdot \mathbb{B}_k, \Phi_{k,n}\rangle|^2\right] \le B\|f\|\,. \qquad (15)$$

The bounds are a function of how much of the signal is concentrated in the overlap regions and will be tightened for the almost orthogonal windowing systems. The closer the approximation, the better the frame bounds. Developing these signal adaptive frames, their bounds and the associated frame operators will be a major point of emphasis in future work. We will additionally develop biorthogonal adaptive frames using our B-spline constructions. We conjecture the following:

$$\mathcal{A}_{1-\delta}\|f\|^2 \le \sum_{k\in\mathbb{Z}}\left[\sum_{n\in\mathbb{Z}} |\langle f \cdot \mathbb{A}_k, \Psi_{k,n}\rangle|^2\right] \le \mathcal{A}_{1+\delta}\|f\|^2\,. \qquad (16)$$

Moreover, this $\longrightarrow$ Normalized Tight Frame as $\delta \longrightarrow 0$.

Acknowledgment

The author's research was partially supported by U. S. Army Research Office Scientific Services program, administered by Battelle (TCN 06150, Contract DAAD19-02-D-0001) and U. S. Air Force Office of Scientific Research Grant Number FA9550-12-1-0430.

References

[1] J. J. Benedetto, *Harmonic Analysis and Applications*, CRC Press, Boca Raton, FL, 1997.

[2] L. Borup and M. Neilsen, "Frame Decomposition of Decomposition Spaces" *Journal of Fourier Analysis and Applications* **13** (1), 39-70, 2007.

[3] S. D. Casey, "Windowing systems for time-frequency analysis – to appear in *STSIP SampTA 2011 Special Issue*, 31 pp., 2013.

[4] S. D. Casey, S. Hoyos, and B. M. Sadler, "Adaptive and ultra-wideband sampling via signal segmentation and projection," to be submitted to *Proc. IEEE*, 24 pp., 2013.

[5] R. Coifman and Y. Meyer, "Remarques sur l'analyse de Fourier a fenetre." *CR Acad. Sci. Paris* 312, 259-261, 1991.

[6] H. S. Malvar, "Biorthogonal and nonuniform lapped transforms for transform coding with reduced blocking and ringing artifacts," *IEEE Trans. Signal Process.*, **46** (4) 1043-1053, 1998.

[7] I. J. Schoenberg, *Cardinal Spline Interpolation* (CBMS–NSF Conference Series in Applied Mathematics, 12), SIAM, Philadelphia, PA, 1973.

Identification of Rational Transfer Functions from Sampled Data

Hagai Kirshner, John Paul Ward, Michael Unser
École polytechnique fédérale de Lausanne (EPFL), Lausanne, Switzerland

***Abstract*—We consider the task of estimating an operator from sampled data. The operator, which is described by a rational transfer function, is applied to continuous-time white noise and the resulting continuous-time process is sampled uniformly. The main question we are addressing is whether the stochastic properties of the time series that originates from the sample values of the process allows one to determine the operator. We focus on the autocorrelation property of the process and identify cases for which the sampling operator is injective. Our approach relies on sampling properties of almost periodic functions, which together with exponentially decaying functions, provide the building blocks of the autocorrelation measure. Our results indicate that it is possible, in principle, to estimate the parameters of the rational transfer function from sampled data, even in the presence of prominent aliasing.**

I. Introduction

Models that are based on stochastic differential equations are widely used for describing numerous physical phenomena. We consider in this work stochastic differential equations that have constant coefficients. Such equations are characterized by a rational transfer function and are equivalent to the filtering of white noise. In practice, the available data is discrete, and one is often required to estimate continuous-time parameters from sampled data. The stochastic properties of the time series that originates from such processes depend on the constant coefficients, and the question we are raising here is whether the sampling process is injective in the sense that there is a one-to-one mapping between the continuous-time and the discrete-time models.

Within the context of state-space autoregressive representation, it is known that stochastic differential equations are mapped to stochastic difference equations upon sampling. The z transform description of the difference equation is based on the exponential values of the poles (the roots of the rational transfer function); and for that reason, currently available estimation algorithms assume that there is an ambiguity in determining their imaginary part value, as the exponential function is invariant to $2\pi i$ increments in its argument. In order to overcome this ambiguity, current estimation approaches require high sampling rate values for avoiding aliasing [3]–[10]. They also restrict the imaginary part of the poles to be less than π/T where T is the sampling interval. The z transform description, however, includes non-exponential terms as well, and this fact has not been taken into account so far.

We revisit in this work the ambiguity assumption of sampled autoregressive continuous-time processes, and identify cases for which the sampling operator is injective when applied to the autocorrelation function. We will show that there is no ambiguity even in the presence of prominent aliasing. To this aim, we introduce two alternative descriptions for the poles of the model: one is used for deriving an explicit expression for the autocorrelation function, while the other is used for assigning a Lebesgue measure to subsets of poles. The building blocks of the autocorrelation function are exponentially decaying terms and almost periodic functions; and we exploit this structure for proving uniqueness of the sampled model.

II. The Problem

We consider the following stochastic process

$$x(t) = \int_0^\infty h(t-\tau;\theta)\, w(\tau)\, d\tau, \tag{1}$$

where $w(t)$ is a Gaussian or non-Gaussian white noise process. The shaping filter $h(t;\theta)$ is given in the Fourier domain by

$$H(\omega;\theta) = \frac{1}{\prod_{n=1}^{p}(i\omega - s_n)}, \tag{2}$$

where $\theta = (s_1, s_2, \ldots, s_p) \in \mathbb{C}^p$ is composed of the poles of $H(\omega;\theta)$. The real part of each pole is strictly negative and complex poles appear in conjugate pairs. Assuming that $w(t)$ is white with finite variance, σ^2, and that $t \gg 0$, the autocorrelation function of $x(t)$ is given in the Fourier domain by $\Phi(\omega;\theta,\sigma^2) = \sigma^2\left|H(\omega;\theta)\right|^2$. In this work we investigate the injective property of the sampling operator $x(t) \to \{x(n)\}$ while assuming that p is known. Specifically, we raise the following question: does the time series that originates from the sampled version of $x(t)$ allow one to recover θ?

III. Asymptotic properties of the autocorrelation function

A. Alternative representations to $H(\omega;\theta)$

We introduce two alternative parameter vectors, $\tilde{\theta}$ and $\bar{\theta}$, that will be used for deriving an explicit formula for the autocorrelation function $\varphi(t;\theta)$, and for associating subsets of θ with a measure in $\mathbb{R}^p$. Let $\theta = (s_1, \ldots, s_{2m}, s_{2m+1}, \ldots, s_p)$ where the first $2m$ poles are complex, and conjugate pairs appear sequentially. Additionally, for a given complex pair, we require the one with positive imaginary part to be listed first. Our first alternative representation is based on decay rates and modulation values. It extends the representation of [11] in the following manner,

$$\tilde{\theta} = (a_1, b_1, a_2, b_2, \ldots, a_m, b_m, a_{m+1}, \ldots, a_{p-m}), \tag{3}$$

where $a_1, \dots a_{p-m}$ are the strictly negative real parts of the poles, and $b_1, \dots b_m$ are the strictly positive imaginary parts. The vector $\tilde{\theta}$ is a point in $\mathbb{R}^p$ and this identification can be made unique by imposing a dictionary-type ordering:

- $0 > a_1 \geq a_2 \geq \cdots \geq a_m$;
- $0 > a_{m+1} \geq a_{m+2} \geq \cdots \geq a_{p-m}$;
- if $a_k = a_{k+1}$, then $b_{k+1} \geq b_k$.

The difference in sign between the a's and b's allows us to distinguish the two types of poles, so that there is no confusion.

The second alternative parameter vector $\bar{\theta}$ indicates multiplicities of poles and will be used for obtaining an explicit formula of autocorrelation functions

$$\bar{\theta} = (\bar{s}_1, m_1, \bar{s}_2, m_2, \dots, \bar{s}_L, m_L). \tag{4}$$

The multiplicity of a pole $\bar{s}_l$ is represented by m_l.

Definition 1. *The collection of all parameter vectors θ is $\Omega(p)$. This is also the collection of all parameter vectors $\tilde{\theta}$ or $\bar{\theta}$.*

B. The autocorrelation function

The rational form of $H(\omega;\theta)$ is known to yield an autocorrelation function that is a sum of Hermitian symmetric exponentials, as the result of a decomposition in partial fractions [1]. The explicit formula is obtained as follows.

Proposition 1. *Let $\bar{\theta} = (\bar{s}_1, m_1 \dots, \bar{s}_L, m_L) \in \Omega(p)$. Then,*

$$\varphi(t;\bar{\theta}) = (-1)^p \sum_{\ell=1}^{L} e^{-\lambda_l^{1/2}|t|} \sum_{n=1}^{m_l} \sum_{k=0}^{n-1} d_{l,n,k} \, |t|^{n-1-k}, \tag{5}$$

where

$$\lambda_l = \bar{s}_l^2 \tag{6}$$

$$P(\xi) = \prod_{l=1}^{L} (\xi - \lambda_l)^{m_l} \tag{7}$$

$$c_{l,n} = \lim_{\xi \to \lambda_l} \frac{1}{(m_l - n)!} \frac{\mathrm{d}^{m_l - n}}{\mathrm{d}\xi^{m_l - n}} \left(\frac{(\xi - \lambda_l)^{m_l}}{P(\xi)} \right) \tag{8}$$

$$d_{l,n,k} = \frac{(-1)^n c_{l,n} (n-1+k)!}{(n-1)! k! (n-1-k)! (2\lambda_l^{1/2})^{n+k}}, \tag{9}$$

and $\lambda_l^{1/2} \in \mathbb{C}$ denotes the principal square root of λ_l.

Definition 2. *Two parameter vectors $\theta_1, \theta_2 \in \Omega(p)$ are equivalent if there exists $\alpha \in \mathbb{R}$ such that $\varphi(n;\theta_1) + \alpha \cdot \varphi(n;\theta_2) = 0$ for all $n \in \mathbb{Z}$. If θ_1 is not equivalent to any distinct θ_2, then it is unique.*

When the uniqueness property holds, there is a one-to-one mapping between the autocorrelation function and its sampled version. The sample value of the autocorrelation function can then be estimated from the available sample values of $x(t)$. The uniqueness property is related to linear combinations of autocorrelation functions. In (5), the real parts of the parameters $\lambda_l^{1/2}$ determine exponentially decaying terms, while the imaginary parts determine periods of trigonometric polynomials. In the case of multiple poles, a polynomial term multiplies the complex exponential. Linear combinations of these functions also have the same basic structure. We generalise this structure in the following definition.

Definition 3. *We denote by X the class of functions of the form*

$$\sum_{l=1}^{L} \sum_{m=0}^{M_l} T_{l,m}(|t|) \, |t|^m \, e^{a_l |t|} \tag{10}$$

where $0 > a_1 > a_2 > \cdots > a_L$ and each $T_{l,m}$ is a trigonometric function.

We note that $T_{l,m}(t) \in \mathrm{AP}(\mathbb{R}, \mathbb{R})$, which is the space of almost periodic functions. Of particular interest is the fact that uniform samples of almost periodic functions lie in the normed space of almost periodic sequences $\mathrm{AP}(\mathbb{Z}, \mathbb{R})$ (cf. [2, Proposition 3.35]), and we shall exploit this fact to verify uniqueness.

Definition 4. [2, pp.94-95] *For any integer n, the mean value of $f \in \mathrm{AP}(\mathbb{Z}, \mathbb{R})$ is*

$$M(f) = \lim_{k \to \infty} \frac{f(n+1) + f(n+2) + \cdots + f(n+k)}{k}. \tag{11}$$

Note that we are free to choose any integer n; however, the limit is independent of this choice. A norm for $\mathrm{AP}(\mathbb{Z}, \mathbb{R})$ is given by

$$\|f\|^2_{\mathrm{AP}(\mathbb{Z},\mathbb{R})} = M\left(|f|^2\right). \tag{12}$$

Theorem 1. *If $f \in X$ and $f(n) = 0$ for all integers n, then the functions $T_{l,m}$ must also satisfy $T_{l,m}(n) = 0$.*

The value of Theorem 1 is that it essentially allows us to compare functions from X in a segmented fashion, i.e. according to decay rates. For example, suppose $\varphi(t;\theta)$ contains a term $T(|t|)\,|t|^m\, e^{a|t|}$, where $T(|n|)\,|n|^m$ is not identically 0. Then it can not be equivalent to any autocorrelation function that lacks a term with similar decay. We shall use this result to show that the uniform sampling operator is injective for large sub-collections of $\Omega(p)$.

IV. Uniqueness properties

We consider two subsets of $\Omega(\mathrm{p})$: $H(\omega;\theta)$ is composed of real poles only; and real and imaginary poles with minimal restrictions.

Lemma 1. *The elements of $\Omega(p)$ that are composed entirely of real poles are unique.*

Definition 5. *Let $\Omega(p)^*$ be the collection of parameter vectors $\tilde{\theta}$ satisfying:*

- *$a_{k_1} \neq a_{k_2}$ for $k_1 \neq k_2$;*
- *each b_k is an irrational multiple of π.*

Proposition 2. *As a subset of $\mathbb{R}^p$, the complement of $\Omega(p)^*$ in $\Omega(p)$ has Lebesgue measure 0.*

Proposition 3. *If an admissible vector $\tilde{\theta}_1 \in \Omega(p)^*$ is equivalent to a vector $\tilde{\theta}_2 \in \Omega(p)$, then $\tilde{\theta}_2$ must have the same number of complex pairs of poles as $\tilde{\theta}_1$. Furthermore, the complex pairs should exist at the same decay rates.*

Proposition 4. *Suppose*

$$\tilde{\theta}_1 = (a_1, b_1, \ldots, a_m, b_m, a_{m+1}, \ldots, a_{p-m}) \in \Omega(p)^* \quad (13)$$

is equivalent to

$$\tilde{\theta}_2 = (a_1, \beta_1, \ldots, a_m, \beta_m, \alpha_{m+1}, \ldots, \alpha_{p-m}) \in \Omega(p). \quad (14)$$

Then $a_l = \alpha_l$ for all l.

Proposition 4 implies that any vector of parameters that is equivalent to a vector of parameters in $\Omega(p)^*$ has the same real part values. The autocorrelation function in such a case is given by

$$\varphi(t;\tilde{\theta}_1) = (-1)^p \sum_{l=1}^{m} 2e^{a_l|t|}\big(\Re(\gamma_l)\cos(b_l|t|) - \Im(\gamma_l)\sin(b_l|t|)\big) + \sum_{l=m+1}^{p-m} \gamma_l e^{a_l|t|} \quad (15)$$

where for $l \le m$

$$\gamma_l := \left[-8i(a_l+ib_l)a_l b_l \prod_{l'\neq l, l'\le m} \big((a_l+ib_l)^2 - (a_{l'}+ib_{l'})^2\big) \big((a_l+ib_l)^2 - (a_{l'}-ib_{l'})^2\big) \prod_{l'>m} \big((a_l+ib_l)^2 - a_{l'}^2\big)\right]^{-1}$$

and for $l > m$

$$\gamma_l := \left[-2a_l \prod_{l'\le m} \big(a_l^2 - (a_{l'}+ib_{l'})^2\big)\big(a_l^2 - (a_{l'}-ib_{l'})^2\big) \prod_{l'\neq l, l'>m} (a_l^2 - a_{l'}^2)\right]^{-1}.$$

Theorem 2. *Let f_1 and f_2 be functions of the form*

$$f_1(t) = \gamma_1\cos(b|t|) + \gamma_2\sin(b|t|) \quad (16)$$

$$f_2(t) = \gamma_3\cos(\beta|t|) + \gamma_4\sin(\beta|t|), \quad (17)$$

where $\gamma_1, \gamma_2, \gamma_3, \gamma_4$ are non-zero real numbers, b, β are positive real numbers, and b is an irrational multiple of π. If $f_1(n) = f_2(n)$ for all non-negative integers n, then $\gamma_1 = \gamma_3$, $\gamma_2 = \pm\gamma_4$, and $b = \beta + 2\pi k$ for some integer k.

Lemma 2. *Let $\tilde{\theta}_1 \in \Omega(p)^*$ be of the form* (13) *with corresponding autocorrelation function $\varphi(t;\tilde{\theta}_1)$ as defined in* (15)*. Let $\tilde{\theta}_2 \in \Omega(p)$ be of the form* (13)*, with autocorrelation function $\varphi(t;\tilde{\theta}_2)$ as defined in* (15) *where γ_l is replaced by γ_l' and b_l is replaced by β_l. If $\tilde{\theta}_1$ is equivalent to $\tilde{\theta}_2$, then there is a positive number σ such that*

$$\sigma^2\gamma_l = \gamma_l' \quad or \quad \sigma^2\gamma_l^* = \gamma_l' \quad (18)$$

for every l.

Lemma 2 provides a practical criterion for determining uniqueness. According to the lemma, uniqueness translates into a set of polynomial equations that can be simplified by means of Gröbner basis algorithms. If the reduced Gröbner basis has only trivial solutions, then uniqueness is guaranteed. We utilized this property for obtaining the following results.

Theorem 3. [11] *Every element of $\Omega(1)$ is unique.*

Theorem 4. [11] *Every element of $\Omega(2)^*$ is unique.*

Theorem 5. *Every element of $\Omega(3)^*$ is unique.*

Finding reduced Gröbner bases for $p > 3$ is computationally demanding, and we suggest to exploit Lemma 2 for a limited number of values of k. That is, verifying uniqueness for a finite number of modulation values $b = \beta + 2\pi k$.

V. Conclusion

In this work, we investigated the injective properties of sampled continuous-time stochastic processes. We considered uniform sampling of processes with rational power spectrum and identified cases for which the sampling operator is injective when applied to the autocorrelation function. Our analysis relies on the sampling properties of almost periodic functions, which are the building blocks of the autocorrelation function of such processes. By removing a zero-measure set of vectors of parameters we derived a criterion for the uniqueness of the sampled model, and we proved the injective property of several rational operators. Our results indicate that the ambiguity assumption of sampled autoregressive models does not hold true, and that it is possible in principle to estimate the parameters of the rational operator from sampled data, even in the presence of prominent aliasing.

Acknowledgment

This work was funded in part by the ERC Grant ERC-2010-AdG 267439-FUN-SP.

References

[1] J. B. Boyling, "Green's functions for polynomials in the Laplacian," *Z. Angew. Math. Phys.*, vol. 47, no. 3, pp. 485–492, 1996.

[2] C. Corduneanu, *Almost periodic oscillations and waves*. New York: Springer, 2009.

[3] A. Feuer and G. Goodwin, *Sampling in Digital Signal Processing and Control*. Boston, MA: Birkhäuser, 1996.

[4] J. Gillberg and L. Ljung, "Frequency-domain identification of continuous-time ARMA models from sampled data," *Automatica*, vol. 45, pp. 1371–1378, 2009.

[5] R. Johansson, "Identification of continuous-time models," *IEEE Trans. signal processing*, vol. 42, no. 4, pp. 887–897, April 1994.

[6] E. K. Larsson, M. Mossberg, and T. Söderström, "An overview of important practical aspects of continuous-time ARMA system identification," *Circ. Sys. Sig. Proc.*, vol. 25, no. 1, pp. 17–46, May 2006.

[7] E. K. Larsson, "Limiting sampling results for continuous-time ARMA systems," *International Journal of Control*, vol. 78, no. 7, pp. 461–473, May 2005.

[8] D.-T. Pham, "Estimation of continuous-time autoregressive model from finely sampled data," *IEEE Trans. Sig. Proc.*, vol. 48, no. 9, pp. 2576–2584, September 2000.

[9] T. Söderström, "Computing stochastic continuous-time models from ARMA models," *International Journal of Control*, vol. 53, no. 6, pp. 1311–1326, May 1991.

[10] H. Tsai and K. S. Chan, "Maximum likelihood estimation of linear continuous time long memory processes with discrete time data," *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, vol. 67, no. 5, pp. 703–716, 2005.

[11] J. P. Ward, H. Kirshner, and M. Unser, "Is uniqueness lost for undersampled continuous-time auto-regressive processes?" *IEEE Signal Processing Letters*, vol. 19, no. 4, pp. 183–186, April 2012.

Reconstruction of Signals from Highly Aliased Multichannel Samples by Generalized Matching Pursuit

Ali Özbek[1], *Massimiliano Vassallo*[2], *Kemal Özdemir*[3], *Dirk-Jan van Manen*[1], *Kurt Eggenberger*[4]

[1]Schlumberger Gould Research, Cambridge, UK
[2]WesternGeco London Technology Centre, Gatwick, UK
[3]WesternGeco Oslo Technology Center, Asker, Norway
[4]WesternGeco Houston Technology Center, Houston, TX, USA
Emails: {ozbek, mvassallo, aozdemir, dmanen, keggenberger}@slb.com

***Abstract*—This paper considers the problem of reconstructing a bandlimited signal from severely aliased multichannel samples. Multichannel sampling in this context means that the samples are available after the signal has been filtered by various linear operators. We propose the method of Generalized Matching Pursuit to solve the reconstruction problem. We illustrate the potential of the method using synthetic data that could be acquired using multimeasurement towed-streamer seismic data acquisition technology. A remarkable observation is that high-fidelity reconstruction is possible even when the data are uniformly and coarsely sampled, with the order of aliasing significantly exceeding the number of channels.**

I. Introduction

In multichannel sampling, samples of a signal that was filtered by various linear operators are available. Suppose $\boldsymbol{m}(y)=\boldsymbol{h}(y)*s(y)$, where $\boldsymbol{m}(y)=[m_1,\cdots,m_J]$ are the measurements, and $\boldsymbol{h}(y)=[h_1,\cdots,h_J]$ are the operators. The samples are available at points $y_1,\cdots,y_L$, which may be regularly or irregularly spaced. The objective is to reconstruct bandlimited signal $s(y)$ at arbitrary points y. In Figure 1, we show a slight generalization, where, for each channel j, the measurements are undersampled by a factor of R_j with respect to the bandwidth of s. In the spectral domain, we have $\boldsymbol{m}(k_y)=\boldsymbol{H}(k_y)s(k_y)$, where k_y is the wavenumber (spatial frequency).

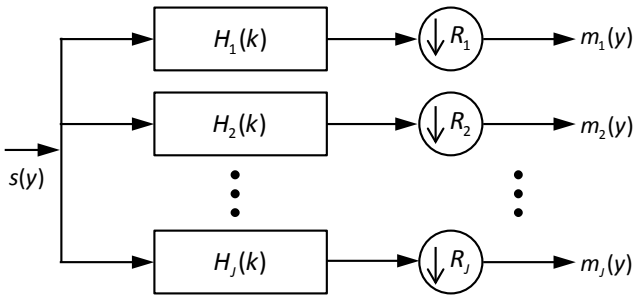


Fig. 1. Multichannel sampling.

The generalized sampling expansion proposed by Papoulis [1] implies that such a linear system, under certain conditions, allows reconstruction of the desired signal when $R_j=J$, $j=1,2,\cdots,J$. However, Papoulis [1] does not provide a readily realizable solution for the inversion of the system. Later, several articles were proposed to study the properties of the generalized sampling expansion, the well-posedness of the system, and a closed-form solution of the inverse problem [2, 3].

In some applications, such as marine seismic data acquisition, the decimation rate R_j can be significantly larger than the number of channels, J. In this case the order of aliasing significantly exceeds the number of channels. In the next section, we discuss a method that has shown promising performance in this setting.

II. Generalized Matching Pursuit

In this section, we describe a parametric matching pursuit method to solve the reconstruction problem that arises in multichannel sampling; we call it *Generalized Matching Pursuit (GMP)*, as its aim is to reconstruct a signal of which no direct samples may be available. Suppose that the unknown signal $s(y)$ is modeled as a sum of parametric basis functions $\beta(y;\boldsymbol{\theta}_n)$ with parameter set $\boldsymbol{\theta}_n$:

$$s(y)=\sum_n \beta(y;\boldsymbol{\theta}_n)\,. \tag{1}$$

There are various basis functions that can be considered; one possibility that is especially convenient for seismic applications is

$$\beta(y;\boldsymbol{\theta}_n)=A_n\exp\left[j\left(k_{y,n}y+\phi_n\right)\right], \tag{2}$$

where the parameter set $\boldsymbol{\theta}_n$ consists of amplitude A_n, phase ϕ_n, and wavenumber $k_{y,n}$. The corresponding measurements would then be

$$\boldsymbol{m}(y)=\sum_n \boldsymbol{H}(k_{y,n})\beta(y;\boldsymbol{\theta}_n)\,. \tag{3}$$

In GMP, the forward linear filters $H_j(k_y)$ are applied to each basis function; the filtered basis functions are then iteratively matched to the multichannel measurements. Iteratively, the basis function that, once forward filtered, jointly best matches all the input signals is used to reconstruct the desired output, with or without the forward filter applied. At the N-th iteration, i.e., after N-1 basis functions have been determined previously, the residual in the measurements is given by

$$\boldsymbol{r}^{N-1}(y)=\boldsymbol{m}(y)-\sum_{n=1}^{N-1}\boldsymbol{H}\left(k_{y,n}\right)\beta(y;\boldsymbol{\theta}_n)\,. \tag{4}$$

If a new term $\beta(y;\boldsymbol{\theta}_N)$ is added to the existing representation of the signal, the residual becomes $\boldsymbol{r}^N(y;\boldsymbol{\theta}_N)=\boldsymbol{r}^{N-1}(y)-\boldsymbol{H}(k_{y,N})\,\beta(y;\boldsymbol{\theta}_N)$, where the parameters of the new term, i.e., $\boldsymbol{\theta}_N$, are to be

determined by minimizing a metric of the residual calculated over measurement locations. One such metric is

$$\mu(\boldsymbol{\theta}_N) = \sum_l \left[\boldsymbol{r}^N(y_l;\boldsymbol{\theta}_N)\right]^H \boldsymbol{C}^{-1}\boldsymbol{r}^N(y_l;\boldsymbol{\theta}_N), \quad (5)$$

where the superscript H represents the Hermitian operator, $\boldsymbol{C}$ is a positive definite matrix, and y_l represents the sensor locations in the y direction. These locations can, in general, be irregularly spaced. The role of matrix $\boldsymbol{C}$ is to weight the contributions of different measurements to the cost function to be minimized. This can take into account the difference of energy content due to the different physics of the input measurements, as well as the signal-to-noise ratio that can vary in time, space, and frequency [10].

For basis functions chosen as in (2), it can be shown that the optimal A_N and ϕ_N can be analytically related to the residuals $\boldsymbol{r}^{N-1}$, the input sample positions y_l, and the optimal wavenumber $k_{y,N}$. Hence, the only remaining parameter to select is

$$k_{y,N} = \arg\max_k \mathcal{L}\left\{A_N\left(\boldsymbol{r}^{N-1}(y_l),k\right), \phi_N\left(\boldsymbol{r}^{N-1}(y_l),k\right)\right\}. \quad (6)$$

We call the objective function $\mathcal{L}$ the generalized Lomb spectrum, in analogy with the single-channel interpolation problem. There, in the case of sinusoidal basis functions, the objective function generated by *Interpolation by Matching Pursuit (IMAP)* with optimal amplitudes in the least-squares sense corresponds to the Lomb spectrum [4, 5, 6].

The GMP iterations can be terminated once the residual energy falls below a predetermined fraction of the input energy.

Next, we illustrate the antialiasing power of GMP for uniformly sampled multichannel data with a very simple multichannel sampling example. In this example, a single sinusoid signal with wavenumber 30 Km^{-1} is uniformly sampled at 25 Km^{-1}. In addition to the signal samples, the spatial gradient samples are available at the same locations. Due to uniform sampling, there is hard-aliasing, i.e., exact periodic replicas in the spectra of each channel. This is a reconstruction problem that cannot be solved by multichannel sinc interpolation [7], since the order of aliasing is greater than two. Figure 2 shows the cost function to select the optimum wavenumber (negative of the generalized Lomb spectrum) at the first iteration. The aliases of the correct wavenumber can be clearly seen. However, simultaneous use of the multichannel measurements in the optimization process results in the correct wavenumber being selected.

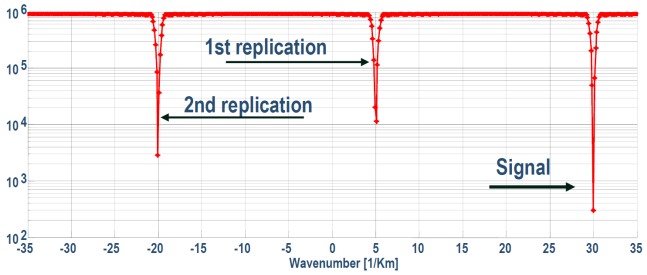


Fig. 2. Cost function for the optimum wavenumber in a hard-aliasing problem resulting from insufficient uniform sampling.

III. Application To Multichannel Sampling

Due to logistical and cost constraints, marine seismic acquisition systems can be deployed to acquire data only along a limited number of parallel lines (i.e., towed streamers) that are coarsely spaced in the crossline direction. Streamers are towed typically with crossline spacing of 75-100 m, resulting in coarse wavefield sampling that contrasts with adequate (non-aliasing) wavefield sampling of 6.25 m along the streamers (inline). Consequently, they do not adequately capture the full spatial bandwidth of the subsurface-scattered wavefield, leading to limitations in accurate subsurface imaging. Furthermore, conventional (pressure-only data) acquisition systems suffer from the ghost effect. The ghost is the reflection from the sea surface that interferes constructively or destructively with the upgoing wavefield (the signal of interest for imaging), reducing the seismic bandwidth at the low and high ends of the spectrum.

To address these critical limitations, a multimeasurement marine seismic acquisition platform was recently introduced. It is equipped with hydrophones to measure the pressure wavefield (P) and accelerometers to measure the particle acceleration vector ($\boldsymbol{A}$). The latter represents the vector spatial gradient of pressure as derived through the particle equation of motion, $\nabla P = -\rho \boldsymbol{A}$, where ρ is the fluid density [8].

A. Example: Reconstructing P from Aliased (P, P_y) Data

An important problem is to reconstruct (interpolate) the total pressure wavefield P at any desired position in the crossline direction from sparse samples of itself and its crossline gradient. P is the sum of the upgoing and downgoing (ghost) wavefields. For this problem, the unknown signal is $s(f,k_x,k_y)=P(f,k_x,k_y)$; the measurement vector is

$$\boldsymbol{m}(t,x,y_l) = \left[P(t,x,y_l) \quad P_y(t,x,y_l)\right]^T, \quad l = 1,2,\cdots,L, \quad (7)$$

where P_y is the crossline gradient of the pressure wavefield; the number of streamers (L) is typically 8-12. The forward linear operator is

$$\boldsymbol{H}(f,k_x,k_y) = \left[1 \quad jk_y\right]^T. \quad (8)$$

Here, f is the temporal frequency; k_x and k_y are the inline and crossline wavenumbers, respectively. As the data are well sampled in the temporal (t) and inline (x) coordinates, we can operate the GMP algorithm outlined in Section 2 for fixed values of f and k_x. The particular form that GMP takes for this reconstruction problem is referred to as *MIMAP (Multichannel Interpolation by Matching Pursuit)* [9].

Figure 3 shows a simple example reproducing linear events with energy up to 65 Hz and various incidence angles first decimated at 75 m and then reconstructed using different techniques. At every receiver position we modeled both the synthetic signal and its horizontal gradient. For the selected geometry, an event propagating horizontally generates first order alias at 10 Hz, and second order alias at 20 Hz, as shown in 2(b). Since MIMAP does not assume that the data comprise linear events in the implementation used for this example, the presence of high orders of aliasing presents a significant challenge for reconstruction.

To show the impact of the antialiasing capabilities of MIMAP, we interpolated the data with two standard techniques in addition to MIMAP: the sinc interpolation, and the multichannel sinc interpolation [7]. Results are shown in Figure 3. In Figure 3(a) we can see a region of the input time-space gather describing the pressure synthetics, sampled at 75 m, and the frequency-wavenumber transform of the overall gather. The high order of aliasing is clearly visible in the *f-k* domain. Figures 3(c) and 3(d) show the results of the single-component conventional sinc

interpolator, bandlimited in the spatial sampling bandwidth. As expected, only frequencies up to 10Hz are not subject to aliasing, and only the events with an incident angle close to zero can be properly interpolated (e.g., the event at 2.6 s). All the rest of the reconstructed information, in fact, corresponds to aliased replicas remapped to incorrect wavenumber positions.

Figures 3(e) and 3(f) show the result of the multichannel sinc interpolation, bandlimited to twice the spatial Nyquist. In this case, we can see that more events are reconstructed correctly in the t-y plot (e.g., events at around 2.4 s, 2.5 s and 2.6 s), and that all the events are reconstructed correctly up to 20 Hz. What is also interesting is that the multicomponent sinc seems to amplify the aliased events that cannot be reconstructed, as visible in the *f-k* gather above 20 Hz. Moreover, the shape of the region not affected by the alias, or affected by a first-order alias only, is clearly recognizable as the properly reconstructed area. Finally, in Figures 3(g) and 3(h), we can observe the results produced by MIMAP, and the removal of aliasing up to very high frequencies can be appreciated. All the events are well reconstructed.

B. Example: Reconstructing P^{up} from Aliased (P, P_y, P_z) Data

Using P, P_y, and P_z data that can be recorded by a multimeasurement streamer, another and more challenging problem would be to reconstruct P^{up} at any desired position without having access to any direct samples of it. This is called the joint interpolation and deghosting problem [10], where the task of separating the wavefield into its down- and upgoing components is performed simultaneously with the task of reconstructing it at any desired position. For this problem, the unknown signal is $s(f,k_x,k_y)=P^{up}(f,k_x,k_y)$; the measurement vector is

$$\boldsymbol{m}(t,x,y)=\left[P(t,x,y)\;\; P_y(t,x,y)\;\; P_z(t,x,y)\right]^T,\;\; l=1,2,\cdots,L, \quad (9)$$

and the forward linear operator that links the measurements to the unknown signal is the ghosting operator defined by

$$\boldsymbol{H}(f,k_x,k_y)=\left[\left(1+\xi e^{j2k_zZ}\right)\;\; k_y\left(1+\xi e^{j2k_zZ}\right)\;\; jk_z\left(1-\xi e^{j2k_zZ}\right)\right]^T. \quad (10)$$

Here, k_z is the vertical wavenumber, Z is the depth of the streamer, and ξ is the reflection coefficient of the sea surface. Through the ghost model, the P_z component brings independent new information on the unknown upgoing wavefield in the crossline direction, which is crucial for this application [10].

Figure 4 shows the application of the GMP technique to solve the joint interpolation and deghosting problem in the crossline direction using synthetic data. The data set was created by finite-difference modeling and simulates a 3D multimeasurement survey over a complex geological structure. The source signature spectrum is flat up to 30 Hz. The streamer depth is 50 m; the unusual depth was chosen to place the pressure ghost notch within the 30-Hz bandwidth. Given total P, P_y and P_z data sampled at 150 m where the data are severely aliased, the reconstructed upgoing pressure wavefield sampled at the desired 25-m interval show both the dealiasing and the deghosting capabilities of this approach.

Figure 4(a) shows the f-k_x-k_y transform of the total pressure wavefield before decimation, with pressure sampled over a 25-m x 25-m spatial grid. We can recognize the lack of energy in the low wavenumbers in the 15-Hz slice, and a circularly shaped notch in the 20-Hz and 25-Hz slices. The events that are not

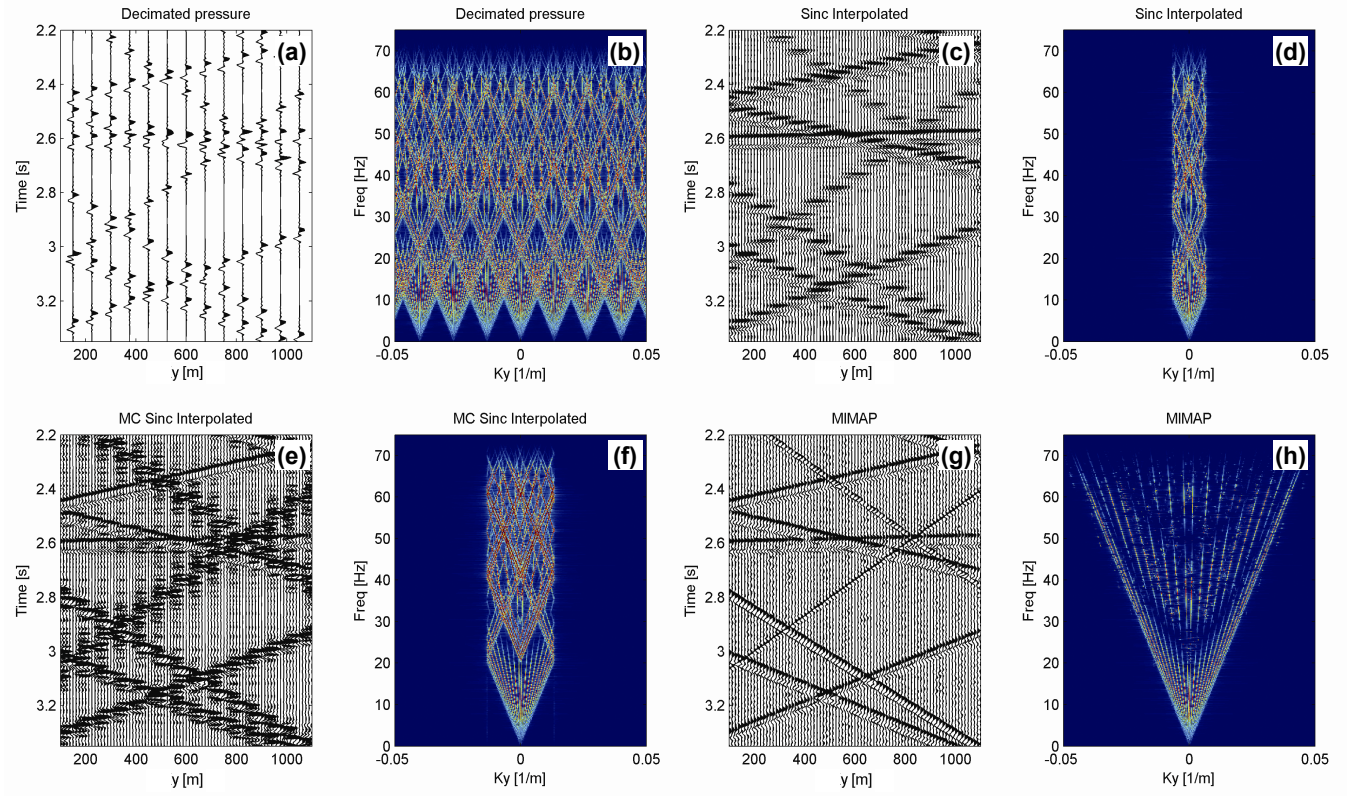


Fig. 3. Example with simple synthetics: close-up of a region of the *t-y* domain and *f-k* transforms of the whole dataset. (a, b) Input pressure, sampled at 75 m; (c, d) pressure reconstructed by using a sinc interpolator; (e, f) pressure reconstructed by using a multichannel sinc interpolator, also having as input the crossline gradients at the samples positions; (g, h) pressure reconstructed with MIMAP, also having as input the crossline gradients at the samples positions.

affected by the notch are still affected by the constructive interference of the ghost. Figure 4(b) shows the f-k_x-k_y transform of the reference upgoing pressure wavefield, sampled over a 25-m x 25-m spatial grid. The f-k_x-k_y transform of the total pressure wavefield after decimation of the data to 150 m in the crossline direction is shown in Figure 4(c). The first-order alias starts just above 5 Hz and the order of the alias grows significantly with frequency. Figure 4(d) shows the f-k_x-k_y transform of the upgoing pressure wavefield reconstructed by GMP, to a 25 m x 25 m spatial grid. The ghost notch is filled and the dealiasing impact of GMP is evident if we compare the output shown here with the spectrum of the input in the previous figures at high frequencies. Comparison of Figures 4(b) and 3(d) confirms the accuracy of joint interpolation and deghosting achieved by GMP.

IV. Summary And Conclusions

The problem of reconstructing a bandlimited signal from highly aliased multichannel samples was considered and a solution proposed in the form of Generalized Matching Pursuit. GMP proceeds by modeling the target signal as a sum of parametric basis functions that are matched to the multichannel data in a simultaneous and iterative fashion through application of the respective linear operators. It was shown that under quite general conditions GMP can achieve high-quality reconstructions of signals aliased by orders significantly higher than the number of different measurements, including the notoriously difficult case of regular undersampling, and signals for which no direct measurements are available.

We should emphasize that the results shown in this paper were obtained without using any priors (e.g., using a low-frequency solution, which is assumed to be unaliased, to constrain a high-frequency solution), which are commonly utilized to interpolate aliased data. In the same vein, the reconstructions were carried out independently at each temporal frequency, i.e., without making any assumptions on local wavefronts being planar.

During the presentation, we intend to show results obtained using real data acquired by multimeasurement towed-streamer seismic data acquisition technology; we had to omit them from this paper due to lack of space.

Acknowledgement

The synthetic data set was generated as part of collaborative projects between Schlumberger, Lawrence Livermore National Laboratory, and Statoil; we thank Shawn Larsen of Lawrence Livermore National Laboratory and Martin Musil and Clément Kostov of Schlumberger. We thank our colleagues Johan Robertsson, Tony Curtis, Smaine Zeroug, Ed Kragh, Phil Christie, Everhard Muyzert, and Ralf Ferber for stimulating discussions. We also thank Statoil for permission to show the synthetic data.

References

[1] A. Papoulis, "Generalized sampling expansion," *IEEE Trans. Circ. Syst.,* vol. 24, pp. 652-654, Nov. 1977.

[2] J. L. Brown, Jr., "Multi-channel sampling of low-pass signals," *IEEE Trans. Circ. Syst.,* vol. 28, pp. 101-106, Feb. 1981.

[3] J. L. Brown, Jr., and S. D. Cabrera, "Multi-Channel Signal Reconstruction Using Noisy Samples," *Proc. IEEE Int. Conf. Acoust. Speech Sig. Proc.*, vol. 3, pp. 1233 – 1236, Albuquerque, NM, April 1990.

[4] N. R. Lomb, "Least squares frequency analysis of unequally spaced data," *Astrophysics and Space Science,* vol. 39, pp. 447–462, 1976.

[5] J. D. Scargle, "Studies in astronomical time series analysis II. statistical aspects of spectral analysis of unevenly sampled data," *Astrophysical Journal*, vol. 263, pp. 835–853, 1982.

[6] K. Özdemir, A. Özbek, and M. Vassallo, "Interpolation of irregularly sampled data by matching pursuit," *Proc. EAGE Conference*, paper G025, Rome, June 2008.

[7] D. A. Linden, "A discussion of sampling theorems," *Proc. IRE*, vol. 47, pp. 1219-1226, 1959.

[8] J. Robertsson, I. Moore, M. Vassallo, A. K. Özdemir, D.-J. Van Manen, and A. Özbek, "On the use of multicomponent streamer recordings for reconstruction of pressure wavefields in the crossline direction," *Geophysics*, vol. 73, pp. A45–A49, 2008.

[9] M. Vassallo, A. Özbek, K. Özdemir, and K. Eggenberger, "Crossline wavefield reconstruction from multicomponent streamer data: Part 1 - Multichannel interpolation by matching pursuit using pressure and its crossline gradient," *Geophysics*, vol. 75, pp. WB53–WB67, 2010.

[10] A. Özbek, M. Vassallo, K. Özdemir, D.-J. Van Manen, and K. Eggenberger, "Crossline wavefield reconstruction from multicomponent streamer data: Part 2 - Joint interpolation and 3D up/down separation by generalized matching pursuit," *Geophysics*, vol. 75, pp. WB69–WB85, 2010.

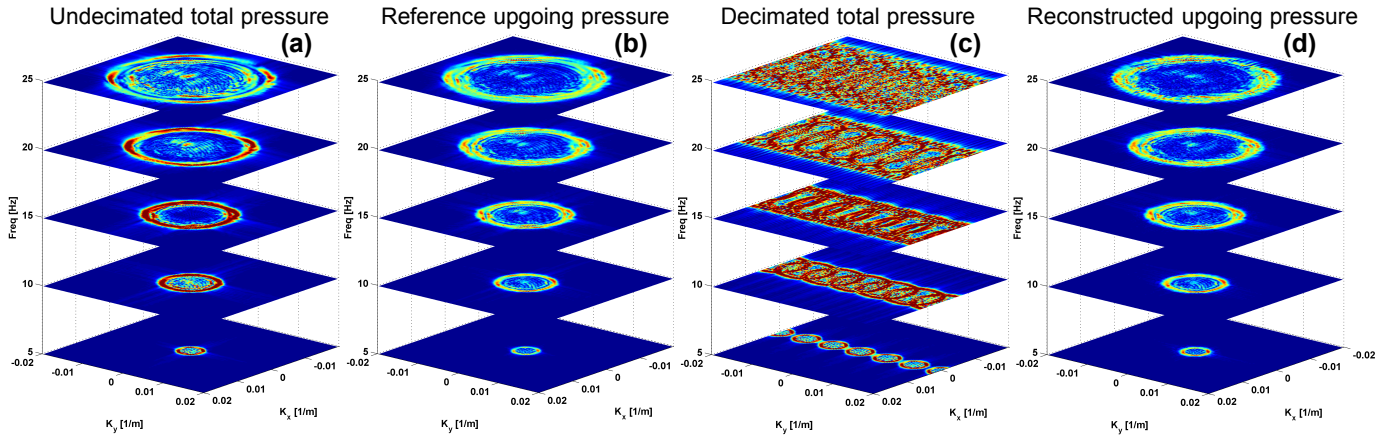


Fig. 4. Seismic synthetics in the 3D spectrum (f-k_x-k_y) domain. (a) reference total pressure wavefield, sampled over a 25- x 25-m grid; (b) reference upgoing pressure wavefield, sampled over the same grid; (c) input total pressure wavefield at 150-m crossline spacing; (d) upgoing pressure wavefield, reconstructed over a 25- x 25-m grid by GMP, processing P, P_y, and P_z at 150 m in the crossline.

Joint Signal Sampling and Detection

Mirosław Pawlak
Dept. of Electrical & Computer Eng.
University of Manitoba
Winnipeg, Manitoba, Canada, R3T 2N2
Email: Miroslaw.Pawlak@ad.umanitoba.ca

Ansgar Steland
Institute of Statistics
RWTH Aachen University
Germany
Email: steland@stochastik.rwth-aachen.de

***Abstract*—In this paper, we examine the joint signal sampling and detection problem when noisy samples of a signal are collected in the sequential fashion. In such a scheme, at the each observation time point we wish to make a decision that the observed data record represents a signal of the assumed target form. Moreover, we are able simultaneously to recover a signal when it departs from the target class. For such a joint signal detection and recovery setup, we introduce a novel algorithm relying on the smooth correction of linear sampling schemes. Given a finite frame of noisy samples of the signal we design a detector being able to test a departure from a target signal as quickly as possible. Our detector is represented as a continuous time normalized partial-sum stochastic process, for which we obtain a functional central limit theorem under weak assumptions on the correlation structure of the noise. The established limit theorems allow us to design monitoring algorithms with the desirable level of the probability of false alarm and able to detect a change with probability approaching one.**

***Index Terms*—joint sampling-detection, parametric signals, nonparametric alternatives**

I. Introduction

The problem of reconstructing an analog signal from its discrete samples plays a critical role in the modern technology of digital data transmission and storage. In fact, the theory of signal sampling and recovery has attracted a great deal of research activities lately, see [8], [9] and the references cited therein. In particular, the problem of signal sampling and recovery from imperfect data has been addressed in a number of recent works [5], [1], [2], [6]. The efficiency of sampling schemes depends strongly on the *a priori* knowledge of an assumed class of signals. For a class of bandlimited signals the signal sampling and recovery theory builds upon the celebrated Whittaker-Shannon interpolation scheme. On the other hand, there exists a class of nonbandlimited signals which can be recovered using the frequency rate below the Nyquist threshold. This is possible since this class is completely specified by a finite dimensional parameter. This parametric class of functions is often referred to as *finite rate innovation* signals [4], [3]. In practice, when only random samples are available it is difficult to verify whether a signal is bandlimited, parametric or belongs to some general function space. This calls for a joint nonparametric detection-reconstruction scheme to verify a type of the signal and simultaneously able to recover it. In fact, the problem of automatic rapid detection of signals differing from a reference (target) signal is important in many fields of signal processing and communication, e.g., in the analysis of radar signals and synchronization procedures the joint detection and reconstruction provides the basis to design effective receivers. The additional difficulty of designing detection/reconstruction procedures comes from the fact that samples are inherently noisy and observed sequentially within a fixed time frame. Hence, at the current frame we have a noisy data set $\{y_i : i \leq k\}$, and a detector should be applied immediately when a new observation y_{k+1} is available to the system. Hence, suppose we are given noisy measurements

$$y_k = f(k\tau) + \epsilon_k, \tag{1}$$

where τ is the sampling period, $\{\epsilon_k\}$ is a zero mean noise process, and $f(\bullet)$ is an unspecified signal which belongs to some signal space. In this paper we are interested in the following on-line detection problem. We are given a reference (target) parametric class of signals $\mathcal{S} = \{f(t;\theta) : \theta \in \Theta\}$, where Θ is a subset of a finite dimensional space, and wish to test the null hypothesis $H_0 : f \in \mathcal{S}$ against an arbitrary alternative $H_1 : f \notin \mathcal{S}$. Throughout the paper, we assume that the signal $f(t)$ of interest is observed over a finite time frame, i.e., $t \in [0, T]$, for some $0 < T < \infty$. Indeed, in practice we can only process a part of the signal which can be otherwise defined over an arbitrary interval. As a result, we are interested in methods relying on a finite data set $\{y_k : k = 0, \ldots, n\}$ obtained from model (1). Concerning the noise process in (1), we admit a wide class of correlated error processes. Our assumption is nonparametric and specifies a certain asymptotic behavior of the noise process. Specifically we assume that $\{\epsilon_k\}$ satisfies the so-called *invariance principle* or *functional central limit theorem* also often referred to as the Donsker's property, see [10] for further details. Hence, the condition on the error process employed in this paper is as follows.

Assumption 1 *Let $\{\epsilon_k\}$ be a weakly stationary stochastic process with zero mean which satisfies a functional central limit theorem, i.e.,*

$$n^{-1/2} \sum_{k=0}^{\lfloor ns \rfloor} \epsilon_k \Rightarrow \sqrt{\eta} B(s),$$

as $n \to \infty$, for some finite constant η.

Here $B(t)$ denotes a standard Brownian motion and $\Rightarrow$ stands for the convergence in distribution. Also $\lfloor x \rfloor$ denotes the greatest integer less or equal to x. It is worth mentioning that the validity of the functional central limit theorem, i.e.,

of Assumption 1 is not limited to the $i.i.d.$ case but also holds for many dependent stationary processes with summable auto-covariances. For instance, it holds for linear processes and mixing processes [10]. The dependence structure of the measured data is controlled by the parameter η appearing in Assumption 1. This parameter is identified with the limit $\lim_{n\to\infty} Var(n^{-1/2}\sum_{k=0}^{n}\epsilon_k)$ and is often referred to as the time-average (long-run) variance of $\{\epsilon_k\}$.

Our test statistic builds upon the signal recovery methods developed in [5], [6], where it has been proved that they possess the consistency property, i.e., they are able to converge to a large class of signals not necessarily being bandlimited. A generic form of such estimates is given by

$$\hat{f}_n(t) = \tau \sum_{j=0}^{n} y_j \mathcal{K}_\Omega(j\tau, t), \tag{2}$$

where $\mathcal{K}_\Omega(u,t)$ is the reconstruction kernel parameterized by the parameter Ω. For the consistency results the parameter Ω and the sampling period τ should depend on the data size n and be selected appropriately. In fact, we need that $\Omega \to \infty$ and $\tau \to 0$ as $n \to \infty$ with the controlled rate. For example, the choice $\Omega = n^{1/3}$ and $\tau = n^{-1}$ would be sufficient to assure the consistency for a wide nonparametric class of signals defined on the a finite interval. Our detection algorithm uses the data observed over the interval $[0,T]$ and therefore we select $\tau = T/n$ and fix Ω to some large number. The kernel $\mathcal{K}_\Omega(u,t) = \sin(\Omega(t-u))/\pi(t-u)$ is particularly important since it is the reproducing kernel for bandlimited signals with the bandwidth Ω. For a broader class of signals we can use generalized kernels $\mathcal{K}_\Omega(u,t) = \sum_k \phi(\Omega(u-k))\psi(\Omega(t-k))$, where $\phi(t)$, $\psi(t)$ can be specified as biorthogonal functions [8].

II. Reconstruction and Detection Algorithms

Our detection technique is relying on the consistent reconstruction method defined in (2). Our asymptotic results assume $n \to \infty$ but we will also provide useful approximations for finite n. Note that n can be regarded as the planed maximum number of observations in the time interval $[0,T]$. In this paper we address the following question: how long do we have to sample the signal, until the available data provide enough evidence to reject the null hypothesis $H_0 : f \in \mathcal{S}$? A specific example of the null hypothesis class $\mathcal{S}$ is a class of signals that are a superposition of shifted versions of a known pulse $h(t)$, i.e., $f(t;\theta) = \sum_{k=1}^{L} a_k h(t-t_k)$. Here θ is a $2L$ dimensional vector of unknown parameters.

Our goal in this paper is to decide whether the null hypothesis is true or not, given a sequentially observed data set drawn from the observation model (1), where $f(t)$ is an unknown signal from a large nonparametric function space. Hence, if the alternative signal is unknown, we propose a detector which can be computed without specifying this signal. We will use a sequential version of the nonparametric estimator (2), which automatically adapts to the unknown alternative signal as sampling proceeds. Specifically, we use $\hat{f}_n(t)$ as a basic building block of our detection method, i.e., we stop our detection process at the first time point $t = k\tau$ if a certain distance measure between $\hat{f}_k(t)$ and the target parametric signal $f(t;\theta_0)$ from $\mathcal{S}$ is is too large. Here θ_0 denotes the "true" parameter if the null hypothesis holds. Since the parameter θ_0 is unknown we replace it in our test statistic with its consistent estimate $\hat{\theta}$, see [7] for an extensive overview of estimation algorithms and their performance for specific classes of parametric signals. In [3] the estimation problem associated with the class of finite rate innovation signals has been also examined. To define our detection scheme, let us introduce the following sequential partial sum process, which represents the sequence of the estimators as a step function

$$\mathcal{F}_n(s,t) = \sqrt{\tau} \sum_{0\le l\le \lfloor ns\rfloor} [y_l - f(l\tau;\hat{\theta})]\mathcal{K}_\Omega(l\tau,t), \tag{3}$$

for $0 < s_0 \le s \le 1$, $t \in [0,T]$. The condition $s_0 \le s$ ensures that at least the first $n_0 = \lfloor ns_0 \rfloor$ observations are used ensuring a certain degree of precision in the reconstruction. This allows us in our asymptotic analysis to replace $\hat{\theta}$ in (3) by θ_0. Then, for $s = k/n$ the value $\mathcal{F}_n(k/n,t)$ can be interpreted as the deviation of $\tau^{-1/2}(\hat{f}_k(t) - E_0\hat{f}_k(t))$, where throughout the paper E_0 and P_0 denote that the expectation and probability are taken under the null hypothesis, i.e., that $f(t) = f(t;\theta_0)$. The interpretation of $\mathcal{F}_n(s,t)$ as a function of one variable is as follows:

- For fixed t the step function $s \mapsto \mathcal{F}_n(s,t)$ describes the sequence of deviations of $\hat{f}_{\lfloor ns\rfloor}(t)$ from $f(t;\hat{\theta})$ as sampling proceeds.
- For fixed s the function $t \mapsto \mathcal{F}_n(s,t)$ is the current estimate of the whole signal, using $\lfloor ns\rfloor$ sampled values.

The sequential nonparametric decision problem for rejecting the hypothesis $H_0 : f \in \mathcal{S}$ can now be handled by the following detector statistics. A *global maximum detector* is defined as follows

$$M_n = \min\left\{ n_0 \le k \le n : \max_{0\le t\le Tk/n} |\mathcal{F}_n(k/n,t)| > c_M \right\}$$

for some appropriately chosen control limit c_M. The detector M_n looks at the largest absolute value of the deviation process. Notice that when calculating the maximum at a candidate time point Tk/n, the maximum is determined for time points t between 0 and Tk/n. That interval corresponds to the time frame where observations are present. For $t > Tk/n$ the estimator $\hat{f}_k(t)$ can be considered as an extrapolation scheme. Alternatively, one can consider a *global integrated detector*

$$I_n = \min\left\{ n_0 \le k \le n : \int_0^{Tk/n} |\mathcal{F}_n(k/n,t)|^2\, dt > c_I \right\}$$

for some appropriately chosen control limit c_I. Without loss of generality, however, we confine our investigation to the detector M_n, which is easy to calculate and interpret. In order to assess the statistical accuracy of the detector M_n we need to establish the limiting distribution of the process $\mathcal{F}_n(s,t)$. This is shown in the next section.

III. Limit Distributions

The statistical accuracy of the aforementioned detection scheme M_n depends critically on the the choice of the threshold parameter c_M. The asymptotic choice of this parameter can be obtained from the limiting distribution of $\mathcal{F}_n(s,t)$. Below we establish that the limiting distribution is a locally stationary Gaussian process $\mathcal{F}(s,t)$ with mean 0 and a certain covariance function.

Theorem 1: Suppose the noise process $\{\epsilon_k\}$ meets Assumption 1. Then under the hypothesis H_0 we have

$$\mathcal{F}_n(s,t) \Rightarrow \mathcal{F}(s,t), \qquad n \to \infty,$$

where the limit stochastic process, $\mathcal{F}(s,t)$ is given by

$$\mathcal{F}(s,t) = \sqrt{T\eta} \int_0^s \mathcal{K}_\Omega(Tz,t)dB(z).$$

As a result, the process $\mathcal{F}(s,t)$ is a locally stationary Gaussian process with the following covariance function

$$cov(\mathcal{F}(s_1,t_1),\mathcal{F}(s_2,t_2)) = T\eta \int_0^{\min(s_1,s_2)} \mathcal{K}_\Omega(Tz,t_1)\mathcal{K}_\Omega(Tz,t_2)dz.$$

The smoothness of the sample paths of the Gaussian process $\mathcal{F}(s,t)$ is determined by smoothness of its variance, i.e., the function $T\eta\int_0^s \mathcal{K}_\Omega^2(Tz,t)dz$. The above result allows us to establish the limit of our detector statistic. Under the conditions of Theorem 1 the following central limit theorem also holds true.

$$M_n/n \Rightarrow \mathcal{M} = \inf\{s \in [s_0,1] : \sup_{0\le t\le sT} |\mathcal{F}(s,t)| > c_M\}.$$

These results allow us to specify the control limit c_M in such a way that the probability of a false alarm in the time frame $[s_0,1]$ is not greater than $\alpha < 1$. For our detector M_n one can proceed as follows. The detection error (under the hypothesis H_0) occurs if $M_n/n < 1$ and $P_0(M_n/n < 1) \to P(\mathcal{M} < 1)$ by the aforementioned result. Since the event $\{\mathcal{M} > z\}$ is equivalent to following one

$$\{\sup_{s_0\le s\le z}\sup_{0\le t\le sT} |\mathcal{F}(s,t)| \le c_M\}, \tag{4}$$

we can obtain a procedure for selecting c_M with an asymptotic detection error being equal to α. In fact, we choose c_M as the $1-\alpha$ quantile of the distribution of the complement of the event in (4) with $z=1$, i.e., the constant c_M is found as the smallest c being the solution of the following inequality

$$P\left(\sup_{s\in[s_0,1]}\sup_{t\in[0,sT]} |\mathcal{F}(s,t)| > c\right) \le \alpha, \tag{5}$$

where the probability is taken with respect to the extrema of the absolute value of the Gaussian process $\mathcal{F}(s,t)$.

The question arises how the above results can be applied in practice. The distribution of the random variable $X = \sup_{s_0\le s\le 1}\sup_{0\le t\le sT}|\mathcal{F}(s,t)|$ required to evaluate the false alarm error can be simulated by Monte Carlo methods using the following algorithm.

1) Generate trajectories of the Gaussian process $\mathcal{F}(s,t)$ on a grid $\{(s_i,t_j) : i = 1,\dots,N, j = 1,\dots,N\}$ where $0 \le s_1 < \cdots < s_N \le 1$ and $0 \le t_1 < \cdots < t_N \le T$.
2) Return X by calculating the maximum of the values $|\mathcal{F}(s_i,t_j)|$ for all (i,j) such that the constraints $s_0 \le s_i \le 1$ and $0 \le t_j \le s_i T$ are satisfied.
3) Repetitions of Step 1 and Step 2 produce realizations of X that can be utilized for estimating $c_M(\alpha)$.

Simulating the process $\mathcal{F}(s,t)$ in Step 1 is feasible, since the covariance function can be evaluated numerically provided that T and η are known. The choice of η is critical for the accuracy of our detectors. We wish to estimate η without assuming which hypothesis holds, i.e., to estimate η using only the available data $\{y_0,\dots,y_k\}$ without the knowledge of the signal shape. Here we can utilize the discrepancies of local means. One of such estimates takes the form

$$\widetilde{\eta}_k = \frac{b_k}{2(L-1)}\sum_{j=1}^{L-1}(A_j - A_{j-1})^2, \tag{6}$$

where $A_j = \sum_{l=jb_k}^{jb_k+b_k-1} y_l/b_k$ is the local mean, $j = 0,1,\dots,L$ and $L+1 = \lfloor (k+1)/b_k \rfloor$ denotes the number of data groups. It can be demonstrated [11] that this estimate can converge to the true η with the rate $O_P(k^{-1/3})$ with virtually no assumptions on the form of the underlying signals.

Having established the asymptotic distributions under the null hypothesis, it remains to see how our detection method behaves when $f \notin \mathcal{S}$, i.e., when the true signal differs from the target parametric signal. We can consider a class of local alternatives for modeling this situation, i.e., let

$$f(t) = f(t;\theta_0) + a_n g(t), \tag{7}$$

where a_n is the sequence tending to zero as $n \to \infty$ and $g(t)$ is a fixed function assumed to be piecewise continuous and bounded. Under this condition and Assumption 1 we can show that under the alternative local hypothesis and the choice $a_n = n^{-1/2}$ the process $\mathcal{F}_n(s,t)$ has the following limit

$$\mathcal{F}^A(s,t) = \mathcal{F}(s,t) + T^{-1/2}\int_0^{sT} \mathcal{K}_\Omega(z,t)g(z)dz, \tag{8}$$

where $\mathcal{F}(s,t)$ is the locally stationary Gaussian process found in Theorem 1. It is worth noting that if the departure from the reference signal $f(t;\theta_0)$ in the local alternative in (7) is of order $a_n = O(n^{-\beta})$, for $\beta > 1/2$, then there is no visible effect on the asymptotic distribution, i.e., $\mathcal{F}_n(s,t) \Rightarrow \mathcal{F}(s,t)$. Thus, even in large samples there is no chance to detect such small departures from the target signal. The rate $\beta = 1/2$ is the right order for getting a non-trivial limit distribution. The result in (8) allows us to evaluate the power $\mathcal{P}_n = P_1(TM_n/n < T)$ of our detector. In fact, the limit in (8) yields

$$\lim_{n\to\infty}\mathcal{P}_n = P\left(\sup_{s_0\le s\le 1}\sup_{0\le t\le sT}|\mathcal{F}^A(s,t)| > c_M\right). \tag{9}$$

This holds for any c_M but the proper value of c_M can be obtained by satisfying the bound for the probability of false alarm in (5). In practise, the probability in (9) can be evaluated by the aforementioned Monte Carlo algorithm.

τ	0.01	0.015	0.02	0.025	0.03
r_τ	0.0924	0.0640	0.0571	0.0480	0.0432

TABLE I
SIMULATED REJECTION RATE FOR VARIOUS SAMPLING INTERVALS τ

IV. SIMULATION STUDIES

In our simulation studies we will focus on the issues related to the choice of the proper control limit and the resulting detector rejection rate and power. This is studied in the context of the length of the sampling interval τ and the problem of the influence that selection of the filter bandwidth Ω has on the detector power. We assume that the target signal is $f_0(t) = \sin(4t)$ on $[0, 2]$. This signal undergoes the jump-point distortion to produce the alternative signal $f_1(t) = f_0(t) + 0.2\mathbf{1}(t \geq 1)$. Taking into account the global maximum norm detector M_n we follow the proposed Monte Carlo algorithm to estimate the proper control limit c_M being the sample 95%-quantile of 50000 simulation replicates. Our base reconstruction algorithm is the post-filtering method [5] utilizing the kernel function $\mathcal{K}_\Omega(u, t) = \sin(\Omega(t-u))/\pi(t-u)$, where Ω is the bandwidth of a low-pass filter. To study the influence of the sampling interval τ on c_M, we applied the above procedure with $s_0 = 0.1$, $\sqrt{\eta} = 0.2$, $\Omega = 10$, $n = 100$. The true rejection rate (the probability of rejection under the null hypothesis) denoted by r_τ was estimated by a Monte Carlo simulation with 50000 repetitions for each given $\tau \in \{0.01, \ldots, 0.03\}$. Since $n\tau = T$ therefore this corresponds to the design intervals ranging from $[0, 1]$ to $[0, 3]$. Note that the fixed value $\tau = 0.02$ was used in the illustrative example. Table I provides the results. It can be seen that there is some influence of the sampling interval on the accuracy of the approximation, but it is still moderate for a rather large range of values of τ. There is an evident drop in the value of the rejection rate for τ larger than 0.01 corresponding to design intervals larger than $[0, 1]$.

Next, we studied the influence of the filter bandwidth Ω of our reconstruction algorithm $\hat{f}_n(t)$ on the detection power (defined in (9)) using the corrected control limit. The parameter η was estimated by the method mentioned in Section III. We employed the fixed alternative $f_1(t) = f_0(t) + 0.1\sin\left(8(t-1) + \frac{\pi}{2}\right), t \in [0, 2]$. This alternative is characterized by the frequency and phase deformation, although the difference between $f_0(t)$ and $f_1(t)$ is small. The results (shown in Figure 1) indicate that there is an optimal value $\Omega^* \in [8, 12]$ that maximizes the detector power. The value of Ω^* is about 10.5 for n ranging from 500 to 1000. The corresponding power for the optimal values of Ω is above 95% ($n = 750$) and 99% ($n = 1000$). This is a quite remarkable fact noting that the L_2 norm of $f_1(t) - f_0(t)$ is as small as 0.0098.

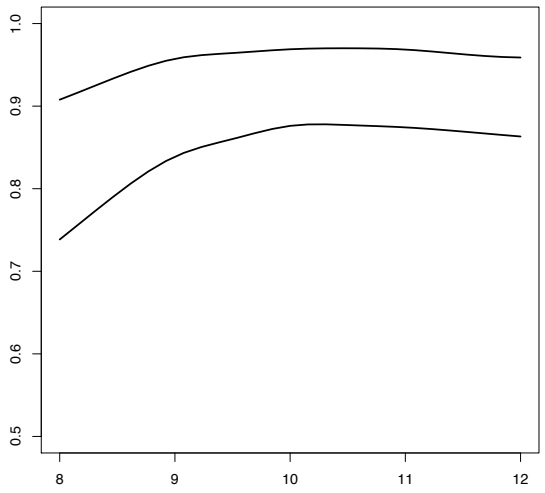


Fig. 1. Simulated power in (9) to detect a change in frequency and phase as a function of $\Omega \in [8, 12]$ for the sample sizes $n = 750$ (bottom curve) and $n = 1000$.

V. CONCLUDING REMARKS

We investigated a new joint sampling-detection procedure for testing the parametric form of a signal observed in the presence of correlated noise. Our detection methods are based on sequentially applied reconstruction algorithms which are related to linear sampling schemes. The asymptotic distribution of our detectors is established via functional central limit theorems and Donsker's invariance principle. This allows us to evaluate the probability of false alarm and the corresponding control limit. The asymptotic performance under local alternatives is also examined.

REFERENCES

[1] A. Aldroubi, C. Leonetti, and Q. Sun. Error analysis of frame reconstruction from noisy samples. *IEEE Trans. Signal Processing*, 56:2311–2315, 2008.
[2] Y.C. Eldar and M. Unser. Non-ideal sampling and interpolation from noisy observations in shift-invariant spaces. *IEEE Trans. Signal Processing*, 54:2636–2651, 2006.
[3] I. Maravic and M.Vetterli. Sampling and reconstruction of signals with finite rate of innovation in the presence of noise. *IEEE Trans. Signal Processing*, 53:2788–2805, 2005.
[4] M.Vetterli, P. Marziliano, and T. Blu. Sampling signals with finite rate of innovation. *IEEE Trans. Signal Processing*, 50:1417–1428, 2002.
[5] M. Pawlak, E. Rafajłowicz, and A. Krzyżak. Postfiltering versus prefiltering for signal recovery from noisy samples. *IEEE Trans. Information Theory*, 49:3195–3212, 2003.
[6] M. Pawlak and U. Stadtmüller. Signal sampling and recovery under dependent noise. *IEEE Trans. Information Theory*, 53:2526–2541, 2007.
[7] P. Stoica and R. Moses. *Spectral Analysis of Signals*. Prentice-Hall, Upper Saddle River, 2005.
[8] M. Unser. Sampling – 50 years after Shannon. *Proceedings of the IEEE*, 88:569–587, 2000.
[9] P.P. Vaidyanathan. Generalizations of the sampling theorems: seven decades after Nyquist. *IEEE Trans. on Ciruits and Systems – I : Fundamental Theory and Applications*, 48:1094–1109, 2001.
[10] W. Whitt. *Stochastic Process Limits*. Springer, New York, 2002.
[11] W.B. Wu and Z. Zhao. Inference of trends in time series. *J.R.Statist. Soc. B*, 69:391–410, 2007.

On Optimal Sampling Trajectories for Mobile Sensing

Jayakrishnan Unnikrishnan and Martin Vetterli
Audiovisual Communications Laboratory, School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland
Email: {jay.unnikrishnan, martin.vetterli}@epfl.ch

***Abstract*—We study the design of sampling trajectories for stable sampling and reconstruction of bandlimited spatial fields using mobile sensors. As a performance metric we use the path density of a set of sampling trajectories, defined as the total distance traveled by the moving sensors per unit spatial volume of the spatial region being monitored. We obtain new results for the problem of designing stable sampling trajectories with minimal path density, that admit perfect reconstruction of bandlimited fields. In particular, we identify the set of parallel lines with minimal path density that contains a set of stable sampling for isotropic fields.**

I. Introduction

Let the square-integrable mapping $f : \mathbb{R}^d \mapsto \mathbb{C}$ denote a d-dimensional time-invariant spatial field, with $f(r)$ representing the field value at a location r in d-dimensional space. The Fourier transform of f is defined as

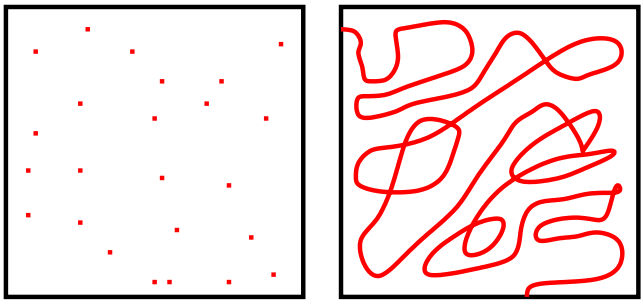
$$F(\omega) = \int_{\mathbb{R}^d} f(r) \exp(-\mathrm{i}\langle \omega, r\rangle) dr, \qquad \omega \in \mathbb{R}^d \tag{1}$$

where i denotes the imaginary unit, and $\langle u, v\rangle$ denotes the scalar product between vectors u and v in $\mathbb{R}^d$. We say that f is bandlimited to some set $\Omega \subset \mathbb{R}^d$, if the Fourier transform F of f is supported on Ω. In this case we write $f \in \mathcal{B}_\Omega$ where $\mathcal{B}_\Omega$ denotes the collection of fields with finite energy bandlimited to Ω, i.e.,

$$\mathcal{B}_\Omega := \{f \in L^2(\mathbb{R}^d) : F(\omega) = 0 \text{ for } \omega \notin \Omega\}. \tag{2}$$

The classical theory of sampling and reconstructing such high-dimensional bandlimited fields dates back to Petersen and Middleton [1] who identified conditions for reconstructing such fields from their measurements on a lattice of points in space. Further research on non-uniform sampling generated more results on conditions for perfect reconstruction from samples taken at non-uniformly distributed spatial locations [2], [3], [4], [5], [6], [7]. Such works primarily deal with the problem of reconstructing the field from measurements taken by a collection of static sensors distributed in space, like that shown in Figure 1(a), and hence the performance metric usually used to quantify the efficiency of a sampling scheme is the spatial density of samples which is exactly equal to the number of sensors required for sampling per unit volume of the spatial region being monitored.

In this paper we consider the problem of reconstructing a bandlimited spatial field (where $d = 2$ or 3) using its samples taken by a mobile sensor that moves along a continuous path through space taking measurements along its path, as shown in Figure 1(b). In such cases it is often relatively inexpensive to increase the spatial sampling rate along the sensor's path while the main cost of the sampling scheme comes from the total distance that needs to be traveled by the moving sensor. Hence it is reasonable to assume that the sensor can record the field values at an arbitrarily high but finite resolution on its path. Furthermore, for such a sampling application, the density of the sampling points in $\mathbb{R}^d$ used in classical sampling theory is not a relevant performance metric. Instead, as we argued in our previous work [8] [9], a more relevant metric is the average distance that needs to be traveled by the sensor per unit spatial volume (or area, for $d = 2$). We call this metric the *path density*. Such a metric is relevant in applications like environmental monitoring using moving sensors [10], [11], where the path density directly measures the distance moved by the sensor per unit area. This metric is also useful in designing k-space trajectories for Magnetic Resonance Imaging (MRI) [5], where the path density captures the total length of the trajectories per unit area in k-space which can be used as a proxy for the total scanning time per unit area in k-space.



(a) Static sampling on points (b) Mobile sampling on a curve

Fig. 1. Two approaches for sampling a field in $\mathbb{R}^2$

In [8] and [9] we introduced the problem of designing sampling trajectories for bandlimited fields that are minimal in path density. We obtained conditions on unions of uniformly spaced straight line trajectories that admit perfect reconstruction of bandlimited fields. From this class of trajectories, we identified those with minimal path density. In this paper we extend our past work to arbitrary configurations of parallel line trajectories. We introduce the notion of trajectories that admit

stable sampling. We identify new designs of trajectories for fields in $\mathbb{R}^d$, $d \geq 3$ that are strictly better in path density than those identified in [9].

The paper is organised as follows. In Section II we describe the formal problem statement, in Section III we present our new results and we conclude with some discussion in Section IV. Below we introduce notations we use frequently in the paper.

Notation: We use $\langle , \rangle$ to denote the canonical inner product, and e_k to denote the unit vector along the k-th coordinate axis. For $u \in \mathbb{R}^d$ we denote the hyperplane orthogonal to u through the origin by $u^\perp = \{x \in \mathbb{R}^d : \langle x, u \rangle = 0\}$. For a set $S \subset \mathbb{R}^d$ we use $|S|$ to denote the volume of S relative to its affine hull, relint(S) to denote the relative interior of S, $S(x)$ to denote its shifted version $S(x) = \{y + x : y \in S\}$, and $\mathcal{P}_{u^\perp} S$ to denote the orthogonal projection of S onto the hyperplane $u^\perp$. We use B_a^d and $B_a^d(x)$ for denoting spherical balls of radius a centered at the origin and $x \in \mathbb{R}^d$ respectively. For a discrete set Λ we use $\#(\Lambda)$ to denote its cardinality.

II. Problem Statement

A *trajectory* p_i in $\mathbb{R}^d$ refers to a curve in $\mathbb{R}^d$. We represent a trajectory by a continuous function $p(.)$ of a real variable taking values on $\mathbb{R}^d$:

$$p : \mathbb{R} \mapsto \mathbb{R}^d.$$

A *trajectory set* P is defined as a countable collection of trajectories:

$$P = \{p_i : i \in \mathbb{I}\} \tag{3}$$

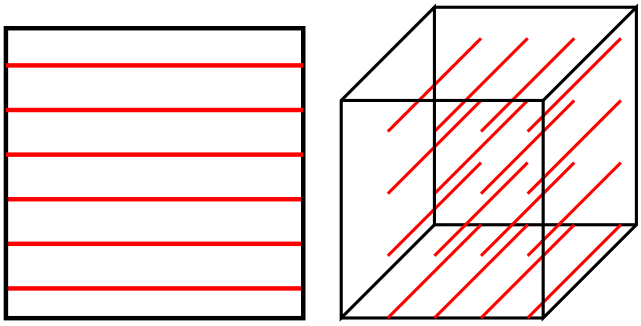
where $\mathbb{I}$ is a countable set of indices and for each $i \in \mathbb{I}$, p_i is a trajectory in the trajectory set P. For any given trajectory set P we denote its *path density* by $\ell(P)$ defined as follows:

$$\ell(P) := \limsup_{a \to \infty} \frac{\sup_{x \in \mathbb{R}^d} \mathcal{D}^P(a, x)}{\text{Vol}_d(a)} \tag{4}$$

where $\mathcal{D}^P(a, x)$ represents the total arc-length of trajectories from P located within the ball $B_a^d(x)$ and $\text{Vol}_d(a)$ represents the volume of the d-dimensional ball. A simple example of a trajectory set in $\mathbb{R}^2$ is a doubly infinite sequence of equispaced parallel lines through $\mathbb{R}^2$ (e.g., see Figure 2(a)). We call such a trajectory set a *uniform set in* $\mathbb{R}^2$. Such a uniform set has a path density equal to $\frac{1}{\Delta}$ (see [9, Lem 2.2]) where Δ is the spacing between the lines. Similarly a *uniform set in* $\mathbb{R}^d$ is defined as a collection of parallel lines in $\mathbb{R}^d$ such that the cross-section forms a $(d-1)$-dimensional lattice, as shown in Figure 2(b).

We say that a set of points $\Lambda \subset \mathbb{R}^d$ is *uniformly discrete* if we have $\inf\{\|x - y\| : x, y \in \Lambda, x \neq y\} > 0$, i.e., there exists $r > 0$ such that for any two distinct points $x, y \in \Lambda$ we have $\|x - y\| > r$.[1] We say that Λ forms a *set of stable sampling*

[1]For example lattices in $\mathbb{R}^d$ are uniformly discrete, but a sequence in $\mathbb{R}^d$ converging to a point in $\mathbb{R}^d$ is not.



(a) Uniform set in $\mathbb{R}^2$ (b) Uniform set in $\mathbb{R}^3$

Fig. 2. Examples of uniform sets in $\mathbb{R}^2$ and $\mathbb{R}^3$.

[4][3] for a set $\Omega \subset \mathbb{R}^d$ if there exists positive scalars A and B such that

$$A\|f\|^2 \leq \sum_{x \in \Lambda} |f(x)|^2 \leq B\|f\|^2, \text{ for all } f \in \mathcal{B}_\Omega. \tag{5}$$

Further, let $\mathcal{A}_\Omega$ denote the collection of all uniformly discrete sets $\Lambda \subset \mathbb{R}^d$ that form sets of stable sampling for Ω. Classical sampling theory is primarily concerned with the elements of $\mathcal{A}_\Omega$, e.g., Nyquist sampling lattices [1].

The following are some desirable properties of sampling trajectory sets.

Definition 2.1: A trajectory set P of the form (3) is called a *stable Nyquist trajectory set* for $\Omega \subset \mathbb{R}^d$ if it satisfies the following conditions:

- **(C1)** [*Nyquist*] There exists a uniformly discrete set Λ of points on the trajectories in P such that Λ forms a stable sampling set for $\mathcal{B}_\Omega$, i.e., $\Lambda \subset \{p_i(t) : i \in \mathbb{I}, t \in \mathbb{R}\}$ and $\Lambda \in \mathcal{A}_\Omega$.
- **(C2)** [*Non-degeneracy*] For any $x \in \mathbb{R}^d$, there is a continuous curve of length no more than $\mathcal{D}^P(a, x) + \text{o}(a^d)$ that contains the portion of the trajectory set P that is located within $B_a^d(x)$.

Here condition **(C2)** is a regularity condition to ensure that the path density metric does indeed capture the total distance traveled per unit area by a single moving sensor using the trajectories in P. We also introduce a special notation for the collection of all stable Nyquist trajectory sets:

Definition 2.2: We use $\mathcal{N}_\Omega$ to denote the collection of all stable Nyquist trajectory sets for Ω, i.e., $\mathcal{N}_\Omega$ is the collection of all trajectory sets P of the form (3) that satisfy conditions **(C1)** and **(C2)**.

Sampling theory for mobile sensing is primarily concerned with identifying trajectory sets in $\mathcal{N}_\Omega$. The key optimization problem that we seek to solve is to identify trajectory sets in $\mathcal{N}_\Omega$ with minimal path density:

$$\min_{P \in \mathcal{N}_\Omega} \ell(P). \tag{6}$$

In [9] and [12] we identified various examples of trajectory sets in $\mathcal{N}_\Omega$, and obtained partial solutions to (6) optimizing over

specific restricted classes of trajectories, such as uniform sets and unions of uniform sets. In this paper we present optimality results from broader classes of trajectory sets.

III. New optimality results for parallel lines

Let P denote a trajectory set composed of parallel lines in $\mathbb{R}^d$. For any $x \in \mathbb{R}^d$ let $N_a^x(P)$ denote the number of lines in P that intersect the d-dimensional ball $B_a^d(x)$ of radius a centered at x. We restrict our attention to trajectory sets that are homogenous in the sense defined below.

Definition 3.1: We say that P is a *homogenous parallel set* if

$$\lim_{a\to\infty} \frac{N_a^x(P)}{|B_a^{d-1}|} \text{ exists and is equal for all } x \in \mathbb{R}^d.$$

Most practically useful parallel trajectory sets such as uniform sets, approximately uniform sets (e.g., with bounded offsets) and their finite unions are homogenous. For $\Omega \subset \mathbb{R}^d$ we use $\mathcal{H}_\Omega$ to denote homogenous parallel sets in $\mathcal{N}_\Omega$. Below, we characterize the path density of homogenous parallel sets.

Lemma 3.1: Any homogenous parallel set P in $\mathbb{R}^d$ satisfies

$$\ell(P) = \lim_{a\to\infty} \frac{N_a^0(P)}{|B_a^{d-1}|}. \qquad \square \tag{7}$$

We provide a proof in the appendix. We now tackle (6) for trajectory sets in $\mathcal{H}_\Omega$ and compact convex symmetric sets Ω. We first establish a lower bound on the path density.

Proposition 3.2: Let $\Omega \subset \mathbb{R}^d$ be a compact convex set with non-empty interior. Assume further that Ω has a point of symmetry at the origin. Let $Q \in \mathcal{H}_\Omega$ be a trajectory set composed of lines parallel to $q \in \mathbb{R}^d$. Then $\ell(Q) \geq \frac{|\Omega \cap q^\perp|}{(2\pi)^{d-1}}$.

Proof: Assume without loss of generality that $q = e_1$, the unit vector along the first coordinate axis. Consider a field of the form $f(x) = \text{sinc}(\epsilon x_1) g(x_2, x_3, \ldots, x_d)$ and g is bandlimited to a closed set Ω^g where $\Omega^g \subset \text{relint}(\Omega \cap q^\perp)$. For ϵ small enough, $f \in \mathcal{B}_\Omega$. For stably recovering f from samples on Q, the non-uniform collection of points at which the lines in Q intersect the hyperplane $e_1^\perp$ must form a set of stable sampling for Ω^g. We know from Landau's result [2] (see also [4, Cor. 1]) that the sampling density of such a set must necessarily be greater than or equal to $\frac{|\Omega^g|}{(2\pi)^{d-1}}$. Thus, by Lemma 3.1 it follows that $\ell(Q) \geq \frac{|\Omega^g|}{(2\pi)^{d-1}}$ for all $\Omega^g \subset \text{relint}(\Omega \cap q^\perp)$. Hence $\ell(Q) \geq \frac{|\Omega \cap q^\perp|}{(2\pi)^{d-1}}$. ∎

Although the result of Proposition 3.2 only provides a lower-bound on the path density, we believe that the techniques used in [13] can be used to construct trajectory sets in $\mathcal{H}_\Omega$ that achieve arbitrarily close to this bound for convex and symmetric Ω. However, in this paper, we only establish the following achievability result, which is tight for some specific choices of Ω as we discuss below.

Proposition 3.3: Let $\Omega \subset \mathbb{R}^d$ be a compact convex set with non-empty interior and a point of symmetry at the origin. Let $\mathcal{S}(\Omega)$ denote the volume of the smallest projection of Ω onto a hyperplane defined as

$$\mathcal{S}(\Omega) := \min_{u \in \mathbb{R}^d : \|u\|=1} |\mathcal{P}_{u^\perp}\Omega|. \tag{8}$$

Let u^* be the minimizer in (8). Then for any $\epsilon > 0$ there exists $P \in \mathcal{H}_\Omega$ such that the lines in P are parallel to u^* and

$$\ell(P) \leq \frac{\mathcal{S}(\Omega)}{(2\pi)^{d-1}} + \epsilon.$$

Sketch of proof: We do not provide a complete proof due to lack of space. The optimal trajectory set is obtained by choosing the lines in P parallel to u^* such that their points of intersection with $(u^*)^\perp$ approximates an optimal set of stable sampling for $\mathcal{P}_{(u^*)^\perp}\Omega$. Such an optimal set can be designed using the results of [13, Cor 4.5]. In this case, the path density of this trajectory set matches the sampling density of the optimal set of sampling which is equal to $|\mathcal{P}_{(u^*)^\perp}\Omega| + \epsilon$. ∎

The following corollary is immediate from the above two results.

Corollary 3.3.1: Let $\Omega \subset \mathbb{R}^d$ be a compact convex set with non-empty interior and a point of symmetry at the origin. Suppose that Ω satisfies the condition

$$\min_{u \in \mathbb{R}^d : \|u\|=1} |\Omega \cap u^\perp| = \mathcal{S}(\Omega). \tag{9}$$

Then

$$\inf_{Q \in \mathcal{H}_\Omega} \ell(Q) = \frac{\mathcal{S}(\Omega)}{(2\pi)^{d-1}}. \qquad \square \tag{10}$$

In words, condition (9) is the requirement that the volume of the smallest section of Ω through the origin is equal to the volume of the smallest projection of Ω onto a hyperplane. This condition holds in the following practically relevant cases:

- $\Omega \subset \mathbb{R}^2$ such that Ω is convex and compact [14, Thm 12.18].
- $\Omega \subset \mathbb{R}^d$ such that Ω is a spherical ball (obvious), or an n-cube [15], or an ellipsoid (can be shown).

However, this condition does not hold in general, a simple counter-example being the regular octahedron in $\mathbb{R}^3$: $\Omega = \{\omega \in \mathbb{R}^3 : \|\omega\|_1 \leq 1\}$. Nevertheless for Ω's that satisfy condition (9), the trajectory set of Proposition 3.3 gives the optimal configuration of parallel lines for sampling fields in $\mathcal{B}_\Omega$. In particular, when Ω is a spherical ball in $\mathbb{R}^d$, the trajectory set of Proposition 3.3 gives the optimal configuration of parallel lines for sampling isotropic fields in $\mathbb{R}^d$. Similarly, for convex and compact sets $\Omega \subset \mathbb{R}^2$, we showed in [9] that the optimum configuration of parallel lines given by Proposition 3.3 is a uniform set in $\mathcal{H}_\Omega$. For general $\mathbb{R}^d$, the result of Proposition 3.3 gives the best known solution to the minimum path density problem of (6). In Section IV we discuss the possibility of extending this result to all of $\mathcal{N}_\Omega$.

IV. Discussion

This work opens up several possible research directions. An obvious question is to check if under the conditions of Proposition 3.3 it is possible to design a trajectory set in $\mathcal{H}_\Omega$ that achieves a path density arbitrarily close to the lower bound. Another direction of interest is to extend the results on parallel lines obtained in this paper to parallel sampling manifolds of higher dimensions, like those considered in [9].

Although we have obtained various optimality results on parallel line trajectories in this paper, our original task of identifying minimal length trajectories for sampling spatial bandlimited fields still remains open. A first case to analyze is the necessary condition on a trajectory set in $\mathcal{N}_\Omega$ composed of arbitrary (not necessarily parallel) straight lines. A generalization of the notion of Fourier frames [4] [5] may be a possible approach towards such a result.

A different question of interest is to examine the definition of $\mathcal{N}_\Omega$. In the current version of this work, while defining the set $\mathcal{N}_\Omega$ we have placed the restriction that a sampling trajectory set in $\mathcal{N}_\Omega$ must contain a uniformly discrete set of points that form a set of stable sampling for Ω. In addition we have the requirement of Condition **(C2)**. Nevertheless, it has recently come to our knowledge that under this definition of $\mathcal{N}_\Omega$ it is possible to design sampling trajectories in $\mathcal{N}_\Omega$ that have arbitrarily small path density. However, this leads to the stability ratio $\frac{B}{A}$ of parameters A and B in the definition of (5) to be arbitrarily high. It is of interest to examine whether a constraint on the ratio $\frac{B}{A}$ can be incorporated in the definition of $\mathcal{N}_\Omega$ to obtain a non-trivial lower bound on the path density of all trajectory sets in $\mathcal{N}_\Omega$. However, it is to be noted that if we restrict ourselves to trajectory sets in $\mathcal{H}_\Omega$, then the problem is still well-posed as evidenced by Proposition 3.2. It would be of interest to examine whether such a non-trivial lower bound on the path density continues to hold if we expand $\mathcal{H}_\Omega$ to all trajectory sets composed of straight lines.

ACKNOWLEDGMENT

This research was supported by ERC Advanced Investigators Grant: Sparse Sampling: Theory, Algorithms and Applications SPARSAM no 247006. We thank Keith Ball, Karlheinz Gröchenig, and José Luis Romero for helpful discussions.

APPENDIX

A. Proof of Lemma 3.1

For simplicity, we prove the result only for $d = 2$, since the same proof idea works for higher dimensions. Without loss of generality assume that the lines in P are parallel to e_2. Since the lines are homogenous we just need to evaluate (4) when x is the origin. We number the lines in P such that for each $i \in \mathbb{Z}^+(\mathbb{Z}^-)$, $\ell_{a,i}$ denotes the length of the portion of the i-th line to the right (left) of the origin that is contained within a disc of radius a centered at the origin. Without affecting the value of the computation we assume that the line indexed by 0 passes through the origin. Let $d_i = \sum_{j=0}^{i} \Delta_j$ where Δ_j denotes the spacing between lines indexed by j and $j+1$. Now let $I_{a,f} = \{i \in \mathbb{Z} : f\epsilon a \leq d_i < (f+1)\epsilon a\}$ for $-\frac{1}{\epsilon} \leq f \leq \frac{1}{\epsilon}$. Let $L_{a,f} = \sum_{i \in I_{a,f}} \ell_{a,i}$ and $N_{a,f} = \#(I_{a,f})$. Clearly $\lim_{a\to\infty} \frac{N_{a,f}}{a\epsilon} = \rho$ where ρ is the right hand side expression in (7). Further, for $f \in [0, \frac{1}{\epsilon}]$,

$$2a(1-(f+1)^2\epsilon^2)^{\frac{1}{2}} N_{a,f} \leq L_{a,f} \leq 2a(1-f^2\epsilon^2)^{\frac{1}{2}} N_{a,f}.$$

Hence

$$\frac{2\epsilon\rho}{\pi}(1-(f+1)^2\epsilon^2)^{\frac{1}{2}} \leq \lim_{a\to\infty} \frac{L_{a,f}}{\pi a^2} \leq \frac{2\epsilon\rho}{\pi}(1-f^2\epsilon^2)^{\frac{1}{2}}.$$

For $f < 0$ the above relation holds with the signs reversed. Thus we see that $\sum_{f=0}^{\frac{1}{\epsilon}} \lim_{a\to\infty} \frac{L_{a,f}}{\pi a^2}$ is bounded between the right hand and left hand Riemann sums that approximate the Riemann integral $\int_0^1 \frac{2\rho}{\pi}(1-x^2)^{\frac{1}{2}} dx$. Since this holds for all ϵ it follows that as we let $\epsilon \to 0$, we get $\lim_{a\to\infty} \frac{\sum_{i\in\mathbb{Z}^+} \ell_{a,i}}{\pi a^2} = \int_0^1 \frac{2\rho}{\pi}(1-x^2)^{\frac{1}{2}} dx$. Following the same steps for negative indices and combining, we get

$$\lim_{a\to\infty} \frac{\sum_{i\in\mathbb{Z}} \ell_{a,i}}{\pi a^2} = \int_{-1}^{1} \frac{2\rho}{\pi}(1-x^2)^{\frac{1}{2}} dx = \rho.$$

□

REFERENCES

[1] D. P. Petersen and D. Middleton, "Sampling and Reconstruction of Wave-Number-Limited Functions in N-Dimensional Euclidean Spaces," *Inform. Contr.*, vol. 5, pp. 279–323, 1962.

[2] H. Landau, "Necessary density conditions for sampling and interpolation of certain entire functions," *Acta Mathematica*, vol. 117, no. 1, pp. 37–52, Jul. 1967. [Online]. Available: http://dx.doi.org/10.1007/BF02395039

[3] ——, "Sampling, data transmission, and the Nyquist rate," *Proceedings of the IEEE*, vol. 55, no. 10, pp. 1701 – 1706, Oct. 1967.

[4] K. Gröchenig and H. Razafinjatovo, "On Landau's necessary density conditions for sampling and interpolation of band-limited functions," *Journal of the London Mathematical Society*, vol. 54, no. 3, pp. 557–565, 1996. [Online]. Available: http://jlms.oxfordjournals.org/content/54/3/557.abstract

[5] J. J. Benedetto and H.-C. Wu, "Nonuniform sampling and spiral MRI reconstruction," in *Proc. SPIE Symp. Wavelets Applications in Signal and Image Processing VIII*, A. L. A. Aldroubi and M. Unser, Eds., vol. 4119, June 2000, pp. 130–141.

[6] A. Beurling, "On balayage of measures in Fourier transforms (seminar, inst. for advanced studies, 1959-60, unpublished)," in *Collected Works of Arne Beurling*, L. Carleson, P. Malliavin, J. Neuberger, and J. Wermer, Eds. Boston: Birkhauser, 1989.

[7] A. Aldroubi and K. Gröchenig, "Nonuniform sampling and reconstruction in shift-invariant spaces," *SIAM Rev.*, vol. 43, no. 4, pp. 585–620, Apr. 2001. [Online]. Available: http://dx.doi.org/10.1137/S0036144501386986

[8] J. Unnikrishnan and M. Vetterli, "Sampling trajectories for mobile sensing," in *Proc. 2011 49th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Allerton House, UIUC, Illinois, USA, Sept. 2011, pp. 1230–1237.

[9] ——, "Sampling high-dimensional bandlimited fields on low-dimensional manifolds," *Information Theory, IEEE Transactions on*, vol. 59, no. 4, pp. 2103–2127, 2013.

[10] ——, "Sampling and reconstruction of spatial fields using mobile sensors," *Signal Processing, IEEE Transactions on*, vol. 61, no. 9, pp. 2328–2340, 2013.

[11] A. Singh, R. Nowak, and P. Ramanathan, "Active learning for adaptive mobile sensing networks," in *Proceedings of Information Processing in Sensor Networks (IPSN), 2006*, 2006.

[12] J. Unnikrishnan and M. Vetterli, "On sampling a high-dimensional bandlimited field on a union of shifted lattices," in *Information Theory Proceedings (ISIT), 2012 IEEE International Symposium on*, July 2012, pp. 1468 –1472.

[13] J. Marzo, "Sampling sequences in spaces of bandlimited functions in several variables,," Ph.D. dissertation, Universitat de Barcelona, April 2008.

[14] F. A. Valentine, *Convex Sets*. New York: McGraw-Hill, 1964.

[15] J. D. Vaaler, "A geometric inequality with applications to linear forms," *Pacific J. Math.*, vol. 83, no. 2, pp. 543–553, 1979.

Phase Retrieval via Structured Modulations in Paley-Wiener Spaces

Fanny Yang, Volker Pohl, Holger Boche
Lehrstuhl für Theoretische Informationstechnik
Technische Universität München, 80290 München, Germany
{fanny.yang, volker.pohl, boche}@tum.de

***Abstract*—This paper considers the recovery of continuous time signals from the magnitude of its samples. It uses a combination of structured modulation and oversampling and provides sufficient conditions on the signal and the sampling system such that signal recovery is possible. In particular, it is shown that an average sampling rate of four times the Nyquist rate is sufficient to reconstruct almost every signal from its magnitude measurements.**

***Index Terms*—Bernstein spaces, Paley-Wiener spaces, phase retrieval, sampling**

I. Introduction

In many applications, only intensity measurements are available to reconstruct a desired signal x. This is widely known as the phase retrieval problem which for example occurs in diffraction imaging applications such as X-ray crystallography, astronomical imaging or speech processing.

In the past, several efforts have been made on the recovery of finite n-dimensional signals from the modulus of their Fourier transform. In general however, they require strong limitations on the signal such as constraints on its z-transform [1] or knowledge of its support [2]. Analytic frame-theoretic approaches were considered in [3], [4] and an algorithm was presented which requires that the number of measurements grows proportionally with the square of the space dimension. Ideas of sparse signal representation and convex optimization where applied in [5], [6] to allow for lower computational complexity. Recently in [7], results in the context of entire functions theory have derived a sampling rate of $4n - 4$.

Note that all of the above approaches addressed finite dimensional signals and the question is whether similar results can be obtained for continuous signals in infinite dimensional spaces. In [8] it was shown that real valued bandlimited signals are completely determined by their magnitude samples taken at twice the Nyquist rate. In the present work we are looking at complex valued continuous signals in Paley-Wiener spaces. Our approach extends ideas from [3], [4], [6] and involves two steps: first we apply a bank of modulators to the signal and sample the subsequent intensity measurements in the Fourier domain. In this step, finite blocks of intensity samples are obtained and a finite dimensional algorithm from [4] can be used to recover the complex signal samples up to a constant phase. Secondly, by ensuring an overlap between subsequent blocks, the unimodular factor in all blocks is matched and well-known interpolation theorems and the inverse Fourier transform are used to obtain the time signal. Therewith we are able to reconstruct the infinite dimensional signals from samples taken at a rate of four times the Nyquist rate, which asymptotically coincides with the value for the finite dimensional case in [4].

This work was partly supported by the German Research Foundation (DFG) under Grant BO 1734/22-1.

Basic notations for sampling and reconstruction in Paley-Wiener spaces are recaptured in Sec. II, Sec. III describes our sampling setup. In Sec. IV we provide sufficient conditions for perfect signal reconstruction from magnitude measurements of the Fourier transform. The paper closes with a short discussion in Sec.V.

II. Sampling in Paley-Wiener Spaces

Let $\mathbb{S} \subseteq \mathbb{R}$ be an arbitrary subset of the real axis $\mathbb{R}$. For $1 \leq p \leq \infty$ we write $\mathcal{L}^p(\mathbb{S})$ for the usual Lebesgue space on $\mathbb{S}$. In particular, $\mathcal{L}^2(\mathbb{S})$ is the Hilbert space of square integrable functions on $\mathbb{S}$ with the inner product

$$\langle x, y\rangle_{\mathcal{L}^2(\mathbb{S})} = \int_{\mathbb{S}} x(\theta)\,\overline{y(\theta)}\,\mathrm{d}\theta\,,$$

where the bar denotes the complex conjugate. In finite dimensional spaces $\langle x, y\rangle = y^* x$ where * denotes the conjugate transpose. Let $T > 0$ be a real number. Throughout this paper $\mathbb{T} = [-T/2, T/2]$ stands for the closed interval of length T, and $\mathcal{PW}_{T/2}$ denotes the *Paley-Wiener space* of entire functions of exponential type $T/2$ whose restriction to $\mathbb{R}$ belongs to $\mathcal{L}^2(\mathbb{R})$. The Paley-Wiener theorem states that to every $\widehat{x} \in \mathcal{PW}_{T/2}$ there is an $x \in \mathcal{L}^2(\mathbb{T})$ such that

$$\widehat{x}(z) = \int_{\mathbb{T}} x(t)\,\mathrm{e}^{\mathrm{i}tz}\,\mathrm{d}t \quad \text{for all } z \in \mathbb{C}\,, \tag{1}$$

and vice versa. If not otherwise noted, our signal space will be $\mathcal{L}^2(\mathbb{T})$, i.e. we consider signals of finite energy which are supported on the finite interval $\mathbb{T}$. These are natural assumptions for signals in reality. In the following we will call x the signal in the *time domain* and $\widehat{x}$ the signal in the *Fourier domain*, since its restriction to the real axis is a Fourier transform.

A sequence $\Lambda = \{\lambda_n\}_{n\in\mathbb{Z}}$ of complex numbers is said to be *complete interpolating* for $\mathcal{PW}_{T/2}$ if and only if

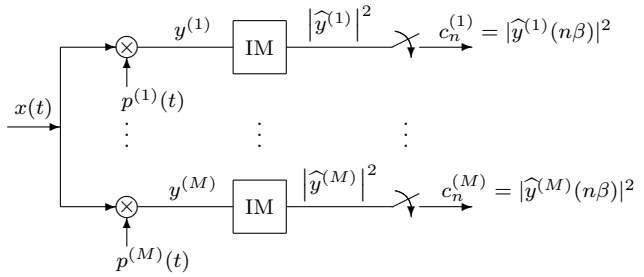


Fig. 1. Measurement setup: In each branch, the unknown signal x is modulated with a different sequence $p^{(m)}$, $m = 1, 2, \ldots, M$. Subsequently, the intensities of the resulting signals $y^{(m)}$ are measured (IM) and uniformly sampled in the frequency domain.

the functions $\{\phi_n(t) := \mathrm{e}^{-\mathrm{i}\lambda_n t}\}_{n\in\mathbb{Z}}$ form a Riesz basis for $\mathcal{L}^2(\mathbb{T})$ [9]. Let $x \in \mathcal{L}^2(\mathbb{T})$ be arbitrary. Then (1) shows that

$$\widehat{x}(\lambda_n) = \langle x, \phi_n \rangle_{\mathcal{L}^2(\mathbb{T})} \quad \text{for all } n \in \mathbb{Z}\,.$$

Since $\{\phi_n\}_{n\in\mathbb{Z}}$ is a Riesz basis for $\mathcal{L}^2(\mathbb{T})$ the signal x can be reconstructed from the samples $\widehat{x}(\Lambda) = \{\widehat{x}(\lambda_n)\}_{n\in\mathbb{Z}}$ by

$$x(t) = \textstyle\sum_{n\in\mathbb{Z}} \langle x, \phi_n \rangle\ \psi_n(t) = \sum_{n\in\mathbb{Z}} \widehat{x}(\lambda_n)\, \psi_n(t)\,, \quad (2)$$

where $\{\psi_n\}_{n\in\mathbb{Z}}$ is the unique dual Riesz basis of $\{\phi_n\}_{n\in\mathbb{Z}}$ [10]. It is well-known that in the Fourier domain

$$\widehat{\psi}_n(z) = \frac{S(z)}{S'(\lambda_n)(z-\lambda_n)} \text{ with } S(z) = z^{\delta_\Lambda} \lim_{R\to\infty} \prod_{\substack{|\lambda_n|<R \\ \lambda_n\neq 0}} \left(1 - \frac{z}{\lambda_n}\right)$$

with $\delta_\Lambda = 1$ if $0 \in \Lambda$ and $\delta_\Lambda = 0$ otherwise. S is an entire function of exponential type $T/2$, and the infinite product converges uniformly on compact subsets of $\mathbb{C}$ if Λ is a complete interpolating sequence (see [11]).

Example 1: The well known Shannon sampling series is obtained for regular sampling with $\lambda_n = n\frac{2\pi}{T}$, $n \in \mathbb{Z}$. Then $S(z) = \sin(\frac{T}{2}z)$ and $\widehat{\psi}_n(z) = \operatorname{sinc}(\frac{T}{2}[z - n\frac{2\pi}{T}])$ where $\operatorname{sinc}(x) := \sin(x)/x$. This corresponds to $x(t) = \sum_{n\in\mathbb{Z}} \widehat{x}(\lambda_n)\, \mathrm{e}^{-\mathrm{i}n\frac{2\pi}{T}t} \mathbb{1}_{\mathbb{T}}(t)$ in the time domain, where $\mathbb{1}_{\mathbb{T}}(t)$ denotes the indicator function on $\mathbb{T}$.

III. Measurement Methodology

We apply a measurement methodology which uses oversampling in connection with structured modulations of the desired signal, inspired by the approach in [6]. Suppose $x \in \mathcal{L}^2(\mathbb{T})$ is the signal of interest. In our sampling scheme in Fig. 1, we assume that x is multiplied with M known modulating functions $p^{(m)}$. In optics, these modulations may be different diffraction gratings between the object (the desired signal) and the measurement device [6]. This way we obtain a collection of M representations (or illuminations) $y^{(m)}$ of x. Afterwards, the modulus of the Fourier spectra $\widehat{y}^{(m)}$ are measured and uniformly sampled with frequency spacing β.

Let $p^{(m)}$ have the following general form

$$p^{(m)}(t) := \textstyle\sum_{k=1}^{K} \overline{\alpha_k^{(m)}} \mathrm{e}^{\mathrm{i}\lambda_k t} \quad (3)$$

where λ_k and $\alpha_k^{(m)}$ are complex coefficients. The samples in the mth branch are then given by

$$\begin{aligned} c_n^{(m)} = |\widehat{y}^{(m)}(n\beta)|^2 &= \left| \sum_{k=1}^{K} \overline{\alpha_k^{(m)}}\, \widehat{x}(n\beta + \lambda_k) \right|^2 \\ &= |\langle \widehat{\mathbf{x}}_n, \boldsymbol{\alpha}^{(m)} \rangle|^2 \end{aligned} \quad (4)$$

with the length K vectors

$$\boldsymbol{\alpha}^{(m)} := \begin{pmatrix} \alpha_1^{(m)} \\ \vdots \\ \alpha_K^{(m)} \end{pmatrix} \text{ and } \widehat{\mathbf{x}}_n := \begin{pmatrix} \widehat{x}(n\beta + \lambda_1) \\ \vdots \\ \widehat{x}(n\beta + \lambda_K) \end{pmatrix}.$$

We will show that if $\boldsymbol{\alpha}^{(m)}$ and the interpolation points $\{\lambda_{n,k} := n\beta + \lambda_k\}_{n\in\mathbb{Z}}^{k=1,\ldots,K}$ are properly chosen, it is possible to reconstruct x from all samples $\mathbf{c} = \{c_n^{(m)}\}_{n\in\mathbb{Z}}^{m=1,\ldots,M}$.

A. Choice of the coefficients $\alpha_k^{(m)}$

The first recovery step determines the vector $\widehat{\mathbf{x}}_n \in \mathbb{C}^K$ from the M intensity measurements $c_n^{(m)}$ for every $n \in \mathbb{Z}$, using a result from [4]. It states that if the family of $\mathbb{C}^K$-vectors $\mathcal{A} = \{\boldsymbol{\alpha}^{(1)}, \ldots, \boldsymbol{\alpha}^{(M)}\}$ constitutes a 2-uniform M/K-tight frame which contains $M = K^2$ vectors or $\mathcal{A}$ is a union of $K+1$ mutually unbiased bases in $\mathbb{C}^K$, then every $\widehat{\mathbf{x}}_n \in \mathbb{C}^K$ can be reconstructed up to a constant phase from the magnitude of the inner products (4). For simplicity, we only discuss the first case here and therefore fix $M = K^2$.

Condition A: *A sampling system as in Fig. 1 is said to satisfy Condition A if $\mathcal{A}$ constitutes a 2-uniform M/K-tight frame.*

Then reconstruction will be based on the following formula

$$Q_{\widehat{\mathbf{x}}_n} = \frac{(K+1)}{K} \sum_{m=1}^{M} c_n^{(m)} Q_{\boldsymbol{\alpha}^{(m)}} - \frac{1}{K} \sum_{m=1}^{M} c_n^{(m)} I \quad (5)$$

with rank-1 matrices $Q_{\mathbf{x}} = \mathbf{x}\mathbf{x}^*$. For $K = 2$ a valid choice for $\mathcal{A}$ reads [4]

$$\boldsymbol{\alpha}^{(1)} = \begin{pmatrix} a \\ b \end{pmatrix},\ \boldsymbol{\alpha}^{(2)} = \begin{pmatrix} b \\ a \end{pmatrix},\ \boldsymbol{\alpha}^{(3)} = \begin{pmatrix} a \\ -b \end{pmatrix},\ \boldsymbol{\alpha}^{(4)} = \begin{pmatrix} -b \\ a \end{pmatrix}$$

with $a = \sqrt{\frac{1}{2}(1 - \frac{1}{\sqrt{3}})}$ and $b = \mathrm{e}^{\mathrm{i}5\pi/4}\sqrt{\frac{1}{2}(1 + \frac{1}{\sqrt{3}})}$.

B. Choice of the interpolation points

Now it is necessary to find conditions which allow unique interpolation from the known samples. Let $\{\lambda_k\}_{k=1}^K$ be ordered increasingly by their real parts. For each $n \in \mathbb{Z}$, the vector $\widehat{\mathbf{x}}_n$ contains the values of $\widehat{x}$ at K distinct interpolation points in the complex plane

$$\boldsymbol{\lambda}_n^a := \{\lambda_{n,k}^a\}_{k=1}^K \quad \text{with} \quad \lambda_{n,k}^a = n\beta + \lambda_k\,, \quad n \in \mathbb{Z}\,. \quad (6)$$

Therein, the parameter $a \in \mathbb{N}$ denotes the number of overlapping points of consecutive sets (6) (cf. also Fig.2). More precisely, we require for every $n \in \mathbb{Z}$ that

$$\lambda_{n,i}^a = \lambda_{n-1,K-i+1}^a \quad \text{for all } i = 1, \ldots, a\,. \quad (7)$$

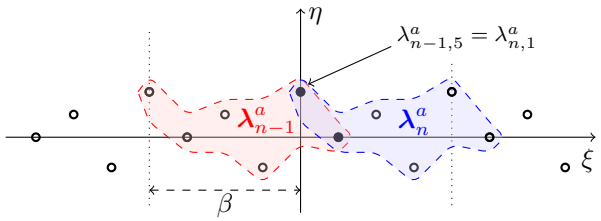


Fig. 2. Illustration for the choice of interpolation points in the complex plane for $K = 6$ in (3) and an overlap $a = 2$.

In the following $\Lambda^a_{O,n} = \boldsymbol{\lambda}^a_n \cap \boldsymbol{\lambda}^a_{n+1}$ is the set of overlapping interpolation points between $\boldsymbol{\lambda}^a_n$ and $\boldsymbol{\lambda}^a_{n+1}$, and we define the overall interpolation sequence

$$\Lambda^a := \bigcup\nolimits_{n\in\mathbb{Z}} \boldsymbol{\lambda}^a_n \,.$$

In general we allow for $a \geq 1$, but we will see that $a = 1$ is generally sufficient for reconstruction.

As explained in Sec. II, $x \in \mathcal{L}^2(\mathbb{T})$ can be perfectly reconstructed by (2) if Λ^a is complete interpolating for $\mathcal{PW}_{T/2}$. This gives a second condition on our sampling system:

Condition B: *A sampling system as in Fig. 1 is said to satisfy Condition B if the coefficients $\{\lambda_k\}_{k=1}^K$ in (3) are such that Λ^a is complete interpolating for $\mathcal{PW}_{T/2}$ and satisfies (7) for a certain $1 \leq a < K$.*

In general it is hard to characterize sets which fulfill this condition. One famous example is the set of zeros of a sine-type function of type $\tilde{T}/2 \geq T/2$ which is β-periodic (see, e.g., [9], [11]). Such sine-type functions are entire functions f of exponential type $\tilde{T}/2$ with simple and isolated zeros and for which there exist positive constants A, B, H such that

$$A\,\mathrm{e}^{\frac{\tilde{T}}{2}|\eta|} \leq |f(\xi + \mathrm{i}\eta)| \leq B\,\mathrm{e}^{\frac{\tilde{T}}{2}|\eta|}\,, \quad \text{for } |\eta| \geq H\,.$$

Note that $\sin(\frac{\tilde{T}}{2}z)$ is a trivial example for a sine-type function (cf. Example 1). Moreover, shifting the zeros of one sine-type functions arbitrarily in their imaginary parts yields the zero set of another sine-type function [12]. The complete interpolating property is also preserved under small shifts in the real part (see Katsnelson's theorem, e.g. in [11]).

IV. Phaseless Signal Recovery

We assume a sampling scheme as described in Section III which satisfies Condition A and B. For this setup, we show that almost every $x \in \mathcal{L}^2(\mathbb{T})$ (up to a set of first category) can be reconstructed from the samples (4). The proof provides an explicit algorithm for perfect signal recovery.

Theorem 1: *Let $x \in \mathcal{L}^2(\mathbb{T})$ be sampled according to the scheme in Section III which satisfies Condition A and B, and let $\mathbf{c} = \{c_n^{(m)}\}_{n\in\mathbb{Z}}^{m=1,\dots,M}$ be the sampling sequence in (4). If the set $\widehat{x}(\Lambda^a_{O,n})$ contains at least one non-zero element for each $n \in \mathbb{Z}$, then x can be perfectly reconstructed from $\mathbf{c}$ up to a constant phase.*

Proof: According to Condition B of the sampling system, Λ^a is complete interpolating for $\mathcal{PW}_{T/2}$. Therefore the signal x can be reconstructed from the vectors $\{\widehat{\mathbf{x}}_n\}_{n\in\mathbb{Z}}$ using (2). It remains to show that $\{\widehat{\mathbf{x}}_n\}_{n\in\mathbb{Z}}$ can be determined from $\boldsymbol{c}$.

Let $n \in \mathbb{Z}$ be arbitrary. Since the sampling system satisfies Condition A, we can use (5) to obtain the rank-1 matrix $Q_n := \widehat{\mathbf{x}}_n \widehat{\mathbf{x}}_n^*$ from the measurements $\{c_n^{(m)}\}_{m=1}^M$. Then $\widehat{\mathbf{x}}_n \in \mathbb{C}^K$ is obtained by factorizing Q_n. However, such a factorization is only unique up to a constant phase factor. If the phase $\phi_{n,i}$ of one element $[\widehat{\mathbf{x}}_n]_i$ is known, the vector $\widehat{\mathbf{x}}_n$ can be completely determined from Q_n by

$$\widehat{x}(n\beta + \lambda_k) = \sqrt{[Q_n]_{k,k}}\,\mathrm{e}^{\mathrm{i}(\phi_{n,i} - \arg([Q_n]_{i,k}))}\,,\ \forall k \neq i\,. \quad (8)$$

Assume that we start the recovery of the sequence $\{\widehat{\mathbf{x}}_n\}_{n\in\mathbb{Z}}$ at a certain $n_0 \in \mathbb{Z}$ and set the constant phase of $\widehat{\mathbf{x}}_{n_0}$ arbitrarily to $\theta_0 \in [-\pi, \pi]$. In the next step, we determine $\widehat{\mathbf{x}}_{n_0+1}$. After the factorization of Q_{n_0+1}, we use the nonempty overlap to carry over the phase from n_0 to $n_0 + 1$. Since by assumption the overlapping point, say $\lambda^a_{n_0+1,i}$, can be chosen such that it is non-zero, the propagation of the constant phase can be ensured. Thus, we can completely determine $\widehat{\mathbf{x}}_{n_0+1}$ and successively all $n = n_0 \pm 1, n_0 \pm 2, \dots$ using (8) to obtain $\widehat{x}(\Lambda^a)\,\mathrm{e}^{\mathrm{i}\theta_0}$. The arbitrary setting of the phase of the initial vector $\widehat{\mathbf{x}}_{n_0}$ yields a constant phase shift θ_0 for all $\widehat{\mathbf{x}}_n$ which persists after the reconstruction of the time signal as in (2). ■

Theorem 1 states that $x \in \mathcal{L}^2(\mathbb{T})$ can only be reconstructed if $\widehat{x} \in \mathcal{PW}_{T/2}$ has at most $a - 1$ zeros on the overlapping interpolation sets $\Lambda^a_{O,n}$. However, this restriction is not too limiting. On the one hand, it is not hard to see that the subset of all $x \in \mathcal{L}^2(\mathbb{T})$ which does not satisfy this condition is of first category [13]. On the other hand, it is known that the zeros of an entire function of exponential type can not be arbitrarily dense. For example, defining $\mathcal{Z}_n := \{z \in \mathbb{C} : n\pi/T < |z| \leq (n+1)\pi/T\}$, the result in [14] states that for every $\widehat{x} \in \mathcal{PW}_{T/2}$ there exist only finitely many sets $\mathcal{Z}_n$ which contain more than one zero of $\widehat{x}$. Consequently, choosing the spacing of the interpolation points in the overlapping sets $\Lambda^a_{O,n}$ less than π/T, it is very unlikely that a randomly chosen function from $\mathcal{PW}_{T/2}$ fails to satisfy the condition of Theorem 1, especially for $a > 1$.

When the overall energy of the signal is known, even such pathological cases can be avoided such that the last condition in Theorem 1 always holds true. To this end, we first state a simple variant of a lemma by Duffin, Schaeffer [15].

Lemma 2: *Let $\widehat{x}(z) \in \mathcal{PW}_{T/2}$ be an entire function of $z = \xi + \mathrm{i}\eta$ satisfying $|\widehat{x}(\xi)| \leq M$ on the real axis. Then for every $T' > T$ the function*

$$\widehat{v}(z) = M\cos(\tfrac{T'}{2}z) - \widehat{x}(z) \quad (9)$$

belongs to the Bernstein space $\mathcal{B}^\infty_{T'/2}$ *and there exists a constant* $H = H(T, T')$ *such that* $|\widehat{v}(z)| > 0\ \forall z : |\eta| > H$.

A proof can be found in [13]. The Bernstein space $\mathcal{B}^\infty_{T'/2}$ is the set of all entire functions of exponential type $T'/2$ whose restriction to $\mathbb{R}$ is in $\mathcal{L}^\infty(\mathbb{R})$. Upon this we can establish a corollary for signals which have a known maximal energy W_0.

Corollary 3: *Let* $x \in \mathcal{L}^2(\mathbb{T}) : \|x\|_{\mathcal{L}^2(\mathbb{T})} \leq W_0$ *be sampled according to the scheme in Sec. III. Then there exist interpolation sequences* Λ^a *with overlap* $a \geq 1$ *such that every* x *can be perfectly reconstructed (up to a constant phase) from the measurements* (4).

Sketch of proof: The theorem of Plancherel-Pólya implies that there exists a constant M independent of x such that $|\widehat{x}(\xi)| \leq MW_0$ for all $\xi \in \mathbb{R}$. Using $T' > T$ we can define $\widehat{v}$ by (9) which only has zeros for $|\eta| \leq H$ by Lemma 2. In the measurement scheme this corresponds to adding a cosine to the signal. Subsequently, the function $\widehat{v}$ is modulated and sampled at interpolation points Λ^a, which we choose as the zero set of a sine-type function of type $\tilde{T}/2 > T'/2$. By [12] we can shift the imaginary parts of the interpolation points such that $|\eta_k| > H$ for all k while Λ^a remains to be the zero set of a sine-type function denoted by S. Since $\widehat{v} \in \mathcal{B}^\infty_{T'/2}$ and Λ^a is the set of zeros of a sine-type function, the sequence $\{d_n = \widehat{v}(\lambda_n)\,\mathrm{e}^{\mathrm{i}\theta_0}\}_{n\in\mathbb{Z}}$ is in ℓ^∞, and we apply a generalization of [11, Lec. 21] (see [13]) to reconstruct $\widehat{v}$ from the sequence $\{d_n\}_{n\in\mathbb{Z}}$ by

$$\widehat{v}(z)\,\mathrm{e}^{\mathrm{i}\theta_0} = \textstyle\sum_{n\in\mathbb{Z}} d_n \frac{S(z)}{S'(\lambda_n)} \left[\frac{1}{z-\lambda_n} + \frac{1}{\lambda_n}\right],$$

where the second term in the sum is omitted when $\lambda_n = 0$. Since θ_0 is unknown, we can only obtain

$$\begin{aligned}\tilde{x}(z) &= MW_0 \cos\left(\tfrac{T'}{2}z\right) - \widehat{v}(z)\,\mathrm{e}^{\mathrm{i}\theta_0} \\ &= \widehat{x}(z)\,\mathrm{e}^{\mathrm{i}\theta_0} + MW_0 \cos\left(\tfrac{T'}{2}z\right)(1 - \mathrm{e}^{\mathrm{i}\theta_0}).\end{aligned}$$

However, applying the inverse Fourier transform yields $x(t)\,\mathrm{e}^{\mathrm{i}\theta_0}$ for $t \in \mathbb{T}$ which is the desired signal up to a constant phase since the distributional Fourier transform of a cosine vanishes within $\mathbb{T}$. ∎

V. Discussion and Outlook

To determine the sampling system in Fig.1, one has to fix K, M, a and β. The number $K \geq 2$ can be chosen arbitrarily. Then $M = K^2$ is fixed, and $1 \leq a \leq K-1$. The sampling period β has to be chosen such that the sampling system satisfies Condition B and in particular that Λ^a is complete interpolating for $\mathcal{PW}_{T/2}$. As discussed before, one possible choice could be the zeros of the function $\sin(\frac{\tilde{T}}{2}z)$ with $\tilde{T} > T' > T$. Then $\delta := \lambda_k - \lambda_{k-1} = 2\pi/\tilde{T}$ such that $\beta = (K-a)\,\delta$, and the total sampling rate becomes

$$R(a, K, \tilde{T}) = \frac{M}{\beta} = \frac{K^2}{(K-a)\,\delta} = \frac{K^2}{K-a}\frac{\tilde{T}}{2\pi} = \frac{K^2}{K-a}\frac{\tilde{T}}{T}R_{\mathrm{Ny}}$$

where $R_{\mathrm{Ny}} := T/(2\pi)$ is the Nyquist rate. It is apparent that $R(a, K, \tilde{T})$ grows asymptotically proportional with K and increases with the overlap a. $R(a, K, \tilde{T})$ is bounded below by

$$\inf_{\substack{1\leq a<K,\\ K\geq 1, \tilde{T}>T}} R(a, K, \tilde{T}) = \inf_{\tilde{T}>T} R(1, 2, \tilde{T}) = 4R_{\mathrm{Ny}}\,.$$

Since $\tilde{T}/T$ can be made arbitrarily close to 1 using Theorem 1 and Corollary 3, we can sample at a rate which is almost as small as $4R_{\mathrm{Ny}}$ while still ensuring perfect reconstruction. This corresponds to the findings in [3] for finite dimensional spaces, where it was shown that basically any $x \in \mathbb{C}^N$ can be reconstructed from $M \geq 4N - 2$ magnitude samples.

We note that the above framework can be applied exactly the same way for bandlimited signals. To this end, one only has to exchange the time and frequency domain. Then the modulators in Fig. 1 have to be replaced by linear filters and the sampling of the magnitudes has to be done in the time domain. In future works, our approach will be extended to larger signal spaces [13] and the influence of sampling errors will be investigated in detail.

References

[1] M. H. Hayes, J. S. Lim, and A. V. Oppenheim, "Signal reconstruction from phase or magnitude," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 6, pp. 672–680, Dec. 1980.

[2] J. R. Fienup, "Phase retrieval algorithms: a comparison," *Applied Optics*, vol. 21, no. 15, pp. 2758–2769, Aug. 1982.

[3] R. Balan, P. G. Casazza, and D. Edidin, "On signal reconstruction without phase," *Appl. Comput. Harmon. Anal.*, vol. 20, no. 3, pp. 345–356, May 2006.

[4] R. Balan, B. G. Bodmann, P. G. Casazza, and D. Edidin, "Painless reconstruction from magnitudes of frame coefficients," *J. Fourier Anal. Appl.*, vol. 15, no. 4, pp. 488–501, Aug. 2009.

[5] Y. M. Lu and M. Vetterli, "Sparse spectral factorization: unicity and reconstruction algorithms," in *Proc. 36th Intern. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Prague, Czech Republic, May 2011, pp. 5976–5979.

[6] E. J. Candès, Y. C. Eldar, T. Strohmer, and V. Voroninski, "Phase retrieval via matrix completion," *SIAM J. Imaging Sci.*, vol. 6, no. 1, pp. 199–225, 2013.

[7] B. G. Bodmann and N. Hammen, "Stable phase retrieval with low-redundancy frames," *preprint, arXiv:1302.5487*, Feb. 2013.

[8] G. Thakur, "Reconstruction of bandlimited functions from unsigned samples," *J. Fourier Anal. Appl.*, vol. 17, no. 4, pp. 720–732, Aug. 2011.

[9] R. M. Young, *An introduction to nonharmonic Fourier series.* Cambridge: Academic Press, 2001.

[10] O. Christensen, *An introduction to frames and Riesz bases.* Boston: Birkhäuser, 2003.

[11] B. Y. Levin, *Lectures on entire functions.* Providence, RI: American Mathematical Society, 1997.

[12] B. Y. Levin and I. V. Ostrovskii, "Small perturbations of the set of roots of sine-type functions," *Izv. Akad. Nauk SSSR Ser. Mat*, vol. 43, no. 1, pp. 87–110, 1979.

[13] V. Pohl, F. Yang, and H. Boche, "Phase retrieval of signals with finite support using structured modulations," *preprint, arXiv:1305.2789*, May 2013.

[14] R. Supper, "Zeros of entire functions of finite order," *J. Inequal. Appl.*, vol. 7, no. 1, pp. 49–60, 2002.

[15] R. Duffin and A. C. Schaeffer, "Some properties of functions of exponential type," *Bull. Amer. Math. Soc.*, vol. 44, pp. 236–240, Apr. 1938.

Joint reconstruction of misaligned images from incomplete measurements for cardiac MRI

Gilles Puy*, Gabriele Bonanno†‡, Matthias Stuber†‡, and Pierre Vandergheynst*
* Institute of Electrical Engineering, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
† Department of Radiology, University Hospital (CHUV) and University of Lausanne, Lausanne, Switzerland
‡ Center for Biomedical Imaging (CIBM), Lausanne, Switzerland

***Abstract*—We present a novel method for robust reconstruction of the image of a moving object from incomplete linear measurements. We assume that only few measurements of this object can be acquired at different instants and model the correlation between measurements using global geometric transformations represented by few parameters. Then, we design a method that is able to jointly estimate these transformation parameters and an image of the object, while taking into account possible occlusions of parts of the object during the acquisitions. The reconstruction algorithm minimizes a non-convex functional and generates a sequence of estimates converging to a critical point of this functional. Finally, we show how to apply this algorithm on a real cardiac acquisition for free breathing coronary magnetic resonance imaging.**

I. Introduction

We have recently presented a method to reconstruct jointly a set of images, representing a same scene, from few linear multi-view measurements [1]. The correlation between images is modeled using global parametric transformations, such as homographies, and the proposed algorithm accurately estimates the images and the transformation parameters, while being robust to occlusions. We have shown the efficiency of the algorithm for problems such as super-resolution from multiple frames, or compressed sensing, using numerical simulations.

We show here the potential interest of this method for free breathing coronary magnetic resonance imaging (MRI) [2]. In this application, one wants to obtain a single high resolution image of the heart to visualize the coronaries. To reach this goal, one of the major challenge is to properly compensate for the respiratory motion in the image reconstruction process. Indeed, the acquisition speed in MRI is slow and inevitable motion of the heart occur during the acquisition. To suppress motion due to heart contractions, an ECG signal is usually acquired to ensure that the Fourier measurements are taken after a fixed time delay from the beginning of the cardiac cycle. A few measurements are then taken at each cycle during a period of minimum coronary motion (late diastole). Unfortunately, the number of measurements acquired during one cardiac cycle is too small to accurately reconstruct a high resolution image of the heart. One thus has to combine measurements acquired at different cycles to gather enough information. However, it is mandatory to compensate for the respiratory motion occurring between cardiac cycles to be able to visualize high resolution features.

This work was partly funded by the Hasler Foundation (project number 12080).

As shown in [3] for two-dimensional MRI of the right coronary artery, global translations are already sufficient to reach good image quality. In [3], the estimation of the transformation parameters and the image reconstruction are separated into two separate tasks. We show here that the algorithm presented in [1] can be considered as an alternative for joint registration and reconstruction.

Notations: The Euclidean scalar product of $\mathbb{R}^n$ is denoted $\langle \cdot, \cdot \rangle$ and $\| \cdot \|_2$ is the corresponding ℓ_2-norm. The ℓ_1-norm of a vector $\boldsymbol{x} = (x_i)_{1 \leqslant i \leqslant n} \in \mathbb{R}^n$ is defined as $\|\boldsymbol{x}\|_1 = \sum_{i=1}^n |x_i|$. The transpose operator is denoted $\cdot^\intercal$.

II. Joint registration and reconstruction via non-convex optimization

A. Problem formulation

Let $\boldsymbol{y}_1, \dots, \boldsymbol{y}_l \in \mathbb{R}^m$ be l independent linear observations of a moving object represented by the image $\boldsymbol{x}_0 \in \mathbb{R}^n$, $m \leqslant n$. The j^{th} vector $\boldsymbol{y}_j$ contains the measurements of the object when it is at its j^{th} position. We assume that the acquisition speed is faster than the one of the object, so that we can consider that the object is not moving during each acquisition. However, as the object is moving between two different acquisitions, the image $\boldsymbol{x}_0$ undergoes geometric transformations. In this work, we consider that these transformations are not known in advance and need to be estimated from the measurements. For simplicity, we restrict ourselves to global transformations, such as translations or homographies, that can be represented by few parameters $\boldsymbol{\theta}_j \in \mathbb{R}^q$, $j = 1, \dots, l$. We also assume that the transformed images can be well estimated using interpolation matrices $\mathsf{S}(\boldsymbol{\theta}_j) \in \mathbb{R}^{n \times n}$, $j = 1, \dots, l$, built using, e.g., bicubic splines [4]. Then, to handle more realistic acquisitions, we consider possible occlusions of the object and model them using l foreground images $\boldsymbol{x}_1, \dots, \boldsymbol{x}_l \in \mathbb{R}^n$. The image "viewed" at the j^{th} acquisition is thus $\mathsf{S}(\boldsymbol{\theta}_j)\boldsymbol{x}_0 + \boldsymbol{x}_j$. In summary, denoting by $\mathsf{A}_1, \dots, \mathsf{A}_l \in \mathbb{R}^{m \times n}$ the observation matrices, the measurement model satisfies

$$\begin{bmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_l \end{bmatrix} = \begin{bmatrix} \mathsf{A}_1\mathsf{S}(\boldsymbol{\theta}_1) & \mathsf{A}_1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \mathsf{A}_l\mathsf{S}(\boldsymbol{\theta}_l) & 0 & \dots & \mathsf{A}_l \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_0 \\ \vdots \\ \boldsymbol{x}_l \end{bmatrix} + \begin{bmatrix} \boldsymbol{n}_1 \\ \vdots \\ \boldsymbol{n}_l \end{bmatrix}, \quad (1)$$

where $\boldsymbol{n}_1, \dots, \boldsymbol{n}_l \in \mathbb{R}^m$ model additive measurement noise.

Estimating the images $\boldsymbol{x} = (\boldsymbol{x}_0^\intercal, \dots, \boldsymbol{x}_l^\intercal)^\intercal \in \mathbb{R}^{(l+1)n}$ and the transformation parameters $\boldsymbol{\theta}^\intercal = (\boldsymbol{\theta}_1^\intercal, \dots, \boldsymbol{\theta}_l^\intercal)^\intercal \in \mathbb{R}^{lq}$ using the acquired measurements $\boldsymbol{y} = (\boldsymbol{y}_1^\intercal, \dots, \boldsymbol{y}_l^\intercal)^\intercal \in \mathbb{R}^{lm}$ as sole information is an ill-posed inverse problem. Prior information is needed to restrict the set of admissible solutions. Concerning the images, we can for example search for the ones with a sparse decomposition in a wavelet basis by minimizing the ℓ_1-norm of their wavelet coefficients. Alternatively, we can search for piecewise constant images by minimizing their Total Variation norm. For the transformation parameters, we can for example impose that they belong to compact convex sets $\Theta_j = \{\boldsymbol{\theta}_j \in \mathbb{R}^q : \underline{\boldsymbol{\theta}}_j \leqslant \boldsymbol{\theta}_j \leqslant \bar{\boldsymbol{\theta}}_j\}$, $j = 1, \dots, l$, where $\underline{\boldsymbol{\theta}}_j \in \mathbb{R}^q$ and $\bar{\boldsymbol{\theta}}_j \in \mathbb{R}^q$ are pre-defined upper and lower bounds[1]. Therefore, an estimate $\boldsymbol{x}^*$ and $\boldsymbol{\theta}^*$ of the images and the transformations parameters can be obtained by solving

$$\min_{(\boldsymbol{x},\boldsymbol{\theta})} f(\boldsymbol{x}) + \kappa \, \|\mathsf{A}(\boldsymbol{\theta})\, \boldsymbol{x} - \boldsymbol{y}\|_2^2 \;\text{ subjet to }\; \boldsymbol{\theta} \in \Theta, \quad (2)$$

where $f\colon \mathbb{R}^{(l+1)n} \to \mathbb{R} \cup \{+\infty\}$ is a proper lower-semicontinuous convex function, $\kappa^{-1} > 0$ is a regularizing parameter that should be adjusted with the noise level $\|\boldsymbol{n}\|_2$, $\Theta = \{\boldsymbol{\theta} = (\boldsymbol{\theta}_1^\intercal, \dots, \boldsymbol{\theta}_l^\intercal)^\intercal \in \mathbb{R}^{lq} : \boldsymbol{\theta}_j \in \Theta_j,\ j = 1, \dots, l\}$, and

$$\mathsf{A}(\boldsymbol{\theta}) = \begin{bmatrix} \mathsf{A}_1\mathsf{S}(\boldsymbol{\theta}_1) & \mathsf{A}_1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \mathsf{A}_l\mathsf{S}(\boldsymbol{\theta}_l) & 0 & \dots & \mathsf{A}_l \end{bmatrix} \in \mathbb{R}^{lm \times (l+1)n}.$$

Unfortunately, the minimization problem (2) is non-linear in $\boldsymbol{\theta}$ and finding a global minimizer is not trivial. Nevertheless, based on the recent works of Attouch *et al.*, [5], [6], we developed a novel minimization method for problem (2) that produces a convergent sequence to a critical point $(\boldsymbol{x}^*, \boldsymbol{\theta}^*)$ of the functional $L\colon \mathbb{R}^{(l+1)n} \times \mathbb{R}^{lq} \to \mathbb{R} \cup \{+\infty\}$ defined as

$$L(\boldsymbol{x}, \boldsymbol{\theta}) = f(\boldsymbol{x}) + \kappa \, \|\mathsf{A}(\boldsymbol{\theta})\, \boldsymbol{x} - \boldsymbol{y}\|_2^2 + i_\Theta(\boldsymbol{\theta}), \quad (3)$$

where i_Θ is the indicator function[2] of Θ. Note that $(\boldsymbol{x}^*, \boldsymbol{\theta}^*)$ is not necessarily a global minimizer of L but might only be local minimizer or a saddle point of the objective function. The proposed algorithm generates a sequence of estimates $(\boldsymbol{x}^k, \boldsymbol{\theta}^k)_{k\in\mathbb{N}}$ such that $L(\boldsymbol{x}^{k+1}, \boldsymbol{\theta}^{k+1}) \leqslant L(\boldsymbol{x}^k, \boldsymbol{\theta}^k)$, $\forall k \in \mathbb{N}$, and consists of two main steps.

B. First step of the algorithm

Let $(\boldsymbol{x}^k, \boldsymbol{\theta}^k) \in \mathbb{R}^{(l+1)n} \times \Theta$ be the estimates obtained after k iterations of the algorithm. The first step consists in finding a new estimate $\boldsymbol{x}^{k+1} \in \mathbb{R}^{(l+1)n}$ that decreases the value of the objective function L while keeping $\boldsymbol{\theta}^k$ fixed. We choose here this new estimate as a solution of

$$\min_{\boldsymbol{x}\in\mathbb{R}^{(l+1)n}} L(\boldsymbol{x}, \boldsymbol{\theta}^k) + \frac{\lambda_{\boldsymbol{x}}^k}{2}\, h_\mu(\Psi^\intercal(\boldsymbol{x} - \boldsymbol{x}^k)), \quad (4)$$

[1]Let $\bar{\boldsymbol{\theta}} = (\bar{\theta}_i)_{1\leqslant i\leqslant q} \in \mathbb{R}^q$, $\boldsymbol{\theta} = (\theta_i)_{1\leqslant i\leqslant q} \in \mathbb{R}^q$, $\boldsymbol{\theta} \leqslant \bar{\boldsymbol{\theta}}$ means that $\theta_i \leqslant \bar{\theta}_i$ for all $i \in \{1, \dots, q\}$.

[2]The indicator function of a non-empty closed convex set C is the proper lower semicontinuous convex function that satisfies $i_C(\boldsymbol{x}) = 0$ if $\boldsymbol{x} \in C$ and $i_C(\boldsymbol{x}) = +\infty$ otherwise.

where $\lambda_{\boldsymbol{x}}^k > 0$ acts as a stepsize parameter, $\Psi \in \mathbb{R}^{(l+1)n\times(l+1)p}$ is a block-diagonal matrix built by repeating $l+1$ times a wavelet tight-frame[3] $\mathsf{W} \in \mathbb{R}^{n\times p}$, $p \geqslant n$, on the diagonal, and $h_\mu\colon \mathbb{R}^{(l+1)p} \to \mathbb{R}$ is the Huber function. It is a smooth approximation of the ℓ_1-norm satisfying

$$\forall\, \boldsymbol{\alpha} = (\alpha_i)_{1\leqslant i\leqslant (l+1)p} \in \mathbb{R}^{(l+1)p}, \quad h_\mu(\boldsymbol{\alpha}) = \sum_{i=1}^{(l+1)p} h_i,$$

with

$$h_i = \begin{cases} \alpha_i^2/(2\mu), & \text{if } |\alpha_i| < \mu, \\ |\alpha_i| + \mu/2, & \text{otherwise,} \end{cases} \quad \forall i \in \{1, \dots, (l+1)p\},$$

and $\mu > 0$. In practice, the smoothing parameter μ can be chosen small so that the function h_μ behaves similarly to the ℓ_1-norm. Let us highlight that the minimization problem (4) is convex and can be solved efficiently using, e.g., the algorithm presented in [7].

We noticed experimentally that the addition of the function h_μ in the minimization procedure was improving the accuracy of the estimated signals and transformations parameters by producing a coarse-to-fine scales reconstruction of the images. This function acts as a proximal term and provides, up to some limits, a control on the evolution of the sequence of estimated images $(\boldsymbol{x}^k)_{k\geqslant 0}$. Remembering that the ℓ_1-norm favors the selection of few large coefficients, this function imposes that the next estimate $\boldsymbol{x}^{k+1}$ differs from $\boldsymbol{x}^k$ by a few large wavelet coefficients. The bigger the $\lambda_{\boldsymbol{x}}^k$ parameter is, the fewer the number of wavelet atoms that can be added at each iteration is. In practice, we start form $\boldsymbol{x}^0 = \boldsymbol{0} \in \mathbb{R}^{(l+1)n}$ and with a large value of $\lambda_{\boldsymbol{x}}^k$ at $k = 0$. We then slightly decrease the value of $\lambda_{\boldsymbol{x}}^k$ at each iteration. This allows us to have a coarse-to-fine scales reconstruction of the images, as illustrated in [1].

C. Second step of the algorithm

In the second step of the algorithm, we update the transformation parameters to further decrease the value of the objective function. As the function $\boldsymbol{\theta} \mapsto \|\mathsf{A}(\boldsymbol{\theta})\, \boldsymbol{x}^{k+1} - \boldsymbol{y}\|_2^2$ and i_Θ are separable in $\boldsymbol{\theta}_j$, $j = 1, \dots, l$, we optimize the transformation parameters separately for each observations.

To simplify the notations, we introduce l new functions $Q_j^{k+1}\colon \mathbb{R}^q \to \mathbb{R}$, with $j = 1, \dots, l$, satisfying

$$Q_j^{k+1}(\boldsymbol{\theta}_j) = \|\mathsf{A}_j\mathsf{S}(\boldsymbol{\theta}_j)\boldsymbol{x}_0^{k+1} + \mathsf{A}_j\boldsymbol{x}_j^{k+1} - \boldsymbol{y}_j\|_2^2. \quad (5)$$

One of our goal is to find parameters $\boldsymbol{\theta}_j^{k+1} \in \Theta_j$ such that $Q_j^{k+1}(\boldsymbol{\theta}_j^{k+1}) \leqslant Q_j^{k+1}(\boldsymbol{\theta}_j^k)$. These functions are non-linear in $\boldsymbol{\theta}_j$. To simplify the estimation of the parameters, we instead minimize quadratic approximations of these functions. Assuming that the entries of the matrix $\mathsf{S}(\boldsymbol{\theta}_j)$ are differentiable with respect to the transformation parameters, the first order Taylor expansion of $\mathsf{S}(\boldsymbol{\theta}_j)\boldsymbol{x}_0^{k+1}$ at $\boldsymbol{\theta}_j^k$ is $\mathsf{S}(\boldsymbol{\theta}_j^k)\boldsymbol{x}_0^{k+1} + \mathsf{J}_j^k(\boldsymbol{\theta}_j - \boldsymbol{\theta}_j^k)$ with

$$\mathsf{J}_j^k = \left(\partial_{\theta_{1j}}\mathsf{S}(\boldsymbol{\theta}_j^k)\boldsymbol{x}_0^{k+1}, \dots\ , \partial_{\theta_{qj}}\mathsf{S}(\boldsymbol{\theta}_j^k)\boldsymbol{x}_0^{k+1}\right) \in \mathbb{R}^{n\times q}.$$

[3]It satisfies $\mathsf{W}\mathsf{W}^\intercal = \mathsf{I}_n$, with $\mathsf{I}_n \in \mathbb{R}^{n\times n}$ the identity matrix.

Therefore, $Q_j^{k+1}(\boldsymbol{\theta}_j^k) + P_j^{k+1}(\boldsymbol{\theta}_j)$, with

$$P_j^{k+1}(\boldsymbol{\theta}_j) = \left\langle \nabla Q_j^{k+1}(\boldsymbol{\theta}_j^k), \boldsymbol{\theta}_j - \boldsymbol{\theta}_j^k \right\rangle + \ \|\mathsf{A}_j \mathsf{J}_j^k (\boldsymbol{\theta}_j - \boldsymbol{\theta}_j^k)\|_2^2,$$

and

$$\nabla Q_j^{k+1}(\boldsymbol{\theta}_j^k) = 2\left(\mathsf{A}_j \mathsf{J}_j^k\right)^\intercal \left(\mathsf{A}_j \mathsf{S}(\boldsymbol{\theta}_j^k)\boldsymbol{x}_0^{k+1} + \mathsf{A}_j \boldsymbol{x}_j^{k+1} - \boldsymbol{y}_j\right),$$

is a quadratic approximation of Q_j^{k+1} at $\boldsymbol{\theta}_j^k$.

To update the transformation parameters, we minimize this quadratic approximation to which we add another quadratic term that ensures a decrease of the objective function L. The next estimate of the transformation parameters is

$$\boldsymbol{\theta}_j^{k+1} = \underset{\boldsymbol{\theta}_j \in \Theta_j}{\operatorname{argmin}} \ P_j^{k+1}(\boldsymbol{\theta}_j) + \frac{2^i \lambda_{\boldsymbol{\theta}}}{2} \|\boldsymbol{\theta}_j - \boldsymbol{\theta}_j^k\|_2^2, \qquad (6)$$

where $\lambda_{\boldsymbol{\theta}} > 0$ and i is the smallest positive integer such that

$$Q_j^{k+1}(\boldsymbol{\theta}_j^{k+1}) \leqslant \ Q_j^{k+1}(\boldsymbol{\theta}_j^k) + P_j^{k+1}(\boldsymbol{\theta}_j^{k+1}) + \ \frac{(2^i - 1)\lambda_{\boldsymbol{\theta}}}{2} \|\boldsymbol{\theta}_j^{k+1} - \boldsymbol{\theta}_j^k\|_2^2. \qquad (7)$$

The above condition ensures that $\boldsymbol{\theta}_j^{k+1}$ decrease the value of objective function and is essential for the convergence of the sequence $(\boldsymbol{x}^k, \boldsymbol{\theta}^k)_{k\in\mathbb{N}}$ to a critical point of L.

D. Convergence result

We are now in position to state our convergence result, whose proof can be found in [1].

Theorem 1: Let L be the objective function defined in (3) with $\kappa > 0$. Assume that L is bounded below, that the entries of S_j, with $j = 1, \ldots, l$, are twice continuously differentiable, that $\Psi \in \mathbb{R}^{(l+1)n \times (l+1)p}$ satisfies $\Psi\Psi^\intercal = \mathsf{I}_{(l+1)n}$, and that the stepsizes satisfy $0 < \underline{\lambda} \leqslant \lambda_{\boldsymbol{x}}^k, \lambda_{\boldsymbol{\theta}} \leqslant \bar{\lambda}$ for all $k \in \mathbb{N}$. Then, the sequence of estimates $(\boldsymbol{x}^k, \boldsymbol{\theta}^k)_{k\in\mathbb{N}}$ generated by the algorithm described above is correctly defined and the following statements hold:

1) For all $k \geqslant 0$,

$$L(\boldsymbol{x}^k, \boldsymbol{\theta}^k) - L(\boldsymbol{x}^{k+1}, \boldsymbol{\theta}^{k+1}) \geqslant \frac{\underline{\lambda}}{2}\left[\kappa \ \|\boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}^k\|_2^2 + h_\mu(\Psi^\intercal(\boldsymbol{x}^{k+1} - \boldsymbol{x}^k))\right]. \qquad (8)$$

   Hence $L(\boldsymbol{x}^k, \boldsymbol{\theta}^k)$, $k \in \mathbb{N}$, does not increase.
2) The sequences $(\boldsymbol{x}^{k+1} - \boldsymbol{x}^k)_{k\in\mathbb{N}}$ and $(\boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}^k)_{k\in\mathbb{N}}$ converge. Indeed,

$$\lim_{k\to+\infty} \|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|_2 + \|\boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}^k\|_2 = 0. \qquad (9)$$

3) Assume that L has the Kurdyca-Łojasiewicz property (see Definition 3.2 in [5]). Then, if the sequence $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ is bounded, the sequence $(\boldsymbol{x}^k, \boldsymbol{\theta}^k)_{k\in\mathbb{N}}$ converges to a critical point $(\boldsymbol{x}^*, \boldsymbol{\theta}^*)$ of L.

The last point of Theorem 1 applies if L has the Kurdyca-Łojasiewicz property. As explained in [5], this property is satisfied by several classes of functions. We detail in [1] several examples where the conditions required by Theorem 1 are satisfied. For example, if the interpolation matrices S_j, $j = 1, \ldots, l$, are built using the bicubic interpolation [8] and $f(\boldsymbol{x}) = \|\Phi\boldsymbol{x}\|_1$ for some basis $\Phi \in \mathbb{R}^{(l+1)n\times(l+1)n}$, then the sequence of estimates converges to a critical point of the objective function L for geometric transformations such as translations, affine transformations or "small" homographies.

III. Free breathing coronary MRI

A. Acquisition model

We acquired 2D image data of the right coronary artery in a healthy adult subject with a clinical 3T scanner (Siemens Trio, Erlangen, Germany). The field of view was 320×320 mm and the spatial resolution was $1 \times 1 \times 8$ mm. Our goal is here to reconstruct a high resolution image $\boldsymbol{x}_0$ of the heart containing 320×320 pixels from this set of few Fourier measurements.

Note that in MRI we are dealing with complex images. For simplicity, and to be able to use the proposed algorithm without modifications, we treat separately the real and imaginary parts of the images. Therefore, the vector $\boldsymbol{x}_0$ has size $n = 2 \times 320^2$ and satisfies: $\boldsymbol{x}_0^\intercal = ((\boldsymbol{x}_0^r)^\intercal, (\boldsymbol{x}_0^i)^\intercal)$, where $\boldsymbol{x}_0^r \in \mathbb{R}^{n/2}$ and $\boldsymbol{x}_0^i \in \mathbb{R}^{n/2}$ are the real and imaginary parts of the image. The same convention is used for the foreground images. Note that the j^{th} transformed background image is now obtained by multiplication with the block-diagonal matrix built by repeating twice $\mathsf{S}(\boldsymbol{\theta}_j)$ on the diagonal. The measurements describing these images are then obtained as follows.

At each cardiac cycle, we acquire few complex Fourier coefficients lying along 15 radial lines, as presented in Fig. 1, with each line containing 320 equispaced sampling points. As before, we separate the real and imaginary parts of the measurements and stack them in a single measurement vector: $\boldsymbol{y}_j^\intercal = ((\boldsymbol{y}_j^r)^\intercal, (\boldsymbol{y}_j^i)^\intercal)$, with $\boldsymbol{y}_j^r \in \mathbb{R}^{m/2}$ and $\boldsymbol{y}_j^i \in \mathbb{R}^{m/2}$. The number of acquired measurements at each cycle satisfies $m/n = 4.7\ \%$ and a total of $l = 24$ acquisitions are performed at different cycles. Note that the radial lines along which the measurements are acquired change at each cardiac cycle to cover the Fourier space as much as possible. Let Ω_j be the set of frequencies probed at the j^{th} cycle. We model this acquisition using the complex Fourier matrix $\mathsf{F}_{\Omega_j} \in \mathbb{C}^{m/2\times n/2}$ which estimates the Fourier transform of a discrete complex image on the frequencies Ω_j. The observation matrix A_j then satisfies

$$\mathsf{A}_j = \begin{bmatrix} \mathsf{F}_{\Omega_j}^r & -\mathsf{F}_{\Omega_j}^i \\ \mathsf{F}_{\Omega_j}^i & \mathsf{F}_{\Omega_j}^r \end{bmatrix},$$

where $\mathsf{F}_{\Omega_j}^r \in \mathbb{R}^{m/2\times n/2}$ and $\mathsf{F}_{\Omega_j}^i \in \mathbb{R}^{m/2\times n/2}$ are the real and imaginary part of F_{Ω_j}.

While we are mainly interested in the reconstruction of $\boldsymbol{x}_0$, the l other foreground images have still their place in the measurement model (1). Indeed, as we are imaging one slice of an object moving in a 3D space, trough-plane motion might occur. The l foreground images can compensate for such negative effects. However, we have access to only $4.7\,\%\ n$ measurements to estimate each foreground image. We thus do not expect to obtain an accurate reconstruction of these images. On the contrary, all the $4.7\,\%\ ln$ measurements contribute to

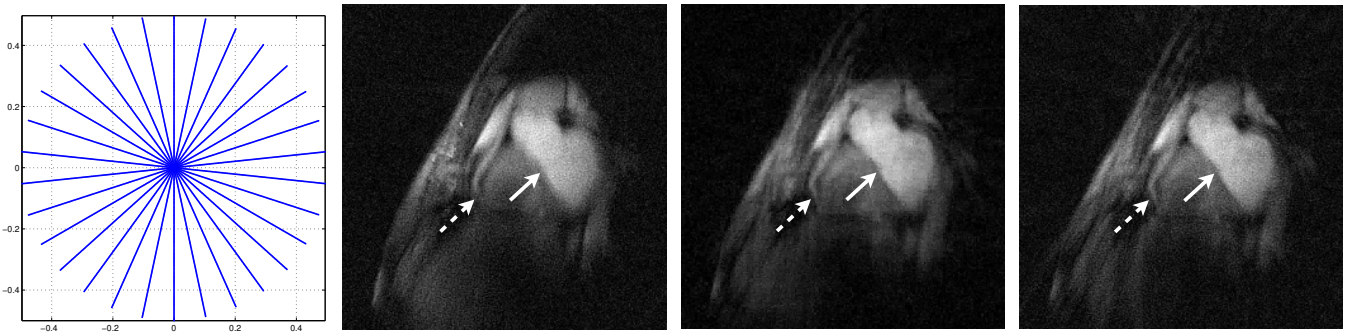


Fig. 1. From left to right: sampling pattern in Fourier space for one cardiac cycle; reconstructed image using the usual method without registration; reconstructed image $\boldsymbol{x}_0^*$ using our method; reconstructed image using the usual method with registration by the estimated parameters $\boldsymbol{\theta}^*$ obtained by our method.

the estimation of $\boldsymbol{x}_0$. Obtaining an accurate reconstruction of this reference image is thus possible.

Let us remark that we considered only one *channel* of the receiving coil in the above measurement model. In total, 32 channels are present and each of them gives access to a local image of the heart. The images of all these channels are usually combined to increase the signal-to-noise ratio of the recovered image and to have a more uniform spatial coverage. Ideally, we should also combine the measurements provided by each channel. However, the problem to solve becomes more challenging and addressing it is beyond the scope of this short abstract. We thus restrict our study to one channel only (chosen to have the best coverage of the heart).

B. Reconstruction results

We run our algorithm with $\kappa = 10^{-1}$, $\mu = 10^{-10}$, and $f(\cdot) = \|\Psi^\intercal \cdot\|_1$, where $\Psi \in \mathbb{R}^{(l+1)n \times (l+1)n}$ is built by repeating $2\,(l+1)$ times the Haar wavelet basis $\mathsf{W} \in \mathbb{R}^{n/2 \times n/2}$ on the diagonal. The transformations between cardiac cycles are assumed to be well modeled by translations. The translation parameters along both dimensions are initialised to 0 and are constrained to be in the set $[-50\,\mathrm{mm}, +50\,\mathrm{mm}]$. Finally, the stepsizes satisfy $\lambda_{\boldsymbol{x}}^k = \max((0.9)^k\, 500, 0.1)$ and $\lambda_{\boldsymbol{\theta}} = 0.1$.

Fig. 1 presents the reconstruction obtained from the acquired measurements with the proposed algorithm, as well as the ones obtained with the usual reconstruction technique *without* and *with* registration with the transformation parameters estimated by our algorithm. The usual reconstruction technique consists of a gridding operation and an inverse Fourier transform [9]. The measurements are also weighted before the gridding operation to compensate for the fact that the low frequencies are more densely sampled than the high frequencies.

Compared to the reconstruction obtained without registration, one can see that the image of the heart is sharper (see arrows) with our reconstruction method. A part the coronary previously hidden becomes visible (dotted arrow), and the borders of the blood pool and the cardiac muscle become better defined, indicating that the translations are accurately estimated. Compared to the reconstruction with registration obtained with the usual technique, our reconstruction contains less noise, though some details are slightly less visible.

IV. Conclusion

We highlighted the interest of a reconstruction technique initially developed for image reconstruction from multi-view measurements, for free breathing coronary MRI. The method reconstructs a high resolution image of the heart from few Fourier measurements and automatically compensates for the motion of the heart occurring during the acquisition. The reconstruction algorithm minimizes a non-convex functional and the generated sequence of estimates converges to a critical point of this functional.

The current technique was designed assuming that the motion can be modeled by global geometric transformations, such as translations or homographies. This is an obvious limitation of the technique which prevents us to use it with more complicated types of motion. However, the requirements of Theorem 1 hold for a large class of transformation models. This leaves us the possibility to choose more general transformations. For example, we could approximate elastic transformations using a parametric model similar to [10] and estimate the corresponding parameters using the proposed algorithm.

References

[1] Puy *et al.*, "Robust image reconstruction from multi-view measurements," *SIAM J. Imaging Sci.*, submitted, 2012. arXiv:1212.3268.

[2] Stehning *et al.*, "Free- breathing whole-heart coronary mra with 3d radial ssfp and self-navigated image reconstruction", *Magn. Reson. Med.*, vol. 54(2), pp. 476–480, 2005.

[3] Bonanno *et al.*, About the performance of multi-dimensional radial self-navigation incorporating compressed sensing for free-breathing coronary mri, *ISMRM conference*, Melbourne, 2012.

[4] Unser, "Sampling50 years after shannon," *Proc. IEEE*, vol. 88(4), pp. 569–587, 2000.

[5] Attouch *et al.*, "Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the kurdyka-lojasiewicz inequality," *Math. Oper. Res.*, vol. 35(2), pp. 438–457, 2010.

[6] Attouch *et al.*, "Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized gauss-seidel methods," *Math. Program.*, 2011.

[7] Beck *et al.*, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imaging Sci.*, vol. 2(1), pp. 183–202, 2009.

[8] Keys, "Cubic convolution interpolation for digital image processing," *IEEE Trans. Acoustics, Speech, Signal Process.*, vol. 29(6), 1981.

[9] Jackson *et al.*, "Selection of a convolution function for fourier inversion using gridding [computerised tomography application]," *IEEE Trans. Medical Imag.*, vol. 10(3), pp. 473–478, 1991.

[10] Kybic *et al.*, "Fast parametric elastic image registration," *IEEE Trans. Image Process.*, vol. 12(11), pp. 1427–1442, 2003.

Localization of point sources in wave fields from boundary measurements using new sensing principle

Zafer Doğan*†, Ivana Jovanovic*†,Thierry Blu‡ and Dimitri Van De Ville*†
*Medical Image Processing Lab (MIPLAB), Institute of Bioengineering, EPFL, CH-1015 Lausanne, Switzerland
†Department of Radiology and Medical Informatics, University of Geneva, CH-1211 Geneva, Switzerland
‡Department of Electronic Engineering, Chinese University of Hong Kong, Shatin N.T. Hong Kong

***Abstract*—We address the problem of localizing point sources in 3D from boundary measurements of a wave field. Recently, we proposed the *sensing principle* which allows extracting volumetric samples of the unknown source distribution from the boundary measurements. The extracted samples allow a non-iterative reconstruction algorithm that can recover the parameters of the source distribution projected on a 2-D plane in the continuous domain without any discretization.**

Here we extend the method for the 3-D localization of multiple point sources by combining multiple 2-D planar projections. In particular, we propose a three-step algorithm to retrieve the locations by means of multiplanar application of the sensing principle. First, we find the projections of the locations onto several 2-D planes. Second, we propose a greedy algorithm to pair the solutions in each plane. Third, we retrieve the 3D locations by least squares regression.

***Index Terms*—Sensing principle, finite-rate-of-innovation (FRI), wave equation, source imaging, inverse problem**

I. Introduction

Inverse source problems have wide applications in signal processing and biomedical imaging. Among these, reconstruction of sparse source distributions from boundary measurements have attracted great attention of many researchers recently. In particular, several mathematical models are studied extensively, such as Poisson's equations for identification of current dipolar sources in electroencephalography (EEG) [1], the steady-state diffusion equation for the determination of a light source function in bioluminescence tomography (BLT) [2] and the wave equation for the recovery of heat absorption profile in photoacoustic tomograpghy (PAT) [3]–[5].

Many advanced techniques for the recovery of source distributions aim at super-resolution by exploiting sparsity properties of the underlying source distribution. For example, the low-dimensional signal subspace plays a key role for the MUSIC-type of algorithms to estimate the location of the absorbing regions [6]. Moreover, compressive sensing approaches have been studied recently for radar imaging applications [7].

Here, we focus on the inverse source problem for the wave equation from the boundary measurements of the field. The problem is ill-posed and thus challenging, and we exploit an explicit sparsity prior on the source model (i.e., a collection of point sources) that makes the problem well-posed [8]. Recently, we proposed the sensing principle that allows extracting volumetric samples of the source distribution with a set of well chosen sensing functions [9], [10]. These samples are then used in a non-iterative FRI-like framework [11] to retrieve the projected positions of the source distribution onto a 2-D plane. The key component of the method is the selection of the sensing functions which are used to extract the samples of the source function through surface integration. We have shown before that the localization of the selected families of sensing functions plays a key role in the accuracy of the estimation [9], [10]. Here we propose a multiplanar application of the sensing principle using a well-localized sensing functions for different projections planes. In particular, we propose a three-step algorithm to retrieve the locations of the pointwise source distribution. First, we extract the projected positions onto several 2-D planes. Second, we propose a greedy approach to pair the solutions between projection planes. Third, we reconstruct the 3-D locations from the 2-D paired solutions by a least squares regression.

The paper is organized as follows. In Section 2, we introduce the setting of the problem. In Section 3, we provide the key components of the sensing principle. In Section 4, we develop the proposed method for a 3-D measurement setup. The feasibility of the proposed method is demonstrated with numerical experiments in Section 5.

II. Problem Formulation

Consider an acoustic source distribution inside a volume Ω. In an acoustically homogeneous medium, the inhomogeneous wave equation is described by

$$\nabla^2 p(\mathbf{r},t) - \frac{1}{c^2}\frac{\partial^2 p(\mathbf{r},t)}{\partial t^2} = -H(\mathbf{r},t), \tag{1}$$

where $H(\mathbf{r},t)$ is a general representation of a spatiotemporal source distribution which we further decompose as the product $H(\mathbf{r},t) = A(\mathbf{r})I(t)$, where $A(\mathbf{r})$ is the spatial part and $I(t)$ is the temporal part of the source. In particular, we assume that the temporal behaviour is usually foreknown and we focus on the spatial part of the source function that we characterise as a pointwise source distribution;

$$A(\mathbf{r}) = \sum_{m=1}^{M} c_m \delta(\mathbf{r} - \mathbf{r_m}), \tag{2}$$

where $c_m \in \mathbb{R}$ is the intensity, and $\mathbf{r_m} \in \Omega$ is the location of M point source. With this parametrization the source distribution is completely described by the positions and intensities of M sources with $4M$ parameters. Hence, the goal of the

inverse problem is to reconstruct the point sources from the measurements of the wave field $p(\mathbf{r}, t)$ on the surface of the volume, $\partial\Omega$.

III. Sensing Principle

Let us consider the time harmonic solutions of (1)

$$\nabla^2 P(\mathbf{r}, \omega) + \frac{\omega^2}{c^2} P(\mathbf{r}, \omega) = -I(\omega) A(\mathbf{r}), \tag{3}$$

which is the inhomogeneous Helmholtz equation. Without loss of generality, we now consider a specific frequency ω. Based on the second Green's identity, we propose the sensing principle that provides a link between the source function and the measurements such that

$$\langle \Psi, A \rangle = \tag{4}$$
$$\frac{1}{I(\omega)} \oiint_{\partial\Omega} [P(\mathbf{r}, \omega) \nabla \Psi(\mathbf{r}, \omega) - \Psi(\mathbf{r}, \omega) \nabla P(\mathbf{r}, \omega)] \cdot \mathbf{e}_{\partial\Omega} dS,$$

where $I(\omega)$ is a constant that we use to compensate the surface integral and $\Psi(\mathbf{r}, \omega)$ is a sensing function satisfying the homogeneous Helmholz equation in the the volume

$$\nabla^2 \Psi(\mathbf{r}, \omega) + \frac{\omega^2}{c^2} \Psi(\mathbf{r}, \omega) = 0 \text{ in } \Omega. \tag{5}$$

This way the sensing principle allows to extract volumetric samples of the source distribution through a surface integral of the sensor measurements of the acoustic field. Finally, we use the extracted samples by the sensing principle, i.e., $\langle \Psi, A \rangle$ the so-called *generalised samples* to retrieve the parameters of the source function.

IV. Algorithm

We propose a three-step algorithm to estimate the 3-D location of the point sources from the observed acoustic field by means of applying the sensing principle.

A. Planar Projection

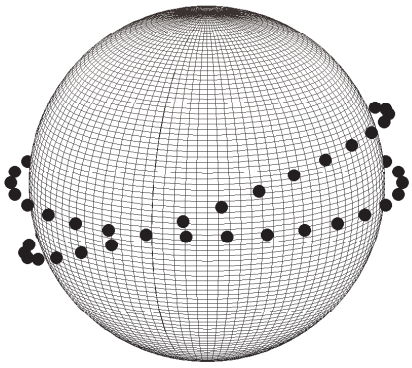


Fig. 1: Poles of the sensing functions in the horizontal XY-plane and after rotation in the X′Y′-plane.

In the first step, we choose a set of sensing functions Ψ satisfying (5) in a general X′Y′Z′coordinate system:

$$\Psi_n(\mathbf{R}\mathbf{r}, \omega) = \frac{e^{j\omega z'/c}}{x' + jy' - a_n}, \; a_n \notin \Omega, \tag{6}$$

where a_n's are the poles of the sensing function on X′Y′-plane located at equidistant angles $a_n = ae^{jn\theta}, n \in [\![0, N-1]\!]$, $|a|$ is greater than the radius of Ω excluding the volume and θ is an arbitrary angle. The matrix $\mathbf{R}$ represents rotation matrix of the coordinate system along the X-axis in a standard right-handed cartesian coordinate system given by

$$\underbrace{\begin{bmatrix} x' \\ y' \\ z' \end{bmatrix}}_{\mathbf{r}'} = \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 1 & \cos\alpha & \sin\alpha \\ 0 & -\sin\alpha & \cos\alpha \end{bmatrix}}_{\mathbf{R}} \underbrace{\begin{bmatrix} x \\ y \\ z \end{bmatrix}}_{\mathbf{r}}. \tag{7}$$

In Fig. 1, we provide a visualisation for the rotation of the poles of the sensing functions on the X′Y′-plane.

Then, we define a polynomial, $Q(X)$ whose roots are the positions of the point sources on the X′Y′-plane:

$$Q(X) = \prod_{m=1}^{M} (X - s'_m) = \sum_{k=0}^{M} q_k X^k \text{ where } s'_m = x'_m + iy'_m \tag{8}$$

where $q_M = 1$. With this selection, the extracted samples of the source function (4) turns into an annihilable equation as follows:

$$\begin{aligned} \langle \Psi_n, A \rangle = \mu_n &= \sum_{m=1}^{M} \frac{c_m e^{i\omega z'_m/c}}{x'_m + iy'_m - a_n} \\ &= \frac{\sum_{m=0}^{M-1} c'_m e^{imn\theta}}{\prod_{m=1}^{M} (x'_m + iy'_m - a_n)} \\ &= \frac{\sum_{m=0}^{M-1} c'_m e^{imn\theta}}{Q(a_n)}, \end{aligned} \tag{9}$$

where c'_m are complex-valued coefficients that do not depend on n nor θ. The sequence $\mu_n \cdot Q(a_n)$, for $n \in [\![0, N-1]\!]$ for some $N \geq 2M + 1$ (i.e., innovation rate given by the FRI sampling [11]), can be annihilated by a known FIR digital filter $h = \{h_k\}$ for $k \in [\![0, N-1]\!]$ characterized as

$$H(z) = \sum_{k \in \mathcal{Z}} h_k z^{-k} = \prod_{k=0}^{M-1} (1 - e^{ik\theta} z^{-1}),$$

where the zeros of the filter are chosen as the poles of (6) on the plane, i.e., $e^{ik\theta}$ for $k \in [\![0, M-1]\!]$. Finally, solving this annihilation system for the coefficients of the polynomial $Q(X)$, the point sources' positions on the X′Y′-plane are found to be the roots of the polynomial Q.

B. Pairing of the Projections

In the second step, we first define an inclusion map so that we can treat the projected points as in $\mathbb{R}^3$. Let $f : \mathbb{R}^2 \to \mathbb{R}^3$ be an inclusion map defined as

$$f\left(\mathbf{r}'_\mathbf{i} = \begin{bmatrix} x'_i \\ y'_i \end{bmatrix}\right) := \begin{bmatrix} x'_i \\ y'_i \cos(\alpha_i) \\ y'_i \sin(\alpha_i) \end{bmatrix} \tag{10}$$

for each projection point $\mathbf{r_i}'$ on a plane defined by the normal $\mathbf{n_i} = \mathbf{R_i^T n_0}$ (See Fig.2) where $\mathbf{R_i}$ is the rotation matrix of the coordinate system for α_i (7) and $\mathbf{n_0} = [0, 0, 1]^T$ is the normal vector of the standard XY-plane. We propose a naive

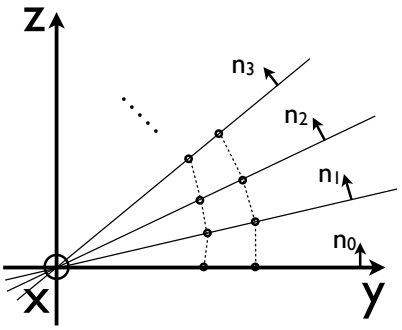


Fig. 2: Visualization of the closest pair of points algorithm for $M = 2$ points projected onto different planes P_i with the normals of the planes $\mathbf{n_i}$

solution for the closest pair problem for two separated sets of points between consecutive projection planes. Indeed, the main idea is to compute the Euclidean distance between all the pairs of points in two sets and then pick the pair with the smallest distance. Let us consider we have P projection planes defined by P_i $i \in [\![0, P-1]\!]$ where each plane has M points to be paired. We assume an initial labelling for the points in plane P_0 with 1 to M. Then, to find the closest pair of points $p \in P_i$ and $q \in P_{i-1}$, we compute the distances between all the $M \times M$ pairs of points and we pick and label the pair with the smallest distance and exclude it from the set. We repeat the same approach for the remaining points. We provide a summary of the method in Algorithm 1 and a visualisation in Fig 2. The method is computed in $O(n^2)$ but can be solved it in $O(n \log n)$ using the recursive divide and conquer approach [13].

Algorithm 1: Closest Pair of Points

Data: $p \in P_i$, for $i \in [[0, P-1]]$
Result: l_p: Labels of $p \in P_i$
begin
  Initialize: Label l_0: 1 to M
  for *i=1* **to** *P-1* **do**
  $P_i^* = P_i$
  while P_i^* *is not empty* **do**
    $p^* = \underset{p \in P_i^*}{\operatorname{argmin}} \underset{q \in P_{i-1}}{\min} ||f(p) - f(q)||^2$
    P_i^*=$P_i^* \backslash \{p^*\}$
  Label l_i : Match the labels of p^* and q

C. 3-D Reconstruction of the Positions

In the third step, we solve for the following least squares problem

$$\hat{\mathbf{r}}_m = \underset{\mathbf{r_m}}{\operatorname{argmin}} \sum_{i=0}^{P-1} ||D_i||^2, \ \forall \mathrm{m} \in [\![1, M]\!] \tag{11}$$

where $||D_i||$ is the distance of the solution r_m to the line that passes through the point $f(r_i')$ and parallel to n_i (See Fig. 3):

$$||D_i|| = \frac{||(\mathbf{r_m} - f(\mathbf{r_i'})) \times \mathbf{n_i}||}{||\mathbf{n_i}||},$$

where $\times$ represents the cross product of the two vectors and $||\mathbf{n_i}|| = 1$. In Fig. 3, we provide a visualisation of the solution.

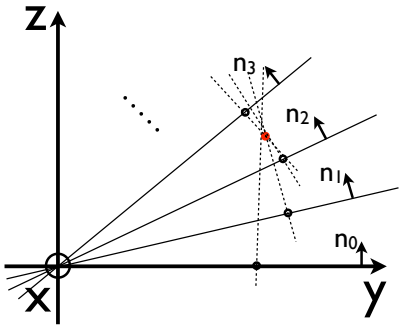


Fig. 3: Visualization of the 3-D reconstruction by least square regression of the distance between the true point (red) and the lines (dashed) defined by the projection points and the normals

V. Experimental Results

We performed numerical experiments to validate our reconstruction algorithm. Specifically, we considered a spherical detection geometry having a radius of 8 cm that is typical for the imaging of breast tissue in a PAT setting using a temporal illumination profile given as $I(t) = \partial/\partial t(e^{-t^2/2\sigma^2})/\sqrt{2\pi\sigma^2}$. The speed of sound is taken as constant $c = 1500$m/s and we assumed that there are 134^2 sensors uniformly positioned on the surface. We focused on the localisation accuracy of our method.

We define the reconstruction error per point source by

$$\mathrm{RMSE} = \sqrt{\frac{1}{M} \sum_{i=1}^{M} ||\mathbf{r}_m - \hat{\mathbf{r}}_m||^2}$$

where $\mathbf{r}_m$ is the true position, $\hat{\mathbf{r}}_m$ is the estimated position. In Fig. 4, we compare the reconstruction accuracy using the frequency samples at 200 KHz of the sensor data at 20 dB for varying number of projections such that the angle between the planes is $\frac{\pi}{2P}$. In Fig. 4, we demonstrate the improvement obtained by increased number of projections in which we achieve about mm reconstruction accuracy among a radius

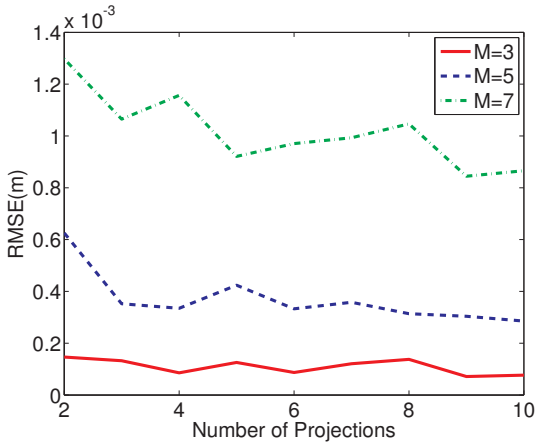


Fig. 4: RMSE results of an average of 50 independent realisations for varying number of projections for $M = 3, 5$ and 7 using the frequency samples at 200 KHz of the sensor data at 20 dB

of 8cm. We conclude that multiplanar approach performs accurate localization once the sensing principle is applied on sufficient number of projection planes, i.e., small projection angle between the planes.

VI. Conclusions

In sum, we proposed a non-iterative algorithm for the detection of point absorbers in three dimensional wave equation from the boundary measurements. The key component of the method is the selection of the sensing function that is used to extract the generalized samples by the surface integral. Here, we demonstrate that a well localised family of sensing functions with the proposed framework to build the solution in 3D from 2D projections can achieve accurate results even for the low SNR regime.

For simplicity of the discussion, we provide the method that combines the projected solutions using a simple rotation of the coordinate system along the X-axis only. However, a general rotation in three dimension can be obtained from three basic rotation matrices along X,Y, and Z-axes. Therefore, the idea can be easily generalized to a framework that combines the projections from any rotation as a composition of the rotations along the three axes.

Sparse models for the inverse source problems from overdetermined boundary field measurements remain as a promising research area of further research. The current work focuses on the systems governed by the wave equation, however the framework can be applied to similar problems encountered in different domains. Moreover, we also consider possibility and feasibility of the proposed method in real applications.

VII. Acknowledgments

This work was supported in part by the Swiss National Science Foundation (under the grants 205330-132808 and PP00P2-123438) and in part by the Center for Biomedical Imaging (CIBM) of the Geneva-Lausanne Universities and the EPFL and in part by the Hong Kong University Grant Council (under the RGC grant CUHK410110).

References

[1] D. Kandaswamy, T. Blu, and D. Van De Ville, "Analytic sensing: Noniterative retrieval of point sources from boundary measurements," *SIAM J. Sci. Comput.*, vol. 31, no. 4, pp. 3179–3194, Jul. 2009. [Online]. Available: http://dx.doi.org/10.1137/080712829

[2] J. Yu, F. Liu, J. Wu, L. Jiao, and X. He, "Fast source reconstruction for bioluminescence tomography based on sparse regularization," *IEEE Trans. Biomed. Engineering*, vol. 57, no. 10, pp. 2583–2586, 2010.

[3] L. V. Wang and S. Hu, "Photoacoustic tomography: In vivo imaging from organelles to organs," *Science*, vol. 335, no. 6075, pp. 1458–1462, 2012.

[4] M. Xu and L. Wang, "Photoacoustic imaging in biomedicine," *Rev. Sci. Instrum.*, vol. 77, no. 4, p. 041101, 2006.

[5] K. Wang and M. Anastasio, "Photoacoustic and thermoacoustic tomography: Image formation principles," in *Handbook of Mathematical Methods in Imaging*, O. Scherzer, Ed. Springer New York, 2011, pp. 781–815. [Online]. Available: http://dx.doi.org/10.1007/978-0-387-92920-0˙18

[6] H. Ammari, E. Bossy, V. Jugnon, and H. Kang, "Mathematical modeling in photoacoustic imaging of small absorbers," *SIAM Rev.*, vol. 52, pp. 677–695, November 2010. [Online]. Available: http://dx.doi.org/10.1137/090748494

[7] A. Fannjiang, P. Yan, and T. Strohmer, "Compressed remote sensing of sparse objects," *SIAM J. Imag. Sci.*, vol. 3, pp. 596–618, 2010.

[8] V. Isakov, "Uniqueness and stability in multi-dimensional inverse problems," *Inverse Problems*, vol. 9, no. 6, p. 579, 1993. [Online]. Available: http://stacks.iop.org/0266-5611/9/i=6/a=001

[9] Z. Dogan, I. Jovanovic, T. Blu, and D. Van De Ville, "Application of a new sensing principle for photoacoustic imaging of point absorbers," *Photons Plus Ultrasound: Imaging and Sensing 2013*, vol. 8581-144, 2013.

[10] Z. Dogan, I. Jovanovic, T. Blu, and D. Van de Ville, "3d reconstruction of wave-propagated point sources from boundary measurements using joint sparsity and finite rate of innovation," in *Biomedical Imaging (ISBI), 2012 9th IEEE International Symposium on*, may 2012, pp. 1575 –1578.

[11] M. Vetterli, P. Marziliano, and T. Blu, "Sampling signals with finite rate of innovation," *Signal Processing, IEEE Transactions on*, vol. 50, no. 6, pp. 1417 –1428, jun 2002.

[12] T. Blu, P.-L. Dragotti, M. Vetterli, P. Marziliano, and L. Coulot, "Sparse sampling of signal innovations," *Signal Processing Magazine, IEEE*, vol. 25, no. 2, pp. 31 –40, march 2008.

[13] M. I. Shamos and D. Hoey, "Closest-point problems," in *Foundations of Computer Science, 1975., 16th Annual Symposium on*, oct. 1975, pp. 151 –162.

Compressive Acquisition of Sparse Deflectometric Maps

Prasad Sudhakar, Laurent Jacques and Adriana Gonzalez
ELEN Department, ICTEAM
Université catholique de Louvain, Belgium.
Email: firstname.secondname@uclouvain.be

Xavier Dubois, Philippe Antoine and Luc Joannes
Lambda-X
Nivelles, Belgium.
Email: firstname.secondname@lambda-x.com

***Abstract*—Schlieren deflectometry aims at measuring deflections of light rays from transparent objects, which is subsequently used to characterize the objects. With each location on a smooth object surface a sparse deflection map (or spectrum) is associated. In this paper, we demonstrate the compressive acquisition and reconstruction of such maps, and the usage of deflection information for object characterization, using a schlieren deflectometer. To this end, we exploit the sparseness of deflection maps and we use the framework of spread spectrum compressed sensing. Further, at a second level, we demonstrate how to use the deflection information optimally to reconstruct the distribution of refractive index inside an object, by exploiting the sparsity of refractive index maps in gradient domain.**

I. Introduction

Schlieren deflectometry is a modality to measure the deflections undergone by light in a transparent object [1]. These deflections are used to characterize the properties of the transparent objects such as the surface curvature, distribution of the refractive index, etc. Unlike interferometry, deflectometry is insensitive to vibrations and hence is very attractive for industrial deployment (*e.g.*, for quality control).

Considering a thin transparent object with an incident parallel beam of light rays, as shown in Fig. 1(left). At each surface location $\boldsymbol{p}$, the function of our interest is a *deflection spectrum* $\tilde{s}_{\boldsymbol{p}}(\theta,\varphi) \in \mathbb{R}_+$, representing the flux of the light deviated in the direction (θ,φ), in a spherical coordinate system. These deflection spectra provide information about the curvature of the object, and hence it is interesting to study them.

For convenience, $\tilde{s}_{\boldsymbol{p}}$ is represented in this paper by its projection in the $\Pi_{\boldsymbol{p}} = \boldsymbol{e}_2\boldsymbol{e}_3$ plane, *i.e.*, according to the projected function $s_{\boldsymbol{p}}(r(\theta),\varphi) = \tilde{s}_{\boldsymbol{p}}(\theta,\varphi)$ with $r(\theta) = \tan\theta$. Moreover, the object surface is assumed sufficiently smooth for being parametrized by a projection of $\boldsymbol{p}$ in the same plane (on a arbitrary fixed origin), so that $\boldsymbol{p}$ is basically a 2-D vector.

An important feature of deflections is that for most objects (*e.g.*, with smooth surfaces), for any location $\boldsymbol{p}$, the flux is limited to range of angles and hence the deflection spectra therefore tend to be naturally *sparse* in plane $\Pi_{\boldsymbol{p}}$ or in some appropriate basis of this domain (*e.g.*, wavelets). Fig. 1(right) shows an example of a discretized deflection spectrum $s_{\boldsymbol{p}}$ for one location of a plano convex lens obtained using the setup described in Sec. II. The white spot in the image signifies that deflections occur at only a few angles (as governed by classical optics) and deflections elsewhere are negligible.

The optical setup described in Sec. II measures the deflection spectrum $s_{\boldsymbol{p}}$ for each location $\boldsymbol{p}$ indirectly by optical comparison with a certain number of programmable modulation patterns. Computationally, these optical comparisons are nothing but inner products between the deflection spectrum and the modulation patterns. By assuming an extreme case of the spectrum being a mere impulse, Phase Shifting Schlieren (PSS) method measures the deflection angles by using multiline phase shifted patterns in the SLM [2]. However, it is a limitation to ignore the richness of the deflection spectrum.

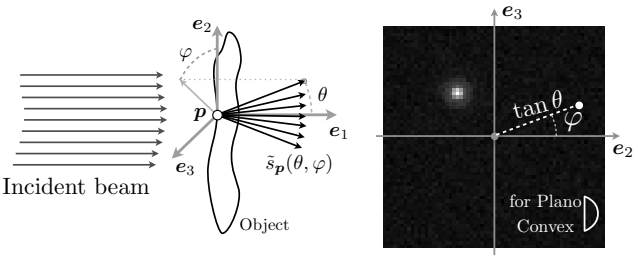


Fig. 1: Left, illustration of a deflection spectrum. Right, a typical (projected) deflection spectrum $s_{\boldsymbol{p}}$ for a plano convex lens of optical power 25.12D.

To this end, aided by the hindsight that each deflection spectra is sparse, we use the framework of *spread spectrum*[1] compressive sensing [3], described in Sec. III, to capture maximum information about the spectrum using relatively few modulation patterns, and then reconstruct the spectrum at each location by solving an inverse problem. In effect, each CCD pixel of our system behaves like a *single pixel camera* [4], but for deflection spectrum.

In Sec. IV, we present the numerical results of reconstructing deflection spectra from deflectometric measurements, after calibrating the system relative to its intrinsic noise. By making further assumptions about the spectra, we show in Sec. V how the deflection information can be obtained without explicit reconstruction of the spectra.

If the object contains regions of varying refractive index, then light undergoes deflections internally and at each surface location only the resultant deflection is measured. Therefore, the deflections provide indirect information about the distribution of the refractive index (henceforth called Refractive Index Map (RIM)). This necessitates measuring deflections for several orientations of the object in order to recover the RIM. Sec. VI briefly describes how the sparsity of RIM helps in its reconstruction using deflections from only few orientations.

PS is supported by the DETROIT project (WIST3), Walloon Region, Belgium. LJ is supported by the Belgian FRS-FNRS fund.

[1] "Spread Spectrum" is not related to the studied deflection "spectrum" but it refers to the signal frequency spectrum.

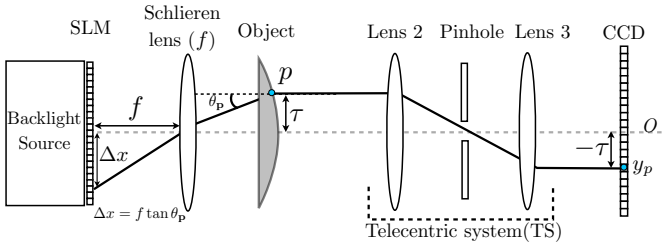


Fig. 2: A 2-D schematic of Schlieren deflectometer.

II. Schlieren Deflectometer

Deflection spectra can be measured by the Schlieren deflectometer, shown in Fig. 2, which consists of *(i)* a Spatial Light Modulator (SLM), *(ii)* the Schlieren lens with focal length f, *(iii)* the Telecentric System (TS) and *(iv)* the Charged Coupled Device (CCD) camera collecting the light.

The object to be analyzed is placed in between the Schlieren lens and the telecentric system. It is shined on its left by a light source and, due to the telecentric system, only the parallel light rays emerging out of the object are collected by the CCD. Moreover, up to a flip around the optical axis, each location $\boldsymbol{p}$ on the object at a distance τ from the optical axis O (dashed line), is probed by a corresponding CCD pixel also at a distance of τ from O. Each location $\boldsymbol{p}$ is thus in one-to-one correspondence with a CCD pixel and we will sometimes consider $\boldsymbol{p}$ as CCD pixel location.

From classical optics, a light ray that is incident on location $\boldsymbol{p}$ at an angle $\theta_{\boldsymbol{p}}$ originates from the light source at a distance of $\Delta x = f \tan\theta_{\boldsymbol{p}}$ from the optical axis. Likewise, the light rays originating at different locations on the source have different incident angles at $\boldsymbol{p}$. Since we can always virtually invert the light propagation in the system, everything works as if the object was shined on its right by a beam of parallel light rays. Therefore, up to a global scaling by f, the SLM plane is actually the local plane $\Pi_{\boldsymbol{p}}$ of the deflection spectrum occurring at $\boldsymbol{p}$. Modulating the SLM corresponds to modulating $s_{\boldsymbol{p}}$, while the light collected in CCD pixel $\boldsymbol{p}$ is just an inner product of $s_{\boldsymbol{p}}$ with the modulation.

If we generate M such modulations $\phi_i \in \mathbb{R}^N$ with $1 \leq i \leq M$ in the SLM of N pixels, considering the discrete nature of the CCD camera (having N_{C} pixels), the discretized deflection spectra are observed through

$$\boldsymbol{y}_k = \boldsymbol{\Phi}\boldsymbol{s}_k + \boldsymbol{n}, \quad 1 \leq k \leq N_{\mathrm{C}}, \tag{1}$$

where $\boldsymbol{\Phi}^T = (\boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_M) \in \mathbb{R}^{N\times M}$ is the sensing matrix, k is a CCD pixel index, $\boldsymbol{s}_k \in \mathbb{R}^N$ is the discretized spectrum at the k^{th} pixel/object location, and $\boldsymbol{n}$ models the measurement noise (assumed Gaussian). Notice that the SLM and the CCD 2-D grids are represented as 1-D spaces for brevity of notation, so that $\boldsymbol{\Phi}$ is then a sensing 2-D matrix acting on 1-D vectors.

To optimize the design of $\boldsymbol{\Phi}$ we rely upon spread spectrum compressed sensing theory.

III. Spread Spectrum Compressive Sensing

In Spread Spectrum Compressive Sensing (SSCS), a signal $\boldsymbol{x} = \boldsymbol{\Psi}\boldsymbol{\alpha} \in \mathbb{C}^N$, having a *sparse* representation in an orthonormal *sparsity basis* $\boldsymbol{\Psi} \in \mathbb{C}^{N\times N}$, *i.e.*, $\|\boldsymbol{\alpha}\|_0 := \#\{j : \alpha_j \neq 0\} \leq K \ll N$ is randomly pre-modulated before sensing [3]. Given a Rademacher or Steinhaus sequence $\boldsymbol{m} \in \mathbb{C}^N, |m_i| = 1$, the sensing process is summarized by

$$\boldsymbol{y} = \boldsymbol{\Gamma}_\Omega^* \boldsymbol{M}\boldsymbol{\Psi}\boldsymbol{\alpha} \ + \ \boldsymbol{n}, \tag{2}$$

where * denotes the conjugate transpose, $\boldsymbol{\Gamma} \in \mathbb{C}^{N\times N}$ is an orthonormal *sensing basis*, $\boldsymbol{\Gamma}_\Omega$ is the $M \times N$ submatrix formed by restricting the columns of $\boldsymbol{\Gamma}$ to those in $\Omega \subset [N] := \{1, \cdots, N\}$, $\boldsymbol{M} = \mathrm{diag}(\boldsymbol{m})$ and $\boldsymbol{n}$ is a Gaussian noise vector.

The signal is reconstructed by solving a convex optimization problem, known as Basis Pursuit De-Noising (BPDN) [5]

$$\widehat{\boldsymbol{\alpha}} := \underset{\widetilde{\boldsymbol{\alpha}}\in\mathbb{C}^N}{\arg\min} \ \|\widetilde{\boldsymbol{\alpha}}\|_1 \text{ subject to } \|\boldsymbol{y} - \boldsymbol{\Phi}\widetilde{\boldsymbol{\alpha}}\|_2 \leq \epsilon, \tag{3}$$

where $\boldsymbol{\Phi} = \boldsymbol{\Gamma}_\Omega^* \boldsymbol{M}\boldsymbol{\Psi}$, and ϵ is a bound on $\|\boldsymbol{n}\|_2 \leq \epsilon$.

For a given ϵ, the number of measurements M required by (3) to find a solution is, in general, governed by the sparsity level K and the *coherence*

$$\mu := \max_{1\leq i,j\leq N} |\langle \boldsymbol{\gamma}_i, \boldsymbol{M}\boldsymbol{\psi}_j \rangle|, \tag{4}$$

where $\boldsymbol{\gamma}_i$ and $\boldsymbol{\psi}_j$ are the columns of sensing and sparsity matrices respectively [3], [6]. Smaller the coherence, lesser is the number of measurements required for successful recovery of the solution, with a high probability.

Defining $C_{\boldsymbol{\Gamma},\boldsymbol{\Psi}} = \max_{1\leq i,j\leq N} \|\boldsymbol{\gamma}_i \circ \boldsymbol{\psi}_j\|_2$, where $\circ$ denotes pointwise product, the mutual coherence μ obeys

$$\mu \leq C_{\boldsymbol{\Gamma},\boldsymbol{\Psi}} \sqrt{2\log(2N^2/\delta)}, \tag{5}$$

with probability at least $1 - \delta$. When $\boldsymbol{\Gamma}$ is a *universal basis*, *i.e.*, when all the entries have the same complex amplitude c, spread spectrum is optimal with $C_{\boldsymbol{\Gamma},\boldsymbol{\Psi}} = c$ and the coherence $\mu \simeq c$, with a very high probability, irrespective of the sparsity basis. Specifically, for Fourier and Hadamard bases, $\mu \simeq 1/\sqrt{N}$ with high probability. We see in next section how to exploit the spread spectrum CS method in our optical setup.

IV. Deflection Spectrum Reconstruction

To apply the ideas of spread spectrum CS to schlieren deflectometry, certain practical aspects have to be considered. Most importantly, as the Spatial Light Modulator (SLM) accepts only real and non-negative valued entries, we use the Hadamard (universal) basis $\boldsymbol{H}$ combined with a random Rademacher vector $\boldsymbol{m}$ with $m_i = \pm 1$ independently with equal probability for sensing.

Further, the sensing basis is biased to have all the entries non-negative and an extra measurement is obtained to remove the bias during reconstruction. The details about obtaining the measurements can be found in [7].

Noise estimation: If there is no test object, then by classical optics the measured deflection spectrum is constant in all CCD pixels and corresponds to a simple disk centered on the origin of the spectrum domain. We denote it as $\boldsymbol{s}^{\mathrm{no}}$. The disk diameter is proportional to the pinhole diameter of the system (see Fig. 2). This prior information aids us in calibrating the system and in estimating the noise level on the measurements.

From actual measurements in the absence of test object, we obtain, on an arbitrary CCD pixel, $\boldsymbol{y}^{\mathrm{no}} = \boldsymbol{\Phi}(\boldsymbol{s}^{\mathrm{no}} + \boldsymbol{n}_{\boldsymbol{s}}) + \boldsymbol{n}_{\boldsymbol{y}}$, where $\boldsymbol{n}_{\boldsymbol{s}}$ and $\boldsymbol{n}_{\boldsymbol{y}}$ are the unknown signal and observation noises. After a small calibration of the SLM origin, and up to

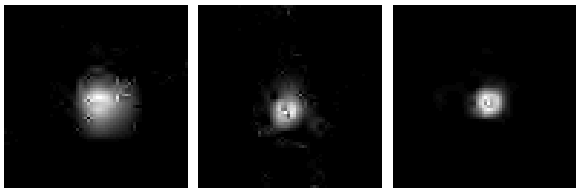


Fig. 3: An example of reconstruction using 2.5%, 10% and 100% of measurements.

a small optimization of the disk height in $\boldsymbol{s}^{\mathrm{no}}$, we can therefore compute a bound on the noise power as $\epsilon = \|\boldsymbol{\Phi}\boldsymbol{n_s} + \boldsymbol{n_y}\|_2 = \|\boldsymbol{y}^{\mathrm{no}} - \boldsymbol{\Phi}\boldsymbol{s}^{\mathrm{no}}\|_2$. We can either obtain this value for every M or estimate it for $M = N$ only and scale the result as $\epsilon(M) = \sqrt{M + 2\sqrt{M}}\ \epsilon(N)/\sqrt{N}$ for $M < N$. This estimate stems from the concentration properties of χ^2_M random variables.

Reconstruction procedure: For the reconstruction, we use the Daubechies 9/7 wavelet basis as our sparsity basis [8] which offers a sparser representation of the spectra than the canonical (Dirac) basis. To reconstruct the spectrum at any location k, an estimate of the sparse wavelet coefficients $\widehat{\boldsymbol{\alpha}}_k$ is obtained by solving (3) with the ϵ estimated above. The spectrum is then estimated by $\hat{\boldsymbol{s}}_k = \boldsymbol{\Psi}^* \widehat{\boldsymbol{\alpha}}_k$. To solve (3), we used the Chambolle-Pock algorithm, a first order primal-dual method for solving convex optimization problems using proximal operators [9]. Compared to a previous work on this subject [7], the reconstruction performance improved by constraining the estimate $\hat{\boldsymbol{s}}_k$ to be non-negative.

For evaluating compressive reconstruction performance, (3) was solved with $M = N$ measurements to obtain the reference reconstruction $\tilde{\boldsymbol{s}}_k$. Reconstructions for $M < N$ were compared with $\tilde{\boldsymbol{s}}_k$ using the (output) Signal-to-Noise Ratio oSNR $:= 20\log_{10}(\|\tilde{\boldsymbol{s}}_k\|_2/\|\tilde{\boldsymbol{s}}_k - \hat{\boldsymbol{s}}_k\|_2)$.

Experimental Results:[2] For experiments, we considered two plano convex lenses of optical powers 10.03D and 60D, and restricted the size of spectrum to 64×64 centered around the SLM origin, resulting in $N = 4096$. For 5 CCD locations, 10 independent reconstruction trials were performed for several values of M, by randomly drawing a new $\Omega \subset [N]$ every time.

Fig. 3 shows an example of deflection spectrum reconstructed using 2.5%, 10% and 100% of measurements, for the lens with $10D$ optical power. Note that the spectrum is well localized and sparse, corroborating our initial observation.

Fig. 4 shows the plot of oSNR versus the number of measurements M/N (in %), averaged over the trials and locations. The curves with square markers correspond to the solutions obtained using additional non-negativity constraints and the rest correspond to the lack of it. The oSNR improves as M/N increases, as expected. Though the absolute values of oSNR seem low, its significance has to be understood in the light of the input SNR, which is approximately computed as iSNR $:= 20\log_{10}(\|\boldsymbol{\Phi}\boldsymbol{s}^{\mathrm{no}}\|_2/\|\boldsymbol{y}^{\mathrm{no}} - \boldsymbol{\Phi}\boldsymbol{s}^{\mathrm{no}}\|_2) \simeq 4.34$ dB. The horizontal dotted line on the plot indicates the iSNR for our experiments, and it is clear that the reconstruction procedure improves the oSNR, beyond the iSNR, thereby demonstrating

[2]Computational resources have been provided by the supercomputing facilities of the Université catholique de Louvain (CISM/UCL) and the Consortium des Équipements de Calcul Intensif en Fédération Wallonie Bruxelles (CÉCI) funded by the F.R.S.-FNRS under convention 2.5020.11.

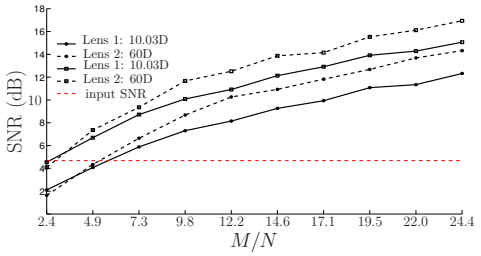


Fig. 4: Average reconstruction oSNR (in dB) as a function of M/N.

the ability of CS reconstruction of deflection spectra in low input SNR regime.

V. Obtaining Deflections without Reconstruction

Reconstruction of deflection spectra is a computationally intensive task and therefore if the objective is only to detect the location of the important feature of the spectrum (in our case, the location of the bright spot), then the idea of compressive domain signal processing can be used [10], [11]. Assuming that a template $\boldsymbol{g}^\rho$ for the feature can be built, the feature can be localized using a matched filtering operation performed directly on the measurements, without reconstructions. These locations provide a first guess of the deflections.

For *compressive spectrum detection*, given deflectometric measurements $\boldsymbol{y}_k$, we simply solve the following [7]

$$\widetilde{\tau}_k = \arg\max_{\tau} |\langle \boldsymbol{\Phi}^T \boldsymbol{y}_k, \boldsymbol{g}^\rho_\tau \rangle|, \tag{6}$$

where $\boldsymbol{g}^\rho_\tau$ is $\boldsymbol{g}^\rho$ translated by τ.

The experimental results showed that the distance between the centroids computed using compressive measurements and full reconstruction becomes sub-pixel for measurements size M/N as low as 4%, and continues to decrease as M increases. The evolution of the centroid estimation error versus the number of measurements M/N is available in [7].

We shall now see how to utilize deflection information for certain meaningful characterization of transparent objects.

VI. Refractive Index Map Reconstruction using Deflection Information

Characterizing a transparent object consisting of heterogeneous optical media by studying its Refractive Index Map (RIM), *i.e.*, the spatial distribution of the refractive index, is an important and challenging task for its manufacturing. In this section we will focus on the task of reconstructing RIM starting from deflection information.

The objective of the work is to demonstrate the relevance of sparsity and compressive sensing ideas for RIM reconstruction, independent of how the deflection maps are acquired (compressively or not). To emphasize that sparsity also helps in efficiently reconstructing RIM of transparent objects, we work with the deflection maps acquired (non-compressively) using the classical phase shifting schlieren method.

As shown in Fig. 5 (left), consider a refractive index map $\mathfrak{n}(\boldsymbol{r}), \boldsymbol{r} \in \mathbb{R}^2$ in the $\boldsymbol{e}_1\boldsymbol{e}_2$ plane (assuming that it is invariant along $\boldsymbol{e}_3$), that characterizes a complex object. For a given incident angle θ of the incoming light rays, schlieren deflectometer measures a two-dimensional map of the effective deflections $\Delta(\theta, \tau)$, where τ is the distance between the origin and the incident light ray under consideration.

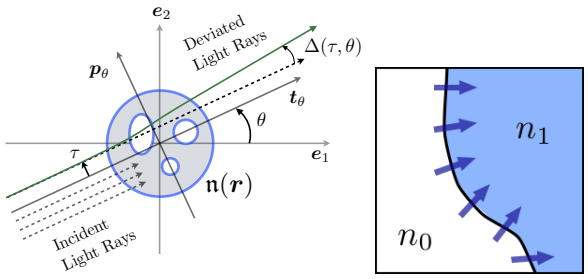


Fig. 5: (left) Model of light deflection through a transparent object. (right) TV model of RIM.

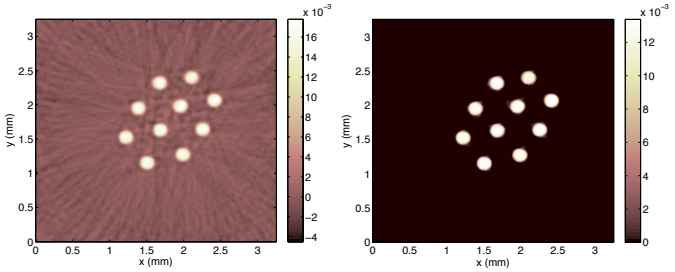


Fig. 6: An example of RIM reconstruction for a bundle of fibres with (left) the FBP and (right) TV minimization approach.

The measured deflection angles turn out to be the integral of the *transverse gradient* of the RIM, along the path of the light assumed to be straight (paraxial approximation) [12]. Notice that unlike the integration of the function itself in usual tomographic settings, here the integration is on the gradient of the function. Upon using the modified (deflectometric) Fourier slice theorem, the Fourier transform $\boldsymbol{y}_\theta$ of $\Delta(\theta,\tau)$, along τ for a fixed θ, provides one "slice" of the two-dimensional *polar* Fourier transform $\hat{\mathfrak{n}}(\boldsymbol{k})$ of $\mathfrak{n}(\boldsymbol{r})$ through the origin, *but each coefficient weighted by its distance to the frequency origin*. This weighting is in fact due to the integration of the gradient.

With a suitable discretization of the quantities and abuse of notations, the vectorized RIM $\mathfrak{n}$ and the vectorized Fourier transform of the deflection angles $\boldsymbol{y}_\theta$ are related by

$$\boldsymbol{y}_\theta = \boldsymbol{\Phi}\mathfrak{n} + \boldsymbol{n}, \tag{7}$$

where $\boldsymbol{\Phi}$ incorporates the Fourier operation and weighting factors arising from the slice theorem [13].

Reconstructing $\mathfrak{n}$ from the $\boldsymbol{y}_\theta$ involves measuring deflections from several incident angles θ and then solving an inverse problem using the forward model (7). To stabilize the inverse problem, suitable prior knowledge on $\mathfrak{n}$ has to be incorporated.

For a wide class of human made transparent objects, the RIM consists of slowly varying regions limited by sharp boundaries, as in Fig. 5(right), and therefore the RIM is sparse in the *gradient domain*. This prior knowledge about sparsity greatly helps us in reducing the number N_θ of incident angles that are needed to satisfactorily reconstruct the RIM.

Algorithmically, the RIM is reconstructed by promoting a solution with least Total Variation (TV) norm $\|\mathfrak{n}\|_{\text{TV}} = \|\nabla\mathfrak{n}\|_{2,1}$ [14], [9], that also respects the forward model (7) for a given noise level. The quality of the solution is further improved by using additional prior knowledge such as the non-negativity of $\mathfrak{n}$ and relevant boundary conditions.

For a test object of a bundle of optical fibres, Fig. 6(left) shows the reconstructed RIM, for the number of incident angles $N_\theta = 60$ (17% out of the possible 360 angles), using the well known Filtered Back Projection (FBP) algorithm that promotes a minimal ℓ_2 norm of the solution [15]. Fig. 6(right) shows the RIM reconstructed using a TV minimization approach, for the same number of angles. The TV reconstruction is better than that of FBP in not only suppressing the artifacts outside the fibre regions, but also in recovering the sharp edges between the fibres and the surroundings.

VII. Conclusions and Perspectives

This paper presents a novel approach for obtaining deflection information of transparent objects using schlieren deflectometer, and using this information to further characterize the objects. It has been demonstrated that suitable sparsity prior not only helps us to compressively acquire and reconstruct deflection maps, but also in efficiently using these deflections to reconstruct refractive index maps.

For further work, it is of foremost importance to understand the noise properties to tune the reconstruction method. Methods have to be developed to fully exploit the rich nature of deflection spectrum for object characterization. We also intend to develop approaches to exploit redundant dictionaries (e.g. undecimated wavelets), analysis-based reconstructions or correlation between neighbouring spectra for their simultaneous reconstruction.

References

[1] G. S. Settles, *Schlieren and Shadowgraph Techniques: Visualizing Phenomena in Transparent Media*. Springer, New York, NY, USA, 2001.

[2] L. Joannes, F. Dubois, and J. C. Legros, "Phase-shifting schlieren: high-resolution quantitative schlieren that uses the phase-shifting technique principle," *Applied optics*, vol. 42, no. 25, pp. 5046–5053, 2003.

[3] G. Puy, P. Vandergheynst, R. Gribonval, and Y. Wiaux, "Universal and efficient compressed sensing by spread spectrum and application to realistic fourier imaging techniques," *EURASIP Journal on Advances in Signal Processing*, vol. 2012, pp. 1–13, 2012.

[4] M. F. Duarte, M. A. Davenport, D. Takhar, J. N. Laska, T. Sun, K. F. Kelly, and R. G. Baraniuk, "Single-pixel imaging via compressive sampling," *IEEE Sig. Proc. Magazine*, vol. 25, no. 2, pp. 83–91, 2008.

[5] J. Tropp and S. Wright, "Computational methods for sparse solution of linear inverse problems," *Proceedings of the IEEE*, 2010.

[6] H. Rauhut, "Compressive sensing and structured random matrices," *Theoretical Found. and Num. Methods for Sparse Recovery*, 2010.

[7] P. Sudhakar, L. Jacques, X. Dubois, P. Antoine, and L. Joannes, "Compressive schlieren deflectometry," submitted to ICASSP 2013. [Online]. Available: http://arxiv.org/abs/1212.0433

[8] S. Mallat, *A Wavelet Tour of Signal Processing: The Sparse Way*, 3rd ed. Academic Press, 2008.

[9] A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *Journal of Mathematical Imaging and Vision*, vol. 40, no. 1, pp. 120–145, May 2011.

[10] M. A. Davenport, M. F. Duarte, M. B. Wakin, J. N. Laska, D. Takhar, K. F. Kelly, and R. G. Baraniuk, "The smashed filter for compressive classification and target recognition," in *Proceedings of Computational Imaging V at SPIE Electronic Imaging*, San Jose, CA, Jan. 2007.

[11] M. A. Davenport, P. T. Boufounos, M. B. Wakin, and R. G. Baraniuk, "Signal processing with compressive measurements," *IEEE Journal on Selected Topics in Signal Processing*, vol. 4, no. 2, pp. 445–460, 2010.

[12] D. Beghuin, J.-L. Dewandel, L. Joannes, E. Foumouo, and P. Antoine, "Optical deflection tomography with the phase-shifting schlieren," *Optics letters*, vol. 35, no. 22, pp. 3745–3747, 2010.

[13] A. Gonzalez, L. Jacques, C. D. Vleeschouwer, and P. Antoine, "Compressive optical deflectometric tomography: A constrained total-variation minimization approach," *CoRR*, vol. abs/1209.0654, 2012.

[14] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Phys. D*, vol. 60, pp. 259–268, Nov. 1992.

[15] *Principles of computerized tomographic imaging*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2001.

Fourier-Laguerre transform, Convolution and Wavelets on the Ball

Jason D. McEwen* and Boris Leistedt*
* Department of Physics and Astronomy, University College London, London WC1E 6BT, U.K.
{jason.mcewen, boris.leistedt.11}@ucl.ac.uk

***Abstract*—We review the Fourier-Laguerre transform, an alternative harmonic analysis on the three-dimensional ball to the usual Fourier-Bessel transform. The Fourier-Laguerre transform exhibits an exact quadrature rule and thus leads to a sampling theorem on the ball. We study the definition of convolution on the ball in this context, showing explicitly how translation on the radial line may be viewed as convolution with a shifted Dirac delta function. We review the exact Fourier-Laguerre wavelet transform on the ball, coined *flaglets*, and show that flaglets constitute a tight frame.**

***Index Terms*—Harmonic analysis, sampling, wavelets, three-dimensional ball.**

I. Introduction

Data often live naturally on the three-dimensional ball. For example, in cosmology the distribution of galaxies that traces the large-scale structure of the Universe is observed on the celestial sphere (*e.g.* [1]), augmented with depth information given by redshift. A spherical shell at a given redshift represents a given epoch in the history of our Universe; thus, such data live naturally on the three-dimensional ball (hereafter referred to as simply the *ball*).

One would like to analyse such data-sets on the ball to study the physics responsible for them. Since many physical processes are manifest on different physical scales, while also spatially localised, wavelet analysis is a power method for this purpose. Recently, two wavelet transforms have been derived on the ball [4], [5]. The former [4] is based on an undecimated wavelet construction, built on the Fourier-Bessel transform. The latter [5] is based on a tiling of harmonic space, built on a Fourier-Laguerre transform, and developed by the authors of the current article. Our approach [5]: (i) yields wavelets that are not isotropic but rather exhibit an angular opening that is invariant under radial translation; (ii) is theoretically exact; and (iii) leads to a fast multiresolution algorithm.

In this article we review our recent work [5] where we consider the Fourier-Laguerre transform and construct wavelets (which we coin *flaglets*) on the ball. Furthermore, we illuminate the translation operator on the radial line, showing how this may be viewed as convolution with a shifted Dirac delta function. We also show that flaglets constitute a tight frame.

II. Fourier-Laguerre transform

The canonical harmonic transform on the ball is the Fourier-Bessel transform, where the basis functions are the eigenfunctions of the Laplacian on the ball. The Fourier-Bessel basis functions separate into the usual spherical harmonic functions on the sphere and the spherical Bessel functions on the radial line. However, the Fourier-Bessel transform suffers from a serious shortcoming. To the best of our knowledge there does not exist a sampling theorem for the Fourier-Bessel transform, since there does not exist an exact quadrature rule for the evaluation of the spherical Bessel transform (the radial part of the Fourier-Bessel transform).

To overcome this limitation we consider the Fourier-Laguerre transform, for which we developed a sampling theorem [5]. The Fourier-Laguerre transform follows by adopting the Laguerre polynomials (the standard orthogonal polynomials on $\mathbb{R}^+$) as the radial basis functions, while keeping the spherical harmonics as the spherical basis functions. We define the Fourier-Laguerre basis functions on the ball $\mathbb{B}^3 = \mathbb{R}^+ \times \mathbb{S}^2$ by

$$Z_{\ell mp}(\boldsymbol{r}) = K_p(r) Y_{\ell m}(\theta, \varphi), \tag{1}$$

with spherical coordinates $\boldsymbol{r} = (r, \theta, \varphi) \in \mathbb{B}^3$, where $r \in \mathbb{R}^+$ denotes radius, $\theta \in [0, \pi]$ colatitude and $\varphi \in [0, 2\pi)$ longitude, and where $\ell, p \in \mathbb{N}_0$ and $m \in \mathbb{Z}$ such that $|m| \leq \ell$. The standard spherical harmonics are denoted by $Y_{\ell m}$ and the normalised spherical Laguerre basis functions are defined on the radial line by

$$K_p(r) \equiv \sqrt{\frac{p!}{(p+2)!}} \frac{e^{-r/2\tau}}{\sqrt{\tau^3}} L_p^{(2)}\left(r/\tau\right), \tag{2}$$

where $L_p^{(2)}$ is the p-th generalised Laguerre polynomial of order two and $\tau \in \mathbb{R}^+$ is a radial scale factor.

A square-integrable signal $f \in \mathrm{L}^2(\mathbb{B}^3)$ can then be decomposed as

$$f(\boldsymbol{r}) = \sum_{p=0}^{\infty} \sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{\ell} f_{\ell mp} Z_{\ell mp}(\boldsymbol{r}), \tag{3}$$

where the harmonic coefficients are given by the usual projection

$$f_{\ell mp} = \langle f | Z_{\ell mp} \rangle_{\mathbb{B}^3} = \int_{\mathbb{B}^3} \mathrm{d}^3\boldsymbol{r} f(\boldsymbol{r}) Z^*_{\ell mp}(\boldsymbol{r}), \tag{4}$$

where $\mathrm{d}^3\boldsymbol{r} = r^2 \sin\theta \, \mathrm{d}r \, \mathrm{d}\theta \, \mathrm{d}\varphi$ is the usual rotation invariant measure in spherical coordinates.[1] We consider band-limited

[1]This measure is a natural choice since it allows the Fourier-Laguerre transform to be related directly to the Fourier-Bessel transform, such that the Fourier-Bessel coefficients can be computed exactly from Fourier-Laguerre coefficients (see [5] for further details).

signals, with angular and radial band-limits L and P, respectively, *i.e.* signals f such that $f_{\ell mp} = 0$, $\forall \ell \geq L$, $\forall p \geq P$. In this case the summations in Eqn. (3) over ℓ and p may be truncated to $L-1$ and $P-1$ respectively.

In practice, computing the Fourier-Laguerre transform involves the evaluation of the integral of Eqn. (4). An exact quadrature rule for the evaluation of this integral for a band-limited function f naturally gives rise to a sampling theorem. Since the Fourier-Laguerre transform is separable in angular and radial coordinates, we may appeal to separate sampling theorems on the sphere and radial line. For the angular part, we adopt the equiangular sampling theorem on the sphere developed recently by one of the authors [7]. Other sampling theorems on the sphere could alternatively be adopted (*e.g.* [2]), however we select the sampling theorem developed by [7] since it leads to the most efficient sampling of the sphere (*i.e.* the fewest number of samples to represent a band-limited signal exactly). For the radial part, we appeal to Gaussian quadrature to develop an exact quadrature rule and, consequently, a sampling theorem [5]. Combining these results we recover a sampling theorem and, equivalently, an exact Fourier-Laguerre transform on $\mathbb{B}^3$. For a band-limited signal all of the information content of the signal is captured in $N = P[(2L-1)(L-1)+1] \sim 2PL^2$ samples on the ball [5].

We have developed the public `FLAG`[2] code [5] to compute the Fourier-Laguerre transform. The `FLAG` code computes exact forward and inverse Fourier-Laguerre transforms at machine precision and is stable to extremely large band-limits, relying on the public `SSHT`[3] code [7] developed by one of the authors for the angular part, which in turn relies on `FFTW`[4]. `FLAG` supports both the C and Matlab programming languages.

III. Convolution on the ball

We review the definition of convolution on the ball [5], highlighting how the translation operator defined on the radial line may be viewed as convolution with a Dirac delta function. By the angular and radial separability of the Fourier-Laguerre transform, we construct a convolution operator on the ball from convolution operators on the sphere and radial line (*e.g.* [3]).

On the sphere, we adopt the usual convolution of $f \in \mathrm{L}^2(\mathbb{S}^2)$ with an axisymmetric kernel $h \in \mathrm{L}^2(\mathbb{S}^2)$ given by the inner product (*e.g.* [8])

$$\begin{aligned}(f \star h)(\theta,\varphi) &\equiv \langle f | \mathcal{R}_{(\theta,\varphi)} h \rangle_{\mathbb{S}^2} \\ &= \int_{\mathbb{S}^2} \mathrm{d}\Omega(\theta',\varphi') f(\theta',\varphi') \left(\mathcal{R}_{(\theta,\varphi)} h\right)^* (\theta',\varphi'),\end{aligned} \tag{5}$$

where $\mathrm{d}\Omega(\theta,\varphi) = \sin\theta \,\mathrm{d}\theta \,\mathrm{d}\varphi$ is the usual rotation invariant measure on the sphere. The translation operator on the sphere is given by the standard three-dimensional rotation: $(\mathcal{R}_{(\alpha,\beta,\gamma)} h)(\theta,\varphi) = h(\mathcal{R}^{-1}_{(\alpha,\beta,\gamma)}(\theta,\varphi))$, with $(\alpha,\beta,\gamma) \in \mathrm{SO}(3)$, where $\alpha \in [0,2\pi)$, $\beta \in [0,\pi]$ and $\gamma \in [0,2\pi)$. We make the association $\theta = \beta$ and $\varphi = \alpha$, *i.e.* $\mathcal{R}_{(\theta,\varphi)} \equiv \mathcal{R}_{(\alpha,\beta,0)}$, and restrict our attention to convolution with axisymmetric functions that are invariant under azimuthal rotation, *i.e.* $\mathcal{R}_{(0,0,\gamma)} h = h$, so that we recover a convolved function $f \star h$ defined on the sphere. In harmonic space, axisymmetric convolution may be written

$$(f \star h)_{\ell m} = \langle f \star h | Y_{\ell m} \rangle_{\mathbb{S}^2} = \sqrt{\frac{4\pi}{2\ell+1}} f_{\ell m} h^*_{\ell 0}, \tag{6}$$

with $f_{\ell m} = \langle f | Y_{\ell m} \rangle_{\mathbb{S}^2}$ and $h_{\ell 0}\delta_{m0} = \langle h | Y_{\ell m} \rangle_{\mathbb{S}^2}$. The generalisation to directional convolution on the sphere is straightforward (see *e.g.* [8]), however we do not present it here since we consider axisymmetric wavelets subsequently.

On the radial line, we consider a convolution operator appropriate for the spherical Laguerre basis. We adopt a convolution similar to that considered by [3] and others (see additional references contained in [3]), although we recover this operator in an alternative manner. Firstly, we define a translation operator $\mathcal{T}$ on the radial line, which is constructed by analogy with the case for the infinite line, for which the standard orthogonal basis is given by the complex exponentials $\phi_\omega(x) = \exp(\mathrm{i}\omega x)$, with $x, \omega \in \mathbb{R}$. Translation of the basis functions on the infinite line is simply defined by the shift of coordinates: $(\mathcal{T}^{\mathbb{R}}_u \phi_\omega)(x) \equiv \phi_\omega(x-u) = \phi^*_\omega(u)\phi_\omega(x)$, with $u \in \mathbb{R}$ and where the final equality follows by the standard rules for exponents. We define translation of the spherical Laguerre basis functions on the radial line by analogy:

$$(\mathcal{T}_s K_p)(r) \equiv K_p(s) K_p(r), \tag{7}$$

where $s \in \mathbb{R}^+$ (since K_p is real we drop the complex conjugation). This leads to a natural harmonic expression for the translation of a radial function $f \in \mathrm{L}^2(\mathbb{R}^+)$:

$$(\mathcal{T}_s f)(r) = \sum_{p=0}^{\infty} f_p K_p(s) K_p(r), \tag{8}$$

implying

$$\left(\mathcal{T}_s f\right)_p = K_p(s) f_p, \tag{9}$$

where $f_p = \langle f | K_p \rangle_{\mathbb{R}^+}$.

With a translation operator to hand, we may define convolution on the radial line of $f, h \in \mathrm{L}^2(\mathbb{R}^+)$ by the inner product

$$(f \star h)(r) \equiv \langle f | \mathcal{T}_r h \rangle_{\mathbb{R}^+} = \int_{\mathbb{R}^+} \mathrm{d}s s^2 f(s) \left(\mathcal{T}_r h\right)(s), \tag{10}$$

from which it follows that radial convolution in harmonic space is given by the product

$$\left(f \star h\right)_p = \langle f \star h | K_p \rangle_{\mathbb{R}^+} = f_p h_p, \tag{11}$$

where $h_p = \langle h | K_p \rangle_{\mathbb{R}^+}$.

Although the definition of the convolution operator on the radial line is complete, we would like to gain further intuition. The action of the translation operator is described in harmonic space through Eqn. (9), which remains somewhat opaque. We would also like to view the translation operator that we have constructed on the radial line in real space.

[2] http://www.flaglets.org/
[3] http://www.jasonmcewen.org/
[4] http://www.fftw.org/

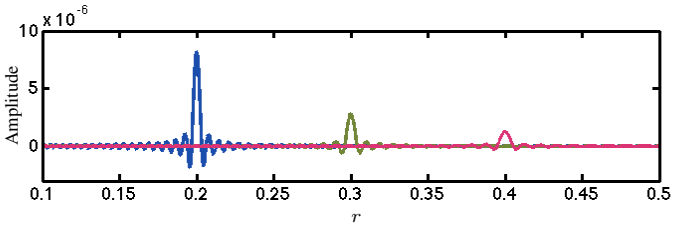


Fig. 1. Band-limited Dirac delta functions plotted on the radial line at positions $s = \{0.2, 0.3, 0.4\}$ (plotted in blue, green and red, respectively). Oscillations are caused by the finite band-limit (here $P = 256$); as $P \to \infty$ oscillations vanish as the band-limited Delta converges to $\delta_s(r) = r^{-2}\delta^{\mathbb{R}}(r-s)$.

In order to recover a real space representation of the radial translation operator we must first consider the Dirac delta function on the radial line. We define the Dirac delta on the radial line at position s by $\delta_s(r) \equiv r^{-2}\delta^{\mathbb{R}}(r-s)$, where $\delta^{\mathbb{R}}$ is the usual Dirac delta defined on the infinite line $\mathbb{R}$. The Dirac delta on the radial line satisfies the following normalisation and sifting properties, respectively:

$$\int_{\mathbb{R}^+} \mathrm{d}r r^2 \delta_s(r) = 1; \tag{12}$$

$$\int_{\mathbb{R}^+} \mathrm{d}r r^2 f(r)\delta_s(r) = f(s). \tag{13}$$

The harmonic expansion of the Dirac delta is given by

$$\delta_s(r) = \sum_{p=0}^{\infty} K_p(s)K_p(r), \tag{14}$$

which follows trivially by the sifting property. For the analysis of band-limited functions, it is sufficient to consider the band-limited Dirac delta (see Fig. 1), where the summation of Eqn. (14) is truncated to $P-1$.

With the Dirac delta function now defined on the radial line, we show that the radial translation operator defined above is simply the convolution of a function with the shifted Dirac delta function:

$$(f \star \delta_s)(r) = \sum_{p=0}^{\infty} f_p K_p(s) K_p(r) = (\mathcal{T}_s f)(r), \tag{15}$$

where the final equality follows by Eqn. (8). Radial convolution and translation are thus the natural analogues of the respective operators defined on the infinite line.

We define the translation operator on the ball by combining the angular and radial translation operators, giving

$$\mathcal{T}_{\boldsymbol{r}} \equiv \mathcal{T}_r \mathcal{R}_{(\theta,\varphi)}. \tag{16}$$

The action of the radial translation operator on functions defined on the ball is shown in Fig. 2. The convolution on the ball of $f \in \mathrm{L}^2(\mathbb{B}^3)$ with an axisymmetric kernel $h \in \mathrm{L}^2(\mathbb{B}^3)$ is then defined by the inner product

$$(f \star h)(\boldsymbol{r}) \equiv \langle f | \mathcal{T}_{\boldsymbol{r}} h \rangle_{\mathbb{B}^3} = \int_{\mathbb{B}^3} \mathrm{d}^3\boldsymbol{s} f(\boldsymbol{s})(\mathcal{T}_{\boldsymbol{r}} h)^*(\boldsymbol{s}), \tag{17}$$

where $\boldsymbol{s} \in \mathbb{B}^3$. In harmonic space, axisymmetric convolution on the ball may be written

$$(f \star h)_{\ell m p} = \langle f \star h | Z_{\ell m p} \rangle_{\mathbb{B}^3} = \sqrt{\frac{4\pi}{2\ell+1}} f_{\ell m p} h^*_{\ell 0 p}, \tag{18}$$

with $f_{\ell m p} = \langle f | Z_{\ell m p} \rangle_{\mathbb{B}^3}$ and $h_{\ell 0 p}\delta_{m0} = \langle h | Z_{\ell m p} \rangle_{\mathbb{B}^3}$.

IV. Flaglets on the Ball

With an exact harmonic transform and a convolution operator defined on the ball in hand, we are now in a position to construct our exact wavelet transform on the ball, which we call the flaglet transform (for Fourier-LAGuerre wavelet transform) [5].

For a function of interest $f \in \mathrm{L}^2(\mathbb{B}^3)$, we define its jj'-th wavelet coefficient $W^{\Psi^{jj'}} \in \mathrm{L}^2(\mathbb{B}^3)$ by the convolution of f with the axisymmetric wavelet, or flaglet, $\Psi^{jj'} \in \mathrm{L}^2(\mathbb{B}^3)$:

$$W^{\Psi^{jj'}}(\boldsymbol{r}) \equiv (f \star \Psi^{jj'})(\boldsymbol{r}) = \langle f | \mathcal{T}_{\boldsymbol{r}} \Psi^{jj'} \rangle_{\mathbb{B}^3}. \tag{19}$$

The scales $j, j' \in \mathbb{N}_0^+$ respectively relate to angular and radial spaces. The wavelet coefficients contain the detail information of the signal only; a scaling function and corresponding scaling coefficients must be introduced to represent the low-frequency, approximate information of the signal. The scaling coefficients $W^{\Phi} \in \mathrm{L}^2(\mathbb{B}^3)$ are defined by the convolution of f with the scaling function $\Phi \in \mathrm{L}^2(\mathbb{B}^3)$:

$$W^{\Phi}(\boldsymbol{r}) \equiv (f \star \Phi)(\boldsymbol{r}) = \langle f | \mathcal{T}_{\boldsymbol{r}} \Phi \rangle_{\mathbb{B}^3}. \tag{20}$$

Provided the flaglets and scaling function satisfy an admissibility property (defined below), the function f may be reconstructed exactly from its wavelet and scaling coefficients by

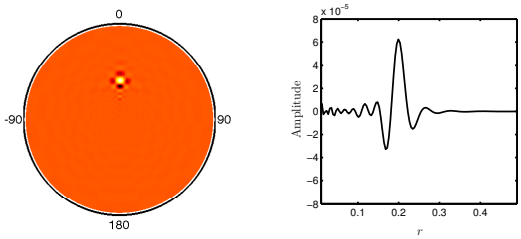
$$\begin{aligned} f(\boldsymbol{r}) = &\int_{\mathbb{B}^3} \mathrm{d}^3\boldsymbol{r}' W^{\Phi}(\boldsymbol{r}')(\mathcal{T}_{\boldsymbol{r}}\Phi)(\boldsymbol{r}') \\ &+ \sum_{j=J_0}^{J} \sum_{j'=J_0'}^{J'} \int_{\mathbb{B}^3} \mathrm{d}^3\boldsymbol{r}' W^{\Psi^{jj'}}(\boldsymbol{r}')(\mathcal{T}_{\boldsymbol{r}}\Psi^{jj'})(\boldsymbol{r}'). \end{aligned} \tag{21}$$

The parameters J_0 and J (J_0' and J') define the minimum and maximum wavelet scales considered respectively for the angular (radial) space and depend on the band-limit of f and the specific definition of the wavelets and scaling function (see [5]).

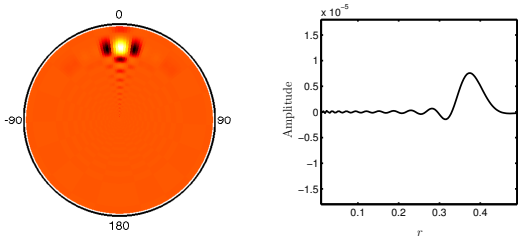
The admissibility condition under which a band-limited function f can be reconstructed exactly is given by the following resolution of the identity:

$$\frac{4\pi}{2\ell+1}\left(|\Phi_{\ell 0 p}|^2 + \sum_{j=J_0}^{J}\sum_{j'=J_0'}^{J'} |\Psi^{jj'}_{\ell 0 p}|^2\right) = 1, \quad \forall \ell, p, \tag{22}$$

where $\Phi_{\ell 0 p}\delta_{m0} = \langle \Phi | Z_{\ell m p} \rangle_{\mathbb{B}^3}$ and $\Psi^{jj'}_{\ell 0 p}\delta_{m0} = \langle \Psi^{jj'} | Z_{\ell m p} \rangle_{\mathbb{B}^3}$. We refer the reader to our previous article [5] for an example of the construction of specific wavelets and scaling functions that satisfy the admissibility condition, where we construct suitable wavelets by tiling the ℓ-p harmonic plane. The resulting wavelets are plotted in Fig. 2.



(a) $\Psi^{jj'}(\boldsymbol{r})$ translated by $r = 0.2$



(b) $\Psi^{jj'}(\boldsymbol{r})$ translated by $r = 0.4$

Fig. 2. Slices of the flaglet $\Psi^{jj'}(\boldsymbol{r})$ with $j = j' = 5$ constructed on the ball of radius $R = 1$ at resolution $P = L = 64$. The three-dimensional flaglets can be visualised by rotating the slices of the left panel (zoomed on a ball of radius $r = 0.5$ for clarity) around the vertical axis passing through the origin. The radial profiles are shown in the right panels. Flaglets are well localised in both real and Fourier-Laguerre spaces. Furthermore, their angular aperture is invariant under radial translation.

We prove that flaglets are a tight frame by showing they satisfy

$$A\|f\|^2_{\mathbb{B}^3} \leq \int_{\mathbb{B}^3} \mathrm{d}^3\boldsymbol{r} |\langle f|\mathcal{T}_{\boldsymbol{r}}\Phi\rangle_{\mathbb{B}^3}|^2 + \sum_{j=J_0}^{J}\sum_{j'=J'_0}^{J'} \int_{\mathbb{B}^3} \mathrm{d}^3\boldsymbol{r} |\langle f|\mathcal{T}_{\boldsymbol{r}}\Psi^{jj'}\rangle_{\mathbb{B}^3}|^2 \leq B\|f\|^2_{\mathbb{B}^3}, \quad (23)$$

with $A = B \in \mathbb{R}^+_*$, for any band-limited $f \in \mathrm{L}^2(\mathbb{B}^3)$, and where $\|\cdot\|^2_{\mathbb{B}^3} \equiv \langle\cdot|\cdot\rangle_{\mathbb{B}^3}$. We adopt a shorthand integral notation in Eqn. (23), although by appealing to our exact quadrature rule these integrals may be replaced by finite sums. Noting the harmonic expression for axisymmetric convolution given by Eqn. (18) and the orthogonality of the Fourier-Laguerre basis functions, it is straightforward to show that the term of Eqn. (23) bounded between inequalities may be written

$$\sum_{p=0}^{P-1}\sum_{\ell=0}^{L-1}\sum_{m=-\ell}^{\ell} \frac{4\pi}{2\ell+1}\left(|\Phi_{\ell 0p}|^2|f_{\ell mp}|^2 + \sum_{j=J_0}^{J}\sum_{j'=J'_0}^{J'} |\Psi^{jj'}_{\ell 0p}|^2|f_{\ell mp}|^2\right) \quad (24)$$
$$= \sum_{p=0}^{P-1}\sum_{\ell=0}^{L-1}\sum_{m=-\ell}^{\ell} |f_{\ell mp}|^2 = \int_{\mathbb{B}^3} \mathrm{d}^3\boldsymbol{r}|f(\boldsymbol{r})|^2 = \|f\|^2_{\mathbb{B}^3},$$

where the second line follows from the admissibility property Eqn. (22). Thus, we find flaglets indeed constitute a tight frame with $A = B = 1$, implying the energy of f is conserved in flaglet space.

We have developed the public FLAGLET[5] code [5] to compute the flaglet transform. The FLAGLET code computes the exact forward and inverse flaglet transform at machine precision, exploiting a fast multiresolution algorithm, and is stable to extremely large band-limits (the computation time and numerical precision of the FLAGLET code is evaluated in detail in [5], where a toy application is also presented). FLAGLET relies on the public code S2LET[6] [6] (to compute wavelet transforms on the sphere), FLAG[5] [5], SSHT[7] [7] and FFTW[8], and supports both the C and Matlab programming languages.

To summarise, flaglets live naturally on the ball (with an angular opening that is invariant under radial translation), yield a theoretically exact wavelet transform on the ball (in both the continuous and discrete settings), and exhibit a fast multiresolution algorithm. It is our hope that flaglets will prove useful for analysing data defined on the ball. Indeed, in the near future we intend to apply flaglets to study the large-scale structure of the Universe traced by the distribution of galaxies.

References

[1] C. P. Ahn et al. The Ninth Data Release of the Sloan Digital Sky Survey: First Spectroscopic Data from the SDSS-III Baryon Oscillation Spectroscopic Survey. *Astrophys. J. Supp.*, 203:21, 2012.

[2] J. R. Driscoll and D. M. Jr. Healy. Computing Fourier transforms and convolutions on the sphere. *Advances in Applied Mathematics*, 15:202–250, 1994.

[3] E. Görlich and C. Markett. A convolution structure for Laguerre series. *Indagationes Mathematicae (Proceedings)*, 85(2):161–171, 1982.

[4] F. Lanusse, A. Rassat, and J.-L. Starck. Spherical 3D isotropic wavelets. *Astron. & Astrophys.*, 540:A92, 2012.

[5] B. Leistedt and J. D. McEwen. Exact wavelets on the ball. *IEEE Trans. Sig. Proc.*, 60(12):6257–6269, 2012.

[6] B. Leistedt, J. D. McEwen, P. Vandergheynst, and Y. Wiaux. S2LET: A code to perform fast wavelet analysis on the sphere. *arXiv:1211.1680*.

[7] J. D. McEwen and Y. Wiaux. A novel sampling theorem on the sphere. *IEEE Trans. Sig. Proc.*, 59(12):5876–5887, 2011.

[8] Y. Wiaux, J. D. McEwen, P. Vandergheynst, and O. Blanc. Exact reconstruction with directional wavelets on the sphere. *Mon. Not. Roy. Astron. Soc.*, 388(2):770–788, 2008.

[5] http://www.flaglets.org/
[6] http://www.s2let.org/
[7] http://www.jasonmcewen.org/
[8] http://www.fftw.org/

Truncation Error in Image Interpolation

Loic Simon
GREYC CNRS UMR 6072
Ecole Nationale Supérieure d'Ingénieurs de Caen
Email:loic.simon@ensicaen.fr

***Abstract*—Interpolation is a fundamental issue in image processing. In this short paper, we communicate ongoing results concerning the accuracy of two landmark approaches: the Shannon expansion and the Discrete Fourier Transform (DFT) interpolation. Among all sources of error, we focus on the impact of spatial truncation. Our estimations are expressed in the form of upper bounds on the Root Mean Squared Error as a function of the distance to the image border. The quality of these bounds is appraised through experiments driven on natural images.**

I. Introduction

Regardless of their very digital nature, images must often be considered continuous. To some extent that we shall discuss, this conceptual "equivalence" is justified by the Shannon-Whittaker theorem. By any means, it is paramount for any application of image processing where sub-pixel operations are performed (such as in optical flow or stereopsis).

However fundamental, the Shannon-Whittaker theorem is by nature deceptive when considering practical circumstances for digital signals are noisy, possibly aliased and more importantly finite. As a result, any practical continuous reconstruction of such signals will be flawed. Among other error sources, one can list photon counting, quantization, aliasing and spatial truncation ([1]). The first three can be harnessed by different means. Photon counting noise can be lowered by increasing the exposure time, while quantization and aliasing are well controlled in recent High Dynamic Range (HDR) cameras.

On the contrary, the last source of error will prove to be much more troublesome. It is indeed the main goal of these notes to alert the readers on this issue. We will also give evidence that this is especially true for images, due to their relatively narrow spatial extension and to their slow spectral decay (mainly when textures are present).

It is rather awkward that the truncation error is often entirely ignored in image processing. It was nonetheless studied in other communities of signal processing. This is for example the case of [2], [3], [4] and [5]. These articles are all dedicated to the truncation error. They include upper bounds valid under diverse circumstances. Unfortunately, because they were developed in different contexts, these results are not so well adapted to images.

In [2] for instance, the signal is assumed bounded. This is certainly true for images since they are encoded on the range between 0 and 255. However in practice their bound yields a large overestimation because its tightness is proportional to the signal dynamic which often exceeds greatly the signal local variability. In [3], [5] the signal is assumed over-sampled, a case often referred to as the guard band assumption in the literature. Such an assumption may be realistic for audio signals but not for images. Note that in the limit where the guard band vanishes, the upper bounds explode inescapably.

While standing no exception to the previous limitations, the study presented in [6] has yet been very inspirational. It considers signals as stationary random processes and proposes two upper-bounds depending upon whether the signal is over-sampled or not. If not, the upper-bound is proportional to the maximum value of the spectrum. This maximum value is generally large and does not lead to a practical upper-bound.

Let us mention also [7], where the problem of Shannon-Whittaker interpolation is directly posed for duration-limited signals. Instead of considering convergence upon an infinite number of samples, the authors let the sampling rate tend to infinity. As a result, no band-limited assumption is required on the signal. The counterpart is that upper bounds are derived and expressed in term of the modulus of continuity of the signal. Such a property can only be known in certain application domains, and certainly not in classical image processing.

All the articles we have mentioned so far concentrate on the Shannon expansion, while in practice, the DFT interpolation is preferred due to a lesser time complexity. To our knowledge, upper-bounds in that case have only been studied in [8]. Their approach is similar to [3] and hence suffers the same limitations. Since the DFT interpolation is equivalent to the exact Shannon expansion under periodic conditions, a periodic plus smooth decomposition [9] may improve its performance.

It is worth noting that the general study of interpolation error can be considered a sub-field of approximation theory. One fundamental and quite powerful result, known as the Strang-Fix conditions [10], relates the capability of a linear shift invariant approximation system to its order of approximation. It was for instance used by Blu et al. (see [11]) to estimate spline based approximation errors. One should note however that these developments concern shift-invariant (and thus infinite) sampling grids. As a result they do not apply to the truncation error. Moreover, it was shown in [12] that in this context at least, the most accurate approximation methods are not interpolating. In a nut shell, imposing a perfect reconstruction of the signal at the sampling position has a negative effect on the overall reconstruction.

For what concerns us, we shall concentrate our efforts on the truncation error and endeavour to obtain realistic estimations of the actual error. Due to lack of space, results shall be presented in a summarized way (*e.g.* without proof and using

the Landau notation). Further details (proofs and tightness analysis) will be included in a forthcoming publication.

II. Notations and Assumptions

In what follows, X_t stands for a random process (RP), where $t \in \mathbb{R}$ might be either a time or space variable. The Fourier transform of a deterministic signal x_t will be denoted by $\mathcal{F}(x)$ and defined as $\mathcal{F}(x)(\omega) := \frac{1}{2\pi}\int e^{i\omega t}x_t dt$.

All RPs are assumed weakly stationary, in other terms with time-invariant first and second order statistics. For such a process X_t, we will generically denote by $\mu := \mathbb{E}[X_t]$ its average, by $R_X(t) := \mathbb{E}[(X_\tau - \mu)(X_{\tau+t} - \mu)]$ its auto-correlation function and by $d\Psi_X(\omega) := \mathcal{F}(R_X)(\omega)$ its power spectral distribution. All RPs are further assumed strictly Nyquist band-limited, which is to say that $d\Psi_X(\{|\omega| \geq \pi\}) = 0$.

Given a RP X_t, we will denote by $X.\Delta_K := \sum_{|k|\leq K} X_k \delta_k$ the sampled version on the finite grid $\{k \in \mathbb{Z}, |k| \leq K\}$. For a fixed $K > 0$, the number of samples will always be denoted by $N = 2K+1$. We consider linear shift-invariant reconstructions from such a sampled version in the form

$$[(X.\Delta_K) * h_K](t) = \sum_{|k|\leq K} X_k h_K(t-k),$$

where $h_K(t)$ is any function referred to as a reconstruction kernel. In this article, we will mainly consider two examples,

- the Shannon kernel $\mathrm{sinc}(t) := \frac{\sin(\pi t)}{\pi t}$ and
- the DFT (or Dirichlet) kernel $\mathrm{sincd}[K](t) := \frac{\sin(\pi t)}{N\sin(\frac{\pi t}{N})}$.

A. Goal

We will appraise the quality of a given reconstruction based on the Root Mean Squared Error (RMSE),

$$RMSE_{[X,h_K]}(t)^2 := \mathbb{E}\left[\left(X_t - [(X.\Delta_K) * h_K](t)\right)^2\right].$$

Resting upon intuitive observations, we shall highlight two predictable features of the RMSE. First, since any interpolation is supposed to perform perfectly at the sampling locations, the RMSE is likely to oscillate, being null at any sample and maximal approximately midway between successive samples. Besides, a RP can be theoretically recovered through the Shannon expansion, if sampled on an infinite grid. Therefore, we expect the error to be tied to the lack of knowledge outside the finite sampling domain, and as such to diminish as we move farther away from the borders. Accordingly, our goal consists in evaluating the decay (up to an oscillating modulation) of the RMSE as the distance varies. We set

$$\delta(t) := \min(K + \frac{1}{2} - t, K + \frac{1}{2} + t). \tag{1}$$

III. Theoretical RMSE bounds

Theorem 1 (Spectral representation of the RMSE): Let X_t be a RP of average μ and power spectrum $d\Psi_X$, $K < \infty$ and h_K a reconstruction system. Then,

$$RMSE_{[X,h_K]}(t)^2 = MSE_{\mu,h_K}(t) + MSE_{d\Psi_X,h_K}(t),$$

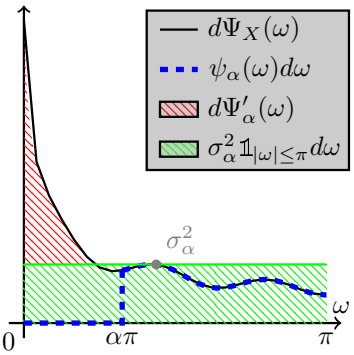


Fig. 1. The spectrum decomposition of Proposition 2.

where

$$MSE_{\mu,h_K}(t) := \mu^2 \left|1 - \Delta_K * h_K(t)\right|^2,$$
$$MSE_{d\Psi,h_K}(t) := \frac{1}{2\pi}\int \left|e^{i\omega t} - [(e^{i\omega\cdot}\Delta_K) * h_K](t)\right|^2 d\Psi(\omega).$$

This theorem merely states that the mean squared error is the sum of the squared errors with respect to the average value of X and with respect to every pure harmonic $e^{i\omega t}$ (weighted by the spectrum of X). The conclusion holds true even if X_t is not band-limited and under mild assumptions (applying to a sequence of h_K's) when $K \to \infty$.

We shall need to evaluate the behavior of each component of the previous decomposition. We refer to them respectively as the average MSE component and the power spectral MSE component. Unlike the previous theorem, the next proposition is specific to strictly Nyquist band-limited RPs.

Proposition 1:

$$MSE_{\mu,h_K}(t) = \mu^2 \left|[\Delta_\infty * \mathrm{sinc}](t) - [\Delta_K * h_K](t)\right|^2,$$
$$MSE_{d\Psi_X,h_K}(t) = \frac{1}{2\pi}\int_{|\omega|\leq\pi} d\Psi_X(\omega) \times \left|[(e^{i\omega\cdot}\Delta_\infty) * \mathrm{sinc}](t) - [(e^{i\omega\cdot}\Delta_K) * h_K](t)\right|^2.$$

Building upon existing works and the analysis of their flaws with respect to specific spectrum characteristics of images, we propose an essential step to obtain realistic bounds. The trick resides in decoupling the low frequencies of the spectrum from a residual component equivalent to band-limited white noise. This process, illustrated in Figure 1, results in

Proposition 2 (Spectrum decomposition): Let $0 \leq \alpha < 1$ and assume that $\mathbb{1}_{|\omega|\geq\alpha\pi} d\Psi_X(\omega) = \psi_\alpha(\omega)d\omega$, with $\psi_\alpha(\omega) \leq \sigma_\alpha^2$. And let $d\Psi'_\alpha$ the positive component of $d\Psi_X - \sigma_\alpha^2 d\omega$. Then,

$$MSE_{d\Psi_X,h_K}(t) \leq MSE_{d\Psi'_\alpha,h_K}(t) + \sigma_\alpha^2 MSE_{\mathbb{1}_{|\omega|\leq\pi}d\omega,h_K}(t).$$

In the previous statement, the first term in the right-hand-side corresponds to an over-sampled signal and the second one to the aforementioned residual band-limited white-noise. In addition, α can be set freely; a freedom we shall exploit to tighten the RMSE bounds which follow.

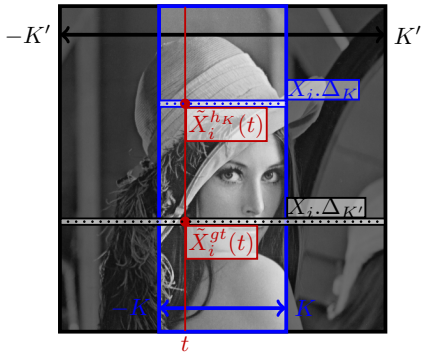
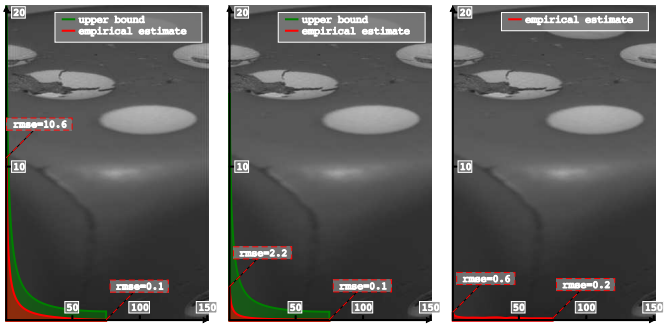


Fig. 2. Construction of the RMSE and spectrum estimators. Interpolation is conducted along the x-axis and averaging along the y-axis.



(a) Shannon (b) DFT (c) B-spline 9

Fig. 3. RMSE estimator and upper bound for the dice image.

Theorem 2: Under the same assumptions and notations as in Proposition 2, and with the kernel h_K associated with either the Shannon expansion or the DFT interpolation, we have

$$RMSE_{[X.\Delta_K * h_K]}(t)^2 = \frac{\sin^2(\pi t)}{\pi^2} \times \begin{pmatrix} \mu_X^2 \mathcal{O}\left(\frac{1}{\delta(t)^2}\right) \\ + \\ {\sigma'}_\alpha^2 \mathcal{O}\left(\frac{1}{\delta(t)^2}\right) \\ + \\ \sigma_\alpha^2 \mathcal{O}\left(\frac{1}{\delta(t)}\right) \end{pmatrix},$$

where

$${\sigma'}_\alpha^2 := \frac{1}{2\pi} \int_{|\omega| \leq \alpha\pi} \frac{2}{1+\cos(\omega)} d{\Psi'}_\alpha(\omega).$$

This provides a decomposition of the mean squared error into a **modulation** times an **envelope**. Following the previous developments, the latter has been decomposed into three terms referred as the **average** envelope, the **low-frequency** envelope and the **white-noise** envelope. This decomposition is valid for the two aforementioned interpolation methods. However the domination constants are different. Precisely, in the DFT, the average envelop is null while the remaining two constants are twice as large as those of the Shannon case.

IV. Experimental settings

In order to scrutinize the correctness of our upper bounds and determine how insightful the information they provide, we have designed an experimental framework[1]. Aspiring to provide practical conclusions on natural images, we could not resort to synthetic signals for which we could have obtained closed-form expressions of the quantities of interest. Instead, given an image exemplar and an interpolation method, our goal would be to estimate the RMSE bound as well as an accurate approximation of the RMSE at varying distances. As illustrated in Figure 2, we will perform interpolation along the horizontal dimension and take advantage of the remaining dimension to perform empirical averages when needed.

The upper bound calculation relies on the image average and two spectral statistics: σ_α and σ'_α. Assuming the spectrum to be absolutely continuous, its density verifies $\psi_X(\omega) = \mathbb{E}\left[|\mathcal{F}(X)(\omega)|\right]^2$. It can thus be estimated at discrete frequencies as an average $\psi_X(\omega_k) \simeq \frac{1}{M}\sum_{i=1}^M |\mathcal{DFT}(X_i)|^2(\omega_k)$.

Assuming that we knew the signal at a given location t, the RMSE could be approached by

$$RMSE^2_{X,h_k}(t) \simeq \frac{1}{M}\sum_{i=1}^M (X_{i,t} - \tilde{X}_{i,t}^{h_K})^2,$$

where to shorten notations $\tilde{X}_{i,t}^{h_K} := [(X_{i,.}\Delta_K) * h_K](t)$. The only challenge here relates to the estimation of the ground-truth interpolated signal. A simple idea would be to subsample an input image, and re-interpolate it with the method under consideration at the missing samples. Obtaining the ground-truth could not be more straightforward. However this scheme does not fulfil other requirements, especially since it violates the Nyquist band-limited assumption.

Instead, as illustrated in Figure 2, starting from an image of half-width K', we restrict the evaluation to a central sub-image of half-width K. That is to say, we apply the interpolation method under test based on this subset of the samples and obtain interpolated samples $\tilde{X}_{i,t}^{h_K}$ in a super-grid of the central region $t \in \{-K, -K+dt, \ldots, K\}$. We then use the whole image to compute (pseudo-)ground-truth samples $\tilde{X}_{i,t}^{gt}$ at the same locations thanks to the Shannon expansion.

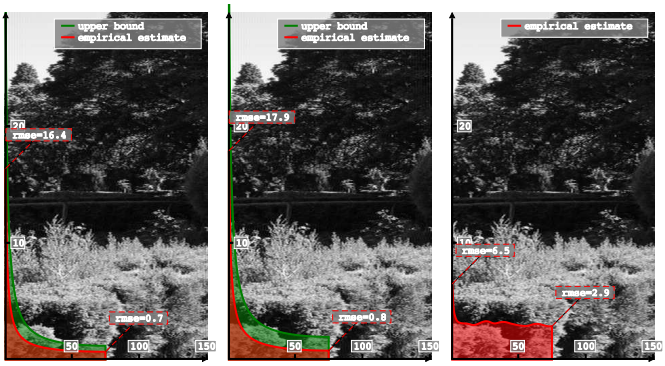
We must point out that the previous strategy has one major drawback. Indeed, since we wish the ground-truth to be much more accurate than the considered interpolation, the margin between the whole image borders and the central region must be large compared to the central extent, *i.e.* $K' \gg K$. Besides, the errors made in $\tilde{X}_{i,t}^{gt}$ and $\tilde{X}_{i,t}^{h_K}$ are due to missing samples, a majority of which are shared. Therefore these errors are correlated and result in a negative bias of the RMSE estimator.

V. Experimental results

Here we present the results on two images. For each image, we plot the RMSE estimator for the Shannon expansion, the DFT and the 9^{th}-order b-spline interpolation, as well as the theoretical upper bound when available.

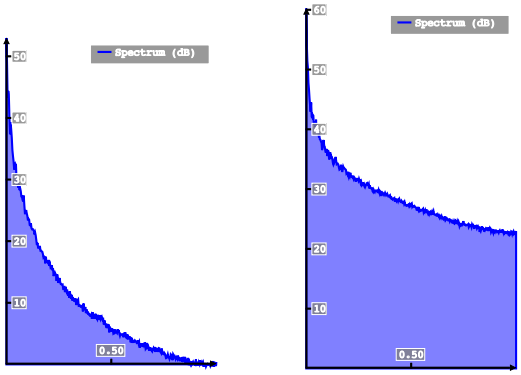
The two images were chosen to illustrate opposite behaviours associated with different spectral contents. Indeed, the first image (Figure 3) is very smooth whereas the second

[1] Available at http://dev.ipol.im/~simonl/ipol_demo/loic_truncate.



(a) Shannon (b) DFT (c) B-spline 9

Fig. 4. RMSE estimator and upper bound for the garden image.



(a) Dice (b) Garden

Fig. 5. Comparison of the spectra in dB for the dice and the garden image.

one (Figure 4) presents various textured regions. We should then expect the white-noise component to be more important in the second case. To confirm this, a visual comparison of the estimated spectra is depicted in Figure 5. In both cases, the upper bound is consistent with the estimator. One cannot fail to observe that the gap between the two curves is greater in the smooth case. This might be explain either by a stronger bias in the estimator or by a lesser sharpness of the upper bound. In any case, consistently with our prediction, the worst-case scenario occurs with highly textured images. It is therefore a great achievement to ensure as tight an estimation in this case. In fact, we have obtained closed-form expressions of the tightness (for white-noise) that confirm our doing so.

Studying closely the values in Figure 4 reveals that for a 150 pixels wide and highly textured image, the interpolation error might very well exceed the quantization (whose RMSE amounts to 0.29) everywhere. The decay of the RMSE as the square root of the inverse distance is then extremely problematic, since it means that to achieve a 2fold decrease of the RMSE the distances must be multiplied by 4. This point brings out dramatic conclusions when considering 16-bits HDR images. Practically, it means that for the same level of accuracy (relatively to the quantization RMSE) the distances must be multiplied by 256^2.

Considering the comparison between the Shannon/DFT methods and B-splines, the most noticeable difference concerns the shape of the RMSE curve. The B-splines error decreases much more quickly and flattens. Unfortunately, the attained value is much larger than in the other methods.

VI. Conclusion

We have presented ongoing results concerning a systematic error which occurs in interpolation. Although similar studies have been published in the past, their knowledge does not seem widely spread among the image processing scientists. More importantly, their applicability to natural images is limited. On the contrary, our study is motivated by practical needs in image processing and is therefore directed toward this specific context. In particular, from the start we took into consideration the possible presence of smooth regions as well as textures.

After presenting the main steps to the theoretical upper bound, we have described an experimental framework and some selected results. A consensus emerged among theoretical and experimental conclusions wherein textured images proved to be a worst-case scenario. The relatively slow decay of the RMSE in that case appears as a major obstacle to highly accurate image processing.

We hope that this paper sheds new light on the legitimacy of the conceptual equivalence of digital and continuous images. It should as well provide a sound starting point to consider accuracy estimations in sub-pixel image processing applications. We do plan to explore such path in the near future.

Acknowledgment

The author would like to thank Jean-Michel Morel for bringing this topic to his attention and the anonymous reviewers for their fruitful suggestions.

References

[1] A. J. Jerri, "The Shannon sampling theorem - its various extensions and applications; a tutorial review," *Proceedings of the IEEE*, vol. 65, 1977.
[2] D. Jagerman, "Bounds for truncation error of the sampling expansion," *SIAM Journal on Applied Mathematics*, vol. 14, 1966.
[3] K. Yao and J. B. Thomas, "On truncation error bounds for sampling representations of band-limited signals," *Aerospace and Electronic Systems, IEEE Transactions on*, 1966.
[4] L. Campbell, "Sampling theorem for the Fourier transform of a distribution with bounded support," *SIAM Journal on Applied Mathematics*, vol. 16, 1968.
[5] J. Brown, "Bounds for truncation error in sampling expansions of band-limited signals," *Information Theory*, vol. 15, 1969.
[6] ——, "Truncation error for band-limited random processes," *Information Sciences*, vol. 1, 1969.
[7] P. Butzer and W. Splettstosser, "A sampling theorem for duration-limited functions with error estimates," *Information and Control*, vol. 34, 1977.
[8] Z. Xu, B. Huang, and X. Li, "On Fourier interpolation error for band-limited signals," *Signal Processing, IEEE Transactions on*, vol. 57, 2009.
[9] L. Moisan, "Periodic plus smooth image decomposition," *Journal of Mathematical Imaging and Vision*, vol. 39, 2011.
[10] G. Strang and G. Fix, "A Fourier analysis of the finite element variational method," *Constructive aspects of functional analysis*, 2011.
[11] T. Blu and M. Unser, "Quantitative fourier analysis of approximation techniques. i. interpolators and projectors," *Signal Processing, IEEE Transactions on*, vol. 47, 1999.
[12] L. Condat, T. Blu, and M. Unser, "Beyond interpolation: Optimal reconstruction by quasi-interpolation," *International Conference on Image Processing*, vol. 1, pp. I–33, 2005.

Optimal Interpolation Laws for Stable AR(1) Processes

Arash Amini and Michael Unser
Biomedical Imaging Group
EPFL, Lausanne, Switzerland
http://bigwww.epfl.ch
Email: {arash.amini, michael.unser}@epfl.ch

***Abstract*—In this paper, we focus on the problem of interpolating a continuous-time AR(1) process with stable innovations using minimum average error criterion. Stable innovations can be either Gaussian or non-Gaussian. In the former case, the optimality of the exponential splines is well understood. For non-Gaussian innovations, however, the problem has been all too often addressed through Monte Carlo methods. In this paper, based on a recent non-Gaussian stochastic framework, we revisit the AR(1) processes in the context of stable innovations and we derive explicit expressions for the optimal interpolator. We find that the interpolator depends on the stability index of the innovation and is linear for all stable laws, including the Gaussian case. We also show that the solution can be expressed in terms of exponential splines.**

I. Introduction

Autoregressive (AR) processes are popular tools for modeling natural phenomena such as speech signals [1]. The processes are usually characterized by an all-pole filter that acts on the innovation process (white excitation noise). They are indexed by the number n of poles of the filter, as AR(n). The AR family contains both stationary and non-stationary models.

The AR processes were historically founded upon Gaussian statistics. Extensions to non-Gaussian scenarios were introduced later, for instance in financial applications, where the data follow a fat-tailed distribution [3], [4]. Besides, fat-tailed distributions are promising models for representing sparse/compressible data [5]. This fact is recently employed in the framework of *sparse stochastic processes* [6], [7] which proposes a unified approach towards Gaussian and non-Gaussian cases.

The estimation problems arising from AR processes are conventionally studied in a finite-dimensional state-space, resulting in the Kalman filter. Under Gaussian statistics, the Kalman filter coincides with the Bayesian estimator (posterior mean estimator) that minimizes the mean-square error. In non-Gaussian scenarios, however, it is common to either apply the Bayesian estimator on approximated posterior distributions [8], [9], [10] or to realize the Bayesian filter numerically [11], [12], [13].

In this paper, we focus on continuous-time AR(1) processes and investigate the problem of Bayesian interpolation between the samples. Our formulation is based on the characteristic forms introduced in [6]. We show that the Bayesian interpolator is linear with respect to the samples when the process follows a symmetric α-stable distribution. The demonstration of linearity is constructive, in the sense that we derive explicit forms for the Bayesian interpolator.

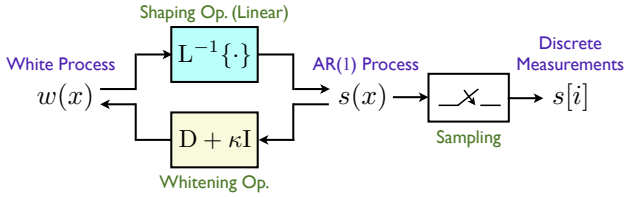


Fig. 1. Generation of the stochastic AR(1) process $s(x)$ based on the excitation white noise $w(x)$. The inverse linear operator L^{-1} includes the possible boundary condition.

II. AR(1) Model

The model in this paper is a special case of [6] adapted for AR(1) processes. The schematic diagram of the continuous-time model is given in Figure 1. The process of interest, s, satisfies the stochastic differential equation

$$\frac{\mathrm{d}}{\mathrm{d}x}s(x)+\kappa s(x)=w(x), \tag{1}$$

where w is a stationary white α-stable excitation with $1\leq\alpha\leq 2$ and $\kappa\in\mathbb{R}$ is a constant. Equation (1) suggests the filter $\mathrm{D}+\kappa\mathrm{I}$ as the whitening operator, where D and I stand for the derivative and identity operators, respectively. This whitening operator has a one-dimensional null space spanned by the function $\mathrm{e}^{-\kappa x}$.

For a proper definition of the process, the shaping operator L^{-1} (inverse of the whitening operator), which transforms the innovations into the main process, needs to be stable. For $\kappa\neq 0$, the system $\mathrm{D}+\kappa\mathrm{I}$ has a unique stable inverse which is shift-invariant and corresponds to the impulse response $\mathrm{e}^{-\kappa x}\chi_{\mathbb{R}_0^+}(x)$ for $\kappa>0$ and $\mathrm{e}^{-\kappa x}\chi_{\mathbb{R}_0^+}(-x)$ for $\kappa<0$, where $\chi_{\mathbb{R}_0^+}(\cdot)$ denotes the characteristic function of the nonnegative real numbers (step function).

For $\kappa=0$, however, there exists no stable inverse. It is shown in [6] that $\mathrm{L}^{-1}=\int_0^x$, which is weakly stable (finite-input finite-output), is a valid choice for $\kappa=0$. Nevertheless, it imposes $s(0)=0$ (boundary condition) and makes the process

s non-stationary. A more general way of setting the boundary condition is given by

$$\mathrm{L}^{-1}\{w\}(x) = \int_0^x w(\tau)\mathrm{d}\tau + \langle w, \phi\rangle, \tag{2}$$

where ϕ is an anti-causal function that decreases rapidly and $\langle w, \phi\rangle = \int w(\tau)\phi(\tau)\mathrm{d}\tau$ in the sense of generalized functions. The anti-causal choice of ϕ shows that, for all $x > 0$, the random variable $\mathrm{L}^{-1}\{w\}(x)$ is statistically independent of $w(\tau)$ for $\tau > x$. This will later help us in simplifying the estimation procedure.

Since the innovation process is white and the impulse response of the shaping operator for $(-\kappa)$ is the time-reversal of the one for $\kappa \neq 0$, we expect to obtain the interpolation results of $(-\kappa)$ by time-reversing the results for κ. Therefore, without loss of generality, we shall assume $\kappa \geq 0$.

Finally, the samples of the AR(1) process are taken at the integers $0, 1, \ldots, m$. They are then used to interpolate the process values in the interval $[0, m]$. For the sake of simplicity, we use $s[k]$ to denote the sample $s(x)|_{x=k}$ for $k = 0, 1, \ldots, m$.

III. Interpolation

Our approach to the interpolation problem is to estimate the process values on a finer grid with spacing T that contains the integers. For this reason, we set $T = \frac{1}{N}$, where N is an arbitrary large positive integer. In this approach, we can get arbitrarily close to any desired point by increasing N. We represent the values $s(x)|_{x=kT}$ for $k = 0, 1, \ldots, mN$, which we want to estimate, by $s_T[k]$. Clearly, $s_T[kN]$ (or $s_1[k]$) represent the known samples and we do not need to estimate them. Since the definition of the process s might include a boundary condition, it is not necessarily stationary, which complicates our analysis. Hence, we prefer to work with the generalized-increment process defined as

$$u_T[k] = s_T[k] - \mathrm{e}^{-\kappa T} s_T[k-1]. \tag{3}$$

To relate the generalized increments to the innovation process, recall that

$$s_T[k] = \mathrm{L}^{-1}w(x)\Big|_{x=kT} = \int_{\eta_\kappa}^{kT} w(\tau)\mathrm{e}^{-\kappa(kT-\tau)}\mathrm{d}\tau + c_{w,\kappa}, \tag{4}$$

where $\eta_\kappa = -\infty$ and $c_{w,\kappa} = 0$ for $\kappa > 0$, and $\eta_\kappa = 0$ and $c_{w,\kappa} = \langle w, \phi\rangle$ for $\kappa = 0$. By substituting $s_T[k]$ and $s_T[k-1]$ from (4) into (3), we see that the null-space term vanishes and we obtain

$$u_T[k] = \int_{(k-1)T}^{kT} w(\tau)\mathrm{e}^{-\kappa(kT-\tau)}\mathrm{d}\tau. \tag{5}$$

The outcome can be written in form of an inner product as

$$u_T[k] = \big\langle w\,,\, \beta_{\kappa,T}(kT - \cdot)\big\rangle, \tag{6}$$

where

$$\beta_{\kappa,T}(x) = \mathrm{e}^{-\kappa x}\Big(\chi_{\mathbb{R}_0^+}(x) - \chi_{\mathbb{R}_0^+}(x-T)\Big). \tag{7}$$

The function $\beta_{\kappa,T}$ is usually known as the exponential B-spline for the grid spacing T [14]. It is supported on $[0, T)$.

A. Preliminary Results

To further proceed in solving the interpolation problem, we need to use a few results regarding the increment process which we state below in the form of 3 lemmas.

Lemma 1: Let $k, k^{'}$ be nonnegative integers and $T, T^{'}$ be positive reals. For the generalized increments u_T we have

(i) $u_T[k]$ and $u_{T'}[k^{'}]$ are independent if $\frac{T}{T'} \notin (\frac{k^{'}-1}{k}, \frac{k^{'}}{k-1})$;
(ii) $u_T[k]$ and $s_{T'}[k^{'}]$ are independent if $\frac{k^{'}}{k-1} \geq \frac{T}{T'}$;
(iii) $u_T[k]$ and $u_T[k^{'}]$ are identically distributed.

Proof From (6) and since $\beta_{\kappa,T}$ is of finite support, we know that the statistics of $u_T[k]$ are completely determined by $w\big((k-1)T < x \leq kT\big)$. Condition (i) guarantees that the parts of the innovation contributing to $u_T[k]$ and $u_{T'}[k^{'}]$ are disjoint. Since the innovation is white, the two are independent. Similarly, Condition (ii) implies disjointness of the innovation parts involved in forming $u_T[k]$ and $s_{T'}[k^{'}]$: the LSI part of $s_{T'}[k^{'}]$, due to the use of causal filters, depends only on $w(x \leq k^{'}T^{'})$, while the boundary condition is fully determined by $w(x \leq 0)$. Thus, for nonnegative $k^{'}$, $s_{T'}[k^{'}]$ is statistically independent of $w(k^{'}T^{'} < x)$. The validity of (iii) is a direct consequence of the stationarity of the innovation. ■

Lemma 2: For any positive integer n, we have that

$$u_{nT}[k] = \sum_{i=0}^{n-1} \mathrm{e}^{-i\kappa T} u_T[kn - i]. \tag{8}$$

Proof We show this property by pointing out the refinement equation of $\beta_{\kappa,nT}$

$$\begin{aligned}\beta_{\kappa,nT}(x) &= \mathrm{e}^{-\kappa x}\big(\chi_{\mathbb{R}_0^+}(x) - \chi_{\mathbb{R}_0^+}(x-nT)\big)\\ &= \sum_{i=0}^{n-1} \mathrm{e}^{-i\kappa T}\mathrm{e}^{-\kappa(x-iT)}\big(\chi_{\mathbb{R}_0^+}(x-iT) - \chi_{\mathbb{R}_0^+}(x-iT-T)\big)\\ &= \sum_{i=0}^{n-1} \mathrm{e}^{-i\kappa T}\beta_{\kappa,T}(x-iT).\end{aligned} \tag{9}$$

Now, it is easy to conclude the claim by applying (9) to (6). ■

Lemma 3: For any positive integer i, we have that

$$s_T[k+i] - \mathrm{e}^{-i\kappa T} s_T[k] = \sum_{\theta=1}^{i} \mathrm{e}^{-(i-\theta)\kappa T} u_T[k+\theta]. \tag{10}$$

Proof The proof requires only the substitution of u_T by its definition in (3). ■

B. Minimum Conditional Mean-Square Error

The well-known minimum mean-square error (MMSE) estimation of a random variable x based on the multidimensional random variable $\mathbf{y}$ (observations) is the function $\hat{x}(\mathbf{y}) = \mathbb{E}\{x|\mathbf{y}\}$ that minimizes the cost $\mathbb{E}\big\{(\hat{x}(\mathbf{y}) - x)^2\big\}$. Note that the averaging applies over both x and $\mathbf{y}$. Consider now that we are estimating x based on a deterministic measurement vector $\mathbf{y}$ that is an observed *realization* of some multivariate random variable. In this case, we should modify the cost to

$\mathbb{E}_x\{(\hat{x}(\mathbf{y})-x)^2|\mathbf{y}\}$, which again results in $\hat{x}(\mathbf{y})=\mathbb{E}\{x|\mathbf{y}\}$ (*i.e.*, the Bayesian estimator). More precisely, the Bayesian estimator $\hat{x}(\mathbf{y})=\mathbb{E}\{x|\mathbf{y}\}$ not only minimizes the average quadratic cost over all realizations, but also minimizes the cost for every individual realization. The distinction is revealed when $\mathbf{y}$ follows a heavy-tail distribution with infinite variance (*e.g.*, a non-Gaussian α-stable). Here, the cost function for each realization $\mathbf{y}$ might be finite while the average over all $\mathbf{y}$ often does not exist. In other words, the conditional expectation defines an optimal estimator for the modified cost, while the MSE might not be defined. It is obvious that the Bayesian estimator coincides with the MMSE estimator when it exists.

With respect to the conditional MSE criterion, the optimal interpolation for $s_T[k]$, using the given samples $s[l]_{l=0}^m$, is given by $\mathbb{E}\{s_T[k]\,|\,s[l]_{l=0}^m\}$. By using Lemma 3, for $0\leq k<m$ and $0<i<N$ where $T=\frac{1}{N}$, we have that

$$\mathbb{E}\Big\{s_T[kN+i]\,\Big|\,\{s[l]\}_{l=0}^m\Big\}-\mathrm{e}^{-i\kappa T}s[k]=\sum_{\theta=1}^{i}\mathrm{e}^{-(i-\theta)\kappa T}\mathbb{E}\Big\{u_T[kN+\theta]\,\Big|\,s[l]_{l=0}^m\Big\}. \quad (11)$$

The one-to-one mapping between the sets $s[l]_{l=0}^m$ and $\{u_1[l]\}_{l=1}^m\cup\{s[0]\}$ allows us to rewrite the conditional expectations as

$$\mathbb{E}\Big\{u_T[kN+\theta]\Big|s[l]_{l=0}^m\Big\}=\mathbb{E}\Big\{u_T[kN+\theta]\,\Big|\,u_1[l]_{l=1}^m,s[0]\Big\}. \quad (12)$$

It follows from (12) and Lemma 1 that $u_T[kN+\theta]$ is independent of $s[0]$ and $u_1[l]_{l=1}^m$ except for $l=k+1$. Thus,

$$\mathbb{E}\Big\{s_T[kN+i]\,\Big|\,s[l]_{l=0}^m\Big\}-\mathrm{e}^{-i\kappa T}s[k]=\textstyle\sum_{\theta=1}^{i}\mathrm{e}^{-(i-\theta)\kappa T}\mathbb{E}\Big\{u_T[kN+\theta]\,\Big|\,u_1[k+1]\Big\}. \quad (13)$$

To simplify the notations, we represent the random variables $u_T[kN+\theta]$ by X_θ and the weights $\mathrm{e}^{-\theta\kappa T}$ by d_θ. Lemma 1 shows that X_θ are i.i.d., and from Lemma 2 we know that

$$u_1[k+1]=\sum_{l=1}^{N}\mathrm{e}^{-(N-l)\kappa T}u_T[kN+l]=\sum_{l=1}^{N}d_{N-l}X_l. \quad (14)$$

Hence,

$$\mathbb{E}\Big\{u_T[kN+i]\,\Big|\,u_1[k+1]\Big\}=\mathbb{E}\Big\{X_i\,\Big|\,\textstyle\sum_{l=1}^{N}d_{N-l}X_l\Big\}. \quad (15)$$

The summary of the results in (11)–(15) is

$$\hat{s}_T[kN+i]=\frac{s[k]}{\mathrm{e}^{i\kappa T}}+\frac{\sum_{\theta=1}^{i}\mathbb{E}\{d_{N-\theta}X_\theta\,|\,\sum_{l=1}^{N}d_{N-l}X_l=u_1[k+1]\}}{\mathrm{e}^{(i-N)\kappa T}}. \quad (16)$$

C. Stable Distributions

Up to this point, our results were generic and applicable to all innovation models. We now concentrate on the symmetric α-stable innovations and try to extract the conditional expectations explicitly. For an α-stable innovation w, the inner product $\langle w,\varphi\rangle$ follows an α-stable distribution for any acceptable test function φ[15]. In particular, the distribution of u_T (or X_θ) is α-stable from (6). If we denote the probability density and characteristic functions (Fourier transform of the density function) of u_T by p_X and $\hat{p}_X$, respectively, the α-stable law implies $\hat{p}_X(\omega)=\exp\big(-\sigma|\omega|^\alpha\big)$ for some positive real σ. Unfortunately, there is no closed form for the density function in general. In addition, the characteristic function of the random variable $\sum_i c_iX_i$, which again follows an α-stable distribution, is given by $\exp\big(-\sigma|\omega|^\alpha\sum_i|c_i|^\alpha\big)$ [15]. This shows that, if $Y_1=d_{N-\theta}X_\theta$ and $Y_2=\sum_{l=1,l\neq\theta}^{N}d_{N-l}X_l$, then we should have

$$\left\{\begin{array}{rcl}\hat{p}_{Y_1}&=&\exp\big(-\sigma|\omega|^\alpha|d_{N-\theta}|^\alpha\big),\\ \hat{p}_{Y_2}&=&\exp\big(-\sigma|\omega|^\alpha\sum_{l=1,l\neq\theta}^{N}|d_{N-l}|^\alpha\big).\end{array}\right. \quad (17)$$

Note that Y_1 and Y_2 are independent and that the conditional expectations in (16) are equal to

$$\mathbb{E}\Big\{d_{N-\theta}X_\theta\,\Big|\,\sum_{l=1}^{N}d_{N-l}X_l=u_1[k+1]\Big\}=\mathbb{E}\Big\{Y_1\,\Big|\,Y_1+Y_2=u_1[k+1]\Big\}=\frac{\int_{\mathbb{R}}y\;p_{Y_1}(y)\;p_{Y_2}\big(u_1[k+1]-y\big)\;\mathrm{d}y}{p_{Y_1+Y_2}\big(u_1[k+1]\big)}. \quad (18)$$

The latter integral can be converted to the Fourier domain by employing Parseval's theorem, which results in

$$\begin{aligned}&\int_{\mathbb{R}}y\;p_{Y_1}(y)\;p_{Y_2}\big(u_1[k+1]-y\big)\;\mathrm{d}y\\&=\int_{\mathbb{R}}\mathcal{F}_y\Big\{y\;p_{Y_1}(y)\Big\}(\omega)\overline{\mathcal{F}_y\Big\{p_{Y_2}\big(u_1[k+1]-y\big)\Big\}(\omega)}\;\mathrm{d}\omega\\&=\int_{\mathbb{R}}\frac{\mathrm{d}}{\mathrm{d}\omega}\hat{p}_{Y_1}(\omega)\overline{\hat{p}_{Y_2}(\omega)}\frac{\mathrm{e}^{-\mathrm{j}\omega u_1[k+1]}}{\mathrm{j}}\;\mathrm{d}\omega\\&=|d_{N-\theta}|^\alpha\int_{\mathbb{R}}\frac{-\sigma\alpha|\omega|^\alpha\mathrm{e}^{-\mathrm{j}\omega u_1[k+1]-\sigma|\omega|^\alpha\sum_{l=0}^{N-1}|d_l|^\alpha}}{\mathrm{j}\omega}\mathrm{d}\omega.\end{aligned} \quad (19)$$

On one hand, the main message from (18) and (19) is that

$$\frac{\mathbb{E}\Big\{d_{N-\theta}X_\theta\,\Big|\,\sum_{l=1}^{N}d_{N-l}X_l=u_1[k+1]\Big\}}{|d_{N-\theta}|^\alpha}=const., \quad (20)$$

where $const.$ does not depend on θ. On the other hand,

$$\sum_{\theta=1}^{N}\mathbb{E}\Big\{d_{N-\theta}X_\theta\,\Big|\,\sum_{l=1}^{N}d_{N-l}X_l=u_1[k+1]\Big\}=u_1[k+1]. \quad (21)$$

Now, by combining (20) and (21), we can evaluate the conditional expectations without performing the integration, as

$$\mathbb{E}\Big\{d_{N-\theta}X_\theta\,\Big|\,\sum_{l=1}^{N}d_{N-l}X_l=u_1[k+1]\Big\}=\frac{|d_{N-\theta}|^\alpha u_1[k+1]}{\sum_{l=1}^{N}|d_{N-l}|^\alpha}. \quad (22)$$

The main result of this paper is given in Theorem 1 which is now easy to verify from (16) and (22).

***Theorem** 1:* For the AR(1) process s associated with the whitening operator $\mathrm{D} + \kappa\mathrm{I}$ with α-stable innovations, the optimal Bayesian interpolation at the point $x^* = k + \lambda$, where $0 \leq \lambda \leq 1$ is a rational number and k is a nonnegative integer, depends only on the neighboring samples $s(x = k)$ and $s(x = k+1)$. Moreover, the dependence is linear and can be expressed as

$$\hat{s}(x^*) = \pi_\lambda \, s(k) + \nu_\lambda \, s(k+1), \tag{23}$$

where

$$\begin{cases} \pi_\lambda &= \mathrm{e}^{(\frac{\alpha}{2}-1)\lambda\kappa} \, \frac{\sinh(\frac{\alpha}{2}(1-\lambda)\kappa)}{\sinh(\frac{\alpha}{2}\kappa)}, \\ \nu_\lambda &= \mathrm{e}^{(\frac{\alpha}{2}-1)(\lambda-1)\kappa} \, \frac{\sinh(\frac{\alpha}{2}\lambda\kappa)}{\sinh(\frac{\alpha}{2}\kappa)}, \end{cases} \tag{24}$$

if $\kappa \neq 0$ and, otherwise,

$$\begin{cases} \pi_\lambda &= 1 - \lambda \\ \nu_\lambda &= \lambda. \end{cases} \tag{25}$$

It is interesting that, for $\kappa = 0$ (Lévy process) and independently of the stability index (α), the optimal interpolator is the simple first-degree B-spline. Also, to compare the result with the classical Gaussian theory, we use $\alpha = 2$ and obtain

$$\begin{cases} \pi_\lambda &= \frac{\sinh(1-\lambda)\kappa}{\sinh\kappa}, \\ \nu_\lambda &= \frac{\sinh\lambda\kappa}{\sinh\kappa}. \end{cases} \tag{26}$$

IV. Simulations

To show the impact of our results, we have applied our interpolator to MATLAB simulated data. For this purpose, we have plotted a realization of an α-stable AR(1) process with $\alpha = \frac{3}{2}$ and $\kappa = 5$ in Figure 2. We have used the values at the integers as the samples for interpolating the process. As is evident in Figure 2, the curves connecting the points deviate from straight lines and are not even piecewise monotonic (*e.g.*, the part corresponding to the interval [9,10]). In fact, the statistics of the model show that, for each pair of adjacent samples, the distribution of the values between them is biased in favor of one of the sides of the line connecting the two samples. It is comforting to observe that the curve of the optimal interpolator is bent towards the same direction. From Figure 2, it is evident that the optimal interpolator takes advantage of knowing the system parameters and better follows the process than the outcome of the uninformed first-degree B-spline.

V. Conclusion

In this paper, we studied the interpolation problem for the first-order autoregressive processes generated from stable innovations, including non-Gaussian ones. We applied the Bayesian estimator which minimizes the mean-square error under Gaussian distributions and conditional mean-square error under stable laws that have infinite variance. We derived explicit forms for the optimal interpolator in a general setting and found that it is linear with respect to the samples.

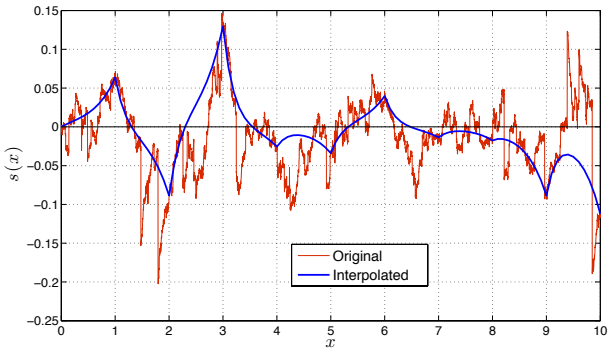


Fig. 2. A realization of the AR(1) process with $\kappa = 5$ and $\alpha = 1.5$, and the interpolated function using the samples at the integers.

Moreover, it depends on the stability index that characterizes stable innovations. Our derivations rely on exponential splines.

Acknowledgment

The research is supported by the European Commission under Grant ERC-2010-AdG 267439-FUN-SP.

References

[1] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. of the IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[2] H. Kirshner, S. Maggio, and M. Unser, "A sampling theory approach for continuous ARMA identification," *IEEE Trans. Sig. Proc.*, vol. 59, no. 10, pp. 4620–4634, Oct. 2011.

[3] P. Embrechts, C. Klüppelberg, and T. Mikosch, *Modeling Extremal Events: For Insurance and Finance (Stochastic Modeling and Applied Probability)*, 2nd ed. Springer, 2008.

[4] C. L. Nikias and M. Shao, *Signal Processing with Alpha-Stable Distributions and Applications*. Wiley-Interscience, 2001.

[5] A. Amini, M. Unser, and F. Marvasti, "Compressibility of deterministic and random infinite sequences," *IEEE Trans. Sig. Proc.*, vol. 59, no. 11, pp. 5193–5201, Nov. 2011.

[6] M. Unser, P. Tafti, and Q. Sun, "A unified formulation of Gaussian *vs.* sparse stochastic processes: Part I—Continuous-domain theory," *arXiv:1108.6150v1 [cs.IT]*, 2011.

[7] M. Unser, P. Tafti, A. Amini, and H. Kirshner, "A unified formulation of Gaussian *vs.* sparse stochastic processes: Part II—Discrete-domain theory," *arXiv:1108.6152v1 [cs.IT]*, 2011.

[8] A. Jazwinski, *Stochastic Processes and Filtering Theory*. Academic Press, 1970.

[9] D. Alspach and H. Sorenson, "Nonlinear Bayesian estimation using Gaussian sum approximations," *IEEE Trans. Auto. Control*, vol. 17, no. 4, pp. 439–448, Aug. 1972.

[10] B. Anderson and J. Moore, *Optimal Filtering*. Englewood Cliffs: Prentice Hall, 1979.

[11] G. Kitagawa, "Non-Gaussian state-space modeling of nonstationary time series," *J. Amer. Stat. Assoc.*, vol. 82, no. 400, pp. 1032–1041, Dec. 1987.

[12] ——, "Monte Carlo filter and smoother for non-Gaussian nonlinear state-space models," *J. Comput. and Graph. Stat.*, vol. 5, no. 1, pp. 1–25, Mar. 1996.

[13] A. Doucet, S. Godsill, and C. Andrieu, "On sequential Monte Carlo sampling methods for Bayesian filtering," *Stat. and Comput.*, vol. 10, no. 3, pp. 197–208, Jul. 2000.

[14] M. Unser and T. Blu, "Cardinal exponential splines: Part I—Theory and filtering algorithms," *IEEE Trans. Sig. Proc.*, vol. 53, no. 4, pp. 1425–1438, Apr. 2005.

[15] G. Samorodnitsky and M. S. Taqqu, *Stable Non-Gaussian Random Processes*. Chapman & Hall/CRC, 1994.

Hierarchical Tucker Tensor Optimization - Applications to Tensor Completion

Curt Da Silva
Seismic Laboratory for Imaging and Modeling
& Department of Mathematics
University of British Columbia
Email: curtd@math.ubc.ca

Felix J. Herrmann
Seismic Laboratory for Imaging and Modeling
& Department of Earth and Ocean Sciences
University of British Columbia
Vancouver, BC, Canada
Email: fherrmann@eos.ubc.ca

***Abstract*—In this work, we develop an optimization framework for problems whose solutions are well-approximated by *Hierarchical Tucker* (HT) tensors, an efficient structured tensor format based on recursive subspace factorizations. Using the differential geometric tools presented here, we construct standard optimization algorithms such as Steepest Descent and Conjugate Gradient for interpolating tensors in HT format. We also empirically examine the importance of one's choice of data organization in the success of tensor recovery by drawing upon insights from the matrix completion literature. Using these algorithms, we recover various seismic data sets with randomly missing sources.**

I. Introduction

Matrix completion has seen a large amount of development in recent years, resulting in algorithms that are very space and time efficient and theoretical guarantees which closely agree with empirical recovery rates. The success of completing a matrix with randomly missing entries via rank-minimizing optimization is a result of assuming a low-rank model on the underlying solution, coupled with a subsampling operator that tends to increase the rank of the underlying matrix.

We use extended notions of low-rank in the case of interpolating a *tensor* with missing entries. Our model is a structured tensor format known as the *Hierarchical Tucker* (HT) format, which efficiently represents a high-dimensional tensor by means of a Kronecker splitting of subspaces, with the set of all such tensors parametrizing a smooth manifold in $\mathbb{R}^{n_1 \times n_2 \times \ldots n_d}$. We extend the largely theoretical results of [1] by imposing a Riemannian metric on the resulting quotient manifold, from which we can derive the Riemannian gradient and develop solvers for minimizing smooth functions defined on this manifold. We will use these efficient, SVD-free solvers in order to interpolate tensors that have a large portion of their entries removed and empirically examine the effect of data organization on the success of recovery for our test seismic cases. Our manifold-optimization approach for completing tensors with missing entries follows a similar spirit to [2]. We present the results of several interpolated seismic frequency slices and demonstrate our ability to recover tensors even amidst high levels of subsampling.

II. Hierarchical Tucker Tensor Format

An important choice of dimension separation, ensuring that the resulting HT tensor is low-rank, is that of a *dimension tree*.

Definition 1. A *dimension tree* for a $d-$dimensional tensor is a nontrivial binary tree such that

- The root node, t_{root}, has the label $\{1, \ldots, d\}$
- The labels for the children of each non-leaf node form a partition of the parent's label, i.e.

$$t_l \sqcup t_r = t, \quad t \notin L$$

where t_l, t_r are the left and right children of the node t, respectively, and L is the set of all leaves of T.

Suppose that we have chosen a set of (positive integer) hierarchical ranks $(k_t)_{t \in T}$ assigned to each node of a dimension tree T, with $k_{\mathrm{t}_{\text{root}}} = 1$. Then we have the following

Definition 2. Let $\mathbb{R}_*^{n \times p}$ and $\mathbb{R}_*^{p \times q \times r}$ denote the set of all $n \times p$ matrices of full rank and $p \times q \times r$ 3-tensors of full *multilinear* rank, respectively.

A $d-$tensor X is said to be in *Hierarchical Tucker* format with associated dimension tree T and hierarchical ranks $(k_t)_{t \in T}$ if there exist parameter matrices/tensors $x = (U_t, B_t)$ with $U_t \in \mathbb{R}_*^{n_t \times k_t}$, $B_t \in \mathbb{R}_*^{k_r \times k_l \times k_t}$ such that $\phi(x) = X$, where

$$\begin{aligned} \operatorname{vec} \phi(x) &= (U_{t_l} \otimes U_{t_r})(B^{(k_l, k_r)}) && t = \mathrm{t}_{\text{root}} \\ U_t &= (U_{t_l} \otimes U_{t_r})(B^{(k_l, k_r)}) && t \notin L \cup \mathrm{t}_{\text{root}} \end{aligned} \tag{1}$$

where k_t is the rank associated to node t and k_l, k_r are the ranks associated to nodes t_l, t_r, respectively. We say that the parameter matrices x are in *Orthogonal Hierarchical Tucker* (OHT) format if (U_t, B_t) also satisfy

$$\begin{aligned} U_t^T U_t &= I_{k_t} \quad \text{for } t \in L \\ (B_t^{(k_l, k_r)})^T B_t^{(k_l, k_r)} &= I_{k_t} \quad \text{for } t \notin L \cup \mathrm{t}_{\text{root}} \end{aligned}$$

Let $\mathcal{H}_{\mathrm{T,k}}$ denote the set of all tensors expressible in HT format with dimension tree T and hierarchical ranks $(k_t)_{t \in T}$.

Note that the intermediate matrices U_t in (1) for $t \notin L$ do not need to be stored: only the matrices U_t for $t \in L$ and so-called *transfer tensors* B_t for $t \notin L$ need to be stored to specify the tensor X completely. Let

$$\mathcal{M} = \bigtimes_{t \in L} \mathbb{R}_*^{n_t \times k_t} \times \bigtimes_{t \in T \setminus L} \mathbb{R}_*^{k_{t_r} \times k_{t_l} \times k_t}$$

be the space of admissible HT parameters. ϕ given in (1) is a smooth function from $\mathcal{M}$ to its image $\mathcal{H}_{\mathrm{T,k}} \subset \mathbb{R}^{n_1 \times \dots n_d}$ that is *not* injective. From the optimization point of view, any optimization problem defined on $\mathcal{H}_{\mathrm{T,k}}$ and parametrized by $\mathcal{M}$ and (1) will have minimizers which are *not* isolated. We will characterize this non-uniqueness, and its remedy, below.

III. Quotient Geometry of the HT format

There is an ambiguity in the representative parameters for a given HT tensor X, which is characterized in [1] as follows. Let $\mathbf{G}$ be the Lie group

$$\mathbf{G} = \{\mathcal{A} = (A_t)_{t \in T} : A_t \in GL(k_t)\ A_{\mathrm{t_{root}}} = 1\}$$

acting on $\mathcal{M}$ via the right action

$$\theta_{\mathcal{A}}(U_t, B_t) := (U_t A_t, (A_{t_r}^{-1}, A_{t_l}^{-1}, A_t^T) \circ B_t)$$

where $(A_1, A_2, A_3) \circ C$ is the multilinear product that premultiplies C by A_i in the i-th dimension. Note that $\phi(x) = \phi(y)$ if and only if there exists a unique $\mathcal{A} \in \mathbf{G}$ such that $y = \theta_{\mathcal{A}}(x)$. The quotient manifold has a unique smooth structure such that $\pi : \mathcal{M} \to \mathcal{M} / \mathbf{G}$ is a smooth submersion. The quotient manifold $\mathcal{M} / \mathbf{G}$ is really our manifold of interest for the purposes of solving optimization problems, since each equivalence class $\pi(x)$ is identified with unique values of $\phi(x)$.

The authors in [1] introduce the following horizontal space

$$\mathcal{H}_x \mathcal{M} := \left\{ (U_t^h, B_t^h) : \begin{array}{c} (U_t^h)^T U_t = 0_{k_t \times k_t} \text{ for } t \in L \\ (B_t^h)^{(k_t)} Q_t (B_t^{(k_t)})^T = 0_{k_t \times k_t} \text{ for } t \notin L \cup \mathrm{t_{root}} \end{array} \right\} \tag{2}$$

where $Q_t = (U_{t_l}^T U_{t_l} \otimes U_{t_r}^T U_{t_r})$, which is shown to be invariant under the action of θ. Eq. (2) allows us to uniquely identify vector fields on $\mathcal{M} / \mathbf{G}$ with horizontal vector fields in $\mathcal{M}$.

For the purposes of interpolation, we are interested in the computing the best fit of our data within the space of HT models, which involves solving a corresponding optimization program on $\mathcal{H}_{\mathrm{T,k}}$. There is a large body of existing research on solving optimization problems on matrix manifolds (see [3] for a comprehensive introduction). Before we can develop such optimization methods, we must first specify a well-defined Riemannian metric on the quotient manifold $\mathcal{M} / \mathbf{G}$.

Fix $x = (U_t, B_t)$, $\eta_x = (\delta U_t, \delta B_t), \zeta_x = (\delta V_t, \delta C_t) \in \mathcal{H}_x \mathcal{M}$. Let $P_t = U_t^T U_t$ for each $t \in T \setminus \mathrm{t_{root}}$, Q_t as above, and let, by abuse of notation, $\delta B_t := \delta B_t^{(k_l, k_r)}$ and similarly for δC_t. One can show that, for the following inner product,

$$\begin{aligned} g_x(\eta_x, \zeta_x) := & \sum_{t \in T} \mathrm{tr}(P_t^{-1} \delta U_t^T \delta V_t) \\ & + \sum_{t \notin L \cup \mathrm{t_{root}}} \mathrm{tr}(P_t^{-1} (\delta B_t)^T Q_t \delta C_t) \\ & + \mathrm{vec}(\delta B_{\mathrm{t_{root}}})^T Q_{\mathrm{t_{root}}} \mathrm{vec}(\delta C_{\mathrm{t_{root}}}) \end{aligned} \tag{3}$$

it holds that $g_x(\eta_x, \zeta_x) = g_{\theta_{\mathcal{A}}(x)}(\eta_{\theta_{\mathcal{A}}(x)}, \zeta_{\theta_{\mathcal{A}}(x)})$ for every $\mathcal{A} \in \mathbf{G}$. Therefore the metric g restricted to vectors in the horizontal space does *not* depend on the representative point for the equivalence class, $x' \in \pi(x)$. Since each $U_t^T U_t$ is symmetric positive definite for each $t \in T \setminus \mathrm{t_{root}}$ and varies smoothly with x, it is easy to see that g_x varies smoothly with x as well. This yields a Riemannian metric that is well-defined on the quotient manifold $\mathcal{M} / \mathbf{G}$ (see 3.6.2 in [3]). Our optimization algorithm will then be implemented on the total space $\mathcal{M}$ rather than the abstract quotient $\mathcal{M} / \mathbf{G}$, with the understanding that points $x \in \mathcal{M}$ will represent their equivalence class $\pi(x) \in \mathcal{M} / \mathbf{G}$ (see [3] for more details).

Require: $x = (U_t, B_t)$, $Z \in \mathbb{R}^{n_1 \times \dots n_d}$
  $\delta U_{\mathrm{t_{root}}} \leftarrow Z$
  for each $t \in T \setminus L$, visiting each node before its children **do**
    $\delta U_{t_l} \leftarrow \frac{\partial U_t}{\partial U_{t_l}}^* \delta U_t$, $\delta U_{t_r} \leftarrow \frac{\partial U_t}{\partial U_{t_r}}^* \delta U_t$,
    $\delta B_t \leftarrow \frac{\partial U_t}{\partial B_t}^* \delta U_t$
  end for
  return $D\phi(x)^* Z = P_{\mathcal{H}_x \mathcal{M}}((\delta U_t)_{t \in L}, (\delta B_t)_{t \in T \setminus L})$

Fig. 1. Algorithm for computing $D\phi(x)^* Z$

When we restrict our parameter matrices to be in OHT, one can see that since $U_t^T U_t = I_{k_t}$ for every $t \in T \setminus \mathrm{t_{root}}$, and so the inner product (3) reduces to the standard Euclidean one. For this reason, and to ensure that the resulting projections on to $\mathcal{H}_x \mathcal{M}$ can be performed efficiently, we restrict our parameters $x = (U_t, B_t)$ to be OHT in the sequel. This is not a hindrance from a theoretical point of view, because any non-orthogonalized parameter set x can be efficiently orthogonalized via Proposition 3 to a parameter set x' such that $\phi(x) = \phi(x')$. It can be shown that the resulting quotient space of orthogonalized parameters is diffeomorphic to $\mathcal{H}_{\mathrm{T,k}}$.

A. Riemannian Gradient

Using this Riemannian metric, we can compute the Riemannian gradient of a smooth function $f : \mathcal{H}_{\mathrm{T,k}} \to \mathbb{R}$ as follows. Let $x \in \mathcal{M}$. Then by the fundamental theorem of linear algebra, since $\mathrm{im}\ D\phi(x) = T_{\phi(x)}\mathcal{H}$, $\ker D\phi(x)^* = T_{\phi(x)}^{\perp}$

Our Riemannian gradient in this case can be easily seen as $Z = D\phi(x)^* \mathrm{grad} f(\phi(x))$, since for any $\xi \in \mathcal{H}_x \mathcal{M}$,

$$\begin{aligned} \langle Z, \xi \rangle &= \langle D\phi(x)^* \mathrm{grad} f(\phi(x)), \xi \rangle \\ &= \langle P_{T_{\phi(x)}\mathcal{H}} \mathrm{grad} f(\phi(x)), D\phi(x)[\xi] \rangle \\ &= Df(\phi(x)) \circ D\phi(x)[\xi] \\ &= Df(\phi(x))[\xi] \end{aligned}$$

The adjoint of $D\phi(x)$ can be computed using that, for $t \in T$,

$$\delta U_t = \frac{\partial U_t}{\partial U_{t_l}} \delta U_{t_l} + \frac{\partial U_t}{\partial U_r} \delta U_{t_r} + \frac{\partial U_t}{\partial B_t} \delta B_t$$

and $D\phi(x)[\xi] = \mathrm{vec}(\delta U_{\mathrm{t_{root}}})$. The adjoint of this recursion, followed by a projection on to (2), gives us Figure 1.

Since U_t in (1) is linear in each variable, one can write out the partial derivatives of U_t with respect to U_{t_l}, U_{t_r} and B_t by considering the possible matricizations of U_t

$$\begin{aligned} U_t^{(k_r)} &= U_r B_t^{(k_r)} (U_{t_l} \otimes I_{k_t})^T \\ U_t^{(k_l)} &= U_l B_t^{(k_l)} (U_{t_r} \otimes I_{k_t})^T \end{aligned}$$

and using the matrix calculus product rule

$$\frac{\partial(AB)}{\partial X} = (B^T \otimes I_{M_A})\frac{\partial A}{\partial X} + (I_{N_B} \otimes A)\frac{\partial B}{\partial X}$$

to isolate for the corresponding differential. We will not go into the full derivation here due to space constraints. The final result is a very simple set of MATLAB commands, which uses code from the SPOT framework [4] and from the hTucker toolbox [5], that requires only matrix-matrix multiplications and permutations of relatively small matrices, which can be performed efficiently.

IV. Optimization Algorithms

Let $\mathcal{M}$ be the space of parameters for the OHT format with the corresponding Lie group of orthogonal matrices $\mathcal{G} \leq \mathbf{G}$ acting on $\mathcal{M}$ via θ. For the purpose of interpolation, we are interested in solving

$$x^* = \underset{x=(U_t,B_t)}{\arg\min}\, f(x) = \|A\phi(x) - b\|_2^2 \quad (4)$$
$$\text{s.t.} \quad U_t^T U_t = I_{k_t}, (B_t^{(k_l,k_r)})^T B_t^{(k_l,k_r)} = I_{k_t}$$

where A is our subsampling operator and b is our subsampled data. For a Steepest Descent-type method, we have a means to compute the Riemannian gradient of f at a point x, which we will denote g_x. In order to move along $-g_x$ for some step size t, we need a retraction on $\mathcal{M}$, which is a first-order approximation to the exponential mapping on $\mathcal{M}$.

Proposition 3. Let $x = (U_t, B_t) \in \mathcal{M}$, $\eta = (\delta U_t, \delta B_t) \in T_x\mathcal{M}$. Then the reorthogonalization mapping R, introduced in [6], and defined by

$$R_x(\eta) = \begin{cases} \text{qf}(U_t + \delta U_t) & \text{if } t \in L \\ \text{qf}((R_{t_l} \otimes R_{t_r})(B_t + \delta B_t)) & \text{if } t \notin \text{t}_{\text{root}} \cup L \\ (R_{t_l} \otimes R_{t_r})(B_t + \delta B_t) & \text{if } t = \text{t}_{\text{root}} \end{cases}$$

where $\text{qf}(X), R_t$ are the Q-factor from the QR factorization of X and R_t is the R-factor from the QR factorization associated to node t, is a retraction on $T\mathcal{M}$.

$R_x(\eta)$ can be computed very efficiently, in the sense that one avoids operating on the full tensor space $\mathbb{R}^{n_1 \times n_2 \times \ldots n_d}$ and instead one performs QR factorizations on relatively small matrices. Since R is a retraction on the tangent bundle $T\mathcal{M}$ and $\mathcal{H}_x\mathcal{M}$ in (2) is a θ-invariant horizontal distribution on $\mathcal{M}$, by 4.1.2 in [3] we have that the mapping $\tilde{R}_{\pi(x)}(\xi_{\pi(x)}) = \pi(R_x(\xi_x))$ is a well-defined retraction on $T(\mathcal{M}/\mathcal{G})$.

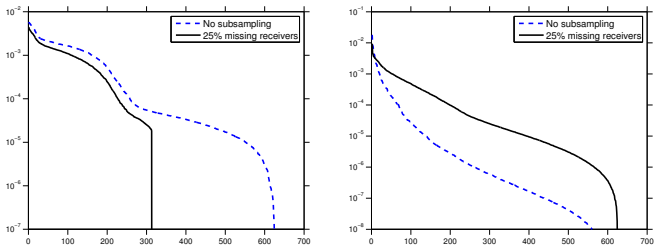
Using this retraction, we formulate the steepest descent algorithm using an Armijo line search in a straightforward manner, presented in Figure 2. We can easily modify this framework to implement other first-order methods such as CG, which we will use for our numerical examples.

V. Multidimensional Subsampling

As we use seismic data examples for our recovery, it should be noted that 3D seismic data is five dimensional, with two source coordinates (x, y), two receiver coordinates (x, y), and time, from which we extract a single, 4D frequency slice by taking the Fourier transform in time and fixing a frequency. Owing to the symmetric nature of seismic data between sources and receivers, we have essentially two choices of underlying dimension tree, both depicted in Figure 3. Namely, we can choose between placing the (src x, src y) dimensions in the rows and (rec x, rec y) dimensions in the columns, or placing the (src x, rec x) dimensions in the rows and (src y, rec y) dimensions in the columns (each choice specifies the rest of the dimension tree). In the case when we are, say, randomly missing sources, the former organization of data has the effect that subsampling will tend to remove rows of this matrix, and hence the singular values will not increase and in fact *are set to zero* at the low end (the worst-case scenario for the purposes of rank-minimizing recovery, e.g. see [7]). On the other hand, the latter organization of data results in a subsampling operator that randomly removes blocks from the underlying matrix, which is a much more favourable situation from a low-rank recovery perspective, as we can see from the singular values of the resulting matrix. The same situation holds for matricizations in the singleton dimensions, adding further degrees of regularity to the computed solution compared to standard matrix completion. Our choice of dimension tree is of great importance in the success of our recovery.

Require: Initial guess $x_0 = (U_t, B_t)$, $0 < c < 1$ sufficient decrease parameter, $0 < \theta < 1$ step size decrease
for $k = 0, 1, 2, \ldots$ until convergence **do**
  $\mathbf{X}_k \leftarrow \phi(x_k)$
  $f_k \leftarrow f(\mathbf{X}_k)$
  $g_k \leftarrow \nabla_x f(\phi(x_k))$ //Riemannian gradient of f at x_k
  $\alpha \leftarrow 1$ //Armijo line search
  while $f(\phi(R_{x_k}(-\alpha g_k))) - f_k > -c\alpha\langle g_k, g_k\rangle$ **do**
    $\alpha \leftarrow \alpha \cdot \theta$
  end while
  $x_{k+1} \leftarrow R_{x_k}(-\alpha g_k)$
end for

Fig. 2. Steepest descent for optimizing a function f over the manifold $\mathcal{H}_{\text{T,k}}$

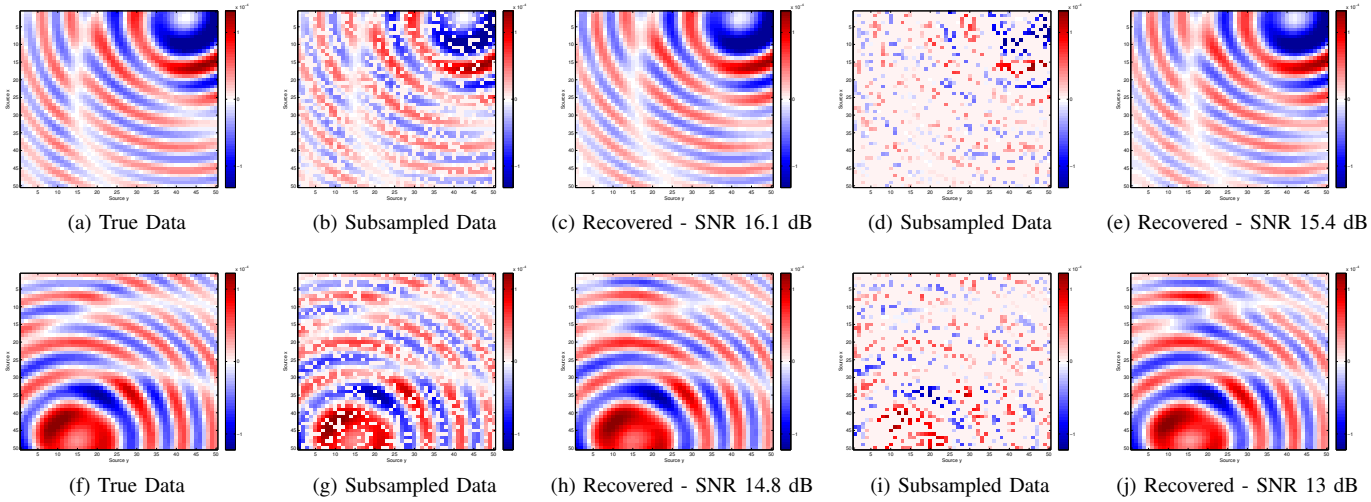


(a) (b)

Fig. 3. Singular values for the *Left*: (src x, src y) Matricization *Right*: (src x, rec x) Matricization of a test data set. *Blue*: Without subsampling, *Black*: With subsampling

VI. Numerical Experiments

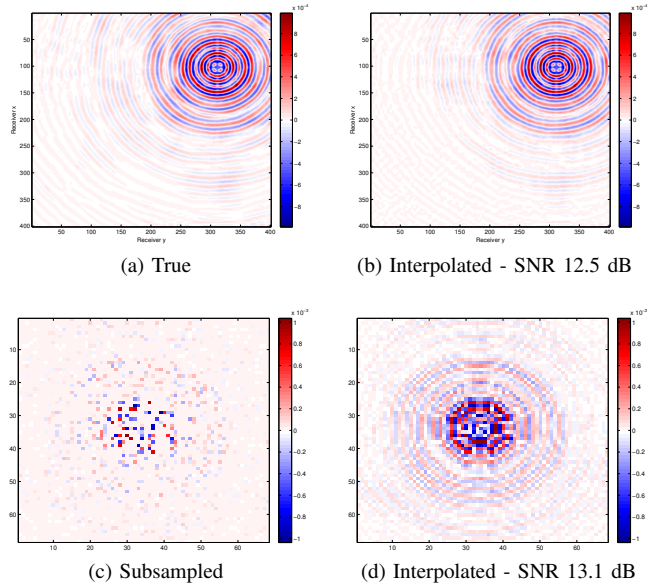
In the following examples, we apply our algorithms to interpolate seismic frequency slices from two test sets. In the first set, we use data generated from a simple single-reflector model, while the second set has been provided to us by British Gas (BG), generated from an unknown model. For our solver, we implement nonlinear CG in this OHT framework, using



(a) True Data (b) Subsampled Data (c) Recovered - SNR 16.1 dB (d) Subsampled Data (e) Recovered - SNR 15.4 dB

(f) True Data (g) Subsampled Data (h) Recovered - SNR 14.8 dB (i) Subsampled Data (j) Recovered - SNR 13 dB

Fig. 4. *Top*: (Rec x, Rec y) = (5,45). (b), (c), (g), (h) are results for 25% source subsampling, (d), (e),(i), (j) are results for 75% source subsampling



(a) True (b) Interpolated - SNR 12.5 dB

(c) Subsampled (d) Interpolated - SNR 13.1 dB

Fig. 5. Interpolated BG Data for 75% missing sources. *Top*: Fixed, unknown source image, *Bottom*: Fixed receiver image

the orthogonalization retraction in (3) and projection onto the horizontal space (2) as a vector transport.

The simple data set has size $D \in \mathbb{R}^{50\times50\times50\times50}$ and we randomly remove source (x, y) pairs from the data set before recovery. We run the resulting algorithm for 200 iterations starting from a random initial guess, which produces the results in Figure 4. Even amidst high levels of missing sources, the HT construction is able to sufficiently regularize the interpolation process to successfully recover each slice for fixed receiver coordinates (known as a common receiver gather in seismic circles).

The BG data set originally has 68 x 68 sources corresponding to 401 x 401 receivers, from which we remove a subset of the sources randomly and interpolate using our CG method. We show a common source gather and a common receiver gather for 75% missing sources in Figure 5. We summarize our results in Figure 6 for interpolating this volume from varying amounts of missing sources.

Missing Sources	SNR - Known	SNR - Interpolated
25%	15.4 dB	14.4 dB
50%	15.7 dB	14.1 dB
75%	17.4 dB	11.6 dB

Fig. 6. SNRs of the data volume restricted to known source locations and interpolated source locations after recovery.

VII. Conclusion

In this work, we have extended the largely theoretical results of [1] to a practical algorithmic framework for solving optimization problems whose solutions lie on a Hierarchical Tucker manifold of fixed dimension tree and hierarchical rank. Our methods easily allow us to interpolate tensors exhibiting this hierarchical low-rank structure from a subset of their entries. There is a large open question as to how one can formulate precise recovery results for this problem to the sufficiently comprehensive level of the recovery results present in the Compressive Sensing and Matrix Completion literature, a question that we leave for future research.

The authors would like to thank the sponsors of the SINBAD consortium for their continued support and particularly BG for providing the test data set.

References

[1] A. Uschmajew and B. Vandereycken, "The geometry of algorithms using hierarchical tensors," *preprint, http://sma.epfl.ch/~vanderey/papers/geom_htucker.pdf*, 2012.

[2] B. Vandereycken, "Low-rank matrix completion by riemannian optimization—extended version," *arXiv.org*, Sep. 2012.

[3] P. Absil, R. Mahony, and R. Sepulchre, *Optimization algorithms on matrix manifolds*. Princeton Univ Press, 2008.

[4] E. van den Berg and M. Friedlander. (2009) The spot operator toolbox. [Online]. Available: http://www.cs.ubc.ca/labs/scl/spot/index.html

[5] D. Kressner and C. Tobler, "htucker–a matlab toolbox for tensors in hierarchical tucker format," *MATHICSE, EPF Lausanne, available at http://sma.epfl.ch/~anchpcommon/publications/htucker.pdf*, 2012.

[6] C. TOBLER, "Low rank tensor methods for linear systems and eigenvalue problems," Ph.D. dissertation, ETH Zürich, 2012.

[7] E. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational mathematics*, vol. 9, no. 6, pp. 717–772, 2009.

Estimation of large data sets on the basis of sparse sampling

Anatoli Torokhti, Phil Howlett
CIAM, School of Inf. Techn. & Math. Sci.
University of South Australia, SA 5095
Email: anatoli.torokhti@unisa.edu.au

Hamid Laga
PBRC, School of Inf. Techn. & Math. Sci.
University of South Australia, SA 5095
Email: hamid.laga@unisa.edu.au

***Abstract*—We propose a new technique which allows us to estimate any random signal from a large set of noisy observed data on the basis of samples of only a few reference signals.**

I. Introduction

A. Motivation

In many applications associated with difficult environments, *a priori* information on signals of interest can be obtained only at *a few* given times $\{t_j\}_1^p \subset T = [a,b] \subset \mathbb{R}$ where $a = t_1 < t_2 < \cdots < t_{p-1} < t_p = b$ whereas it is required to estimate the signals at *any time* $t \in T$. Typical examples are devices deployed in the stratosphere, underground or underwater. The choice of points t_j might be beyond our control (e.g. in geophysics and defence tascs). For any $t \in T$, the signal is a stochastic vector. We consider large sets of such signals where each signal is associated with a particular $t \in T$. The observations are noisy and also *large*. Thus, all we can exploit is noisy observations and a *sparse* information on reference signals given by samples of the signal set at times $\{t_j\}_1^p$.

B. Formalization of the problem

To formalize the problem, we denote by Ω the set of all experimental outcomes[1], by $\mathcal{K}_x = \{\boldsymbol{x}_\omega \mid \omega \in \Omega\}$ a set of reference stochastic signals and by $\mathcal{K}_y = \{\boldsymbol{y}_\omega \mid \omega \in \Omega\}$ a set of observed signals[2]. Note that, theoretically, $\mathcal{K}_x$ and $\mathcal{K}_y$ are infinite signal sets. In practice, however, sets $\mathcal{K}_x$ and $\mathcal{K}_y$ are finite and large, each with, say, N signals. To each random outcome $\omega \in \Omega$ we associate a unique signal pair $(\boldsymbol{x}_\omega, \boldsymbol{y}_\omega)$ where $\boldsymbol{x}_\omega : T \to \mathcal{C}^{0,1}(T, \mathbb{R}^m)$ and $\boldsymbol{y}_\omega : T \to \mathcal{C}^{0,1}(T, \mathbb{R}^n)$[3]. Write

$$\mathcal{P} = \mathcal{K}_x \times \mathcal{K}_y = \{(\boldsymbol{x}_\omega, \boldsymbol{y}_\omega) \mid \omega \in \Omega\} \qquad (1)$$

for the set of all such signal pairs. For each $\omega \in \Omega$, the components $\boldsymbol{x}_\omega = \boldsymbol{x}_\omega(t), \boldsymbol{y}_\omega = \boldsymbol{y}_\omega(t)$ are Lipschitz continuous vector-valued functions on T [1].

We wish to construct an estimator $F^{(p-1)}$ that estimates each reference signal $\boldsymbol{x}_\omega(t)$ in $\mathcal{P}$ from related observed input $\boldsymbol{y}_\omega(t)$ under the restriction that *a priori* information on only a *few* reference signals, $\boldsymbol{x}_\omega(t_1), \ldots, \boldsymbol{x}_\omega(t_p)$, is available where $p \ll N$.

In more detail, this restriction implies the following. Let us denote by $\mathcal{K}_x^{(p)}$ a set of p signals $\boldsymbol{x}_\omega(t_1), \ldots, \boldsymbol{x}_\omega(t_p)$ for which *a priori* information is available. A set of associated observed signals $\boldsymbol{y}_\omega(t_1), \ldots, \boldsymbol{y}_\omega(t_p)$ is denoted by $\mathcal{K}_y^{(p)}$. Then for all $\boldsymbol{y}_\omega(t)$ that do not belong to $\mathcal{K}_y^{(p)}$, $\boldsymbol{y}_\omega(t) \notin \mathcal{K}_y^{(p)}$, estimator $F^{(p-1)}$ is said to be the *blind* estimator [2], [3], [4], [5] since no information on $\boldsymbol{x}_\omega(t) \notin \mathcal{K}_x^{(p)}$ is available. If $\boldsymbol{y}_\omega(t) \in \mathcal{K}_y^{(p)}$ then $F^{(p-1)}$ becomes a *nonblind* estimator since information on $\boldsymbol{x}_\omega(t) \in \mathcal{K}_x^{(p)}$ is available. Thus, depending on $\boldsymbol{y}_\omega(t)$, estimator $F^{(p-1)}$ is classified differently. Therefore, such a procedure of estimating reference signals in $\mathcal{K}_x$ is here called the *almost blind* estimation.

C. Differences from known techniques

We would like to note that the *almost blind* estimation is different from known methods such as nonblind [6]–[18], semiblind and blind techniques [2]–[5], [19]–[22][4]. Indeed, at each particular time $t \in T$, the input of the *almost blind* estimator $F^{(p-1)}$ developed below in this paper, is a random vector $\boldsymbol{y}_\omega(t)$. Thus, for different $t \in T$, the input is a different random vector $\boldsymbol{y}_\omega(t)$ but we

[1] We write $\{\Omega, \Sigma, \mu\}$ for a probability space where $\Sigma \subset \Omega$ is a sigma-algebra of measurable sets known as the event space and μ is a non-negative probability measure with $\mu(\Omega) = 1$.

[2] In an intuitive way, $\mathbf{y}$ can be regarded as a noise-corrupted version of $\mathbf{x}$. For example, $\mathbf{y}$ can be interpreted as $\mathbf{y} = \mathbf{x} + \mathbf{n}$ where $\mathbf{n}$ is white noise. We do not restrict ourselves to this simplest version of $\mathbf{y}$ and assume that the dependence of $\mathbf{y}$ on $\mathbf{x}$ and $\mathbf{n}$ is arbitrary.

[3] The space $\mathcal{C}^{0,1}(T, \mathbb{R}^p)$ is the set of vector-valued Hölder continuous functions $\boldsymbol{f}$ of order 1 with $\boldsymbol{f}(t) \in \mathbb{R}^p$ and $\|\boldsymbol{f}(s) - \boldsymbol{f}(t)\| \leq K|s-t|$. See [1], p. 96.

[4] The literature on these subjects is very abundant. Here, we listed only some related references.

wish to keep *the same estimator* $F^{(p-1)}$ for any $t \in T$, i.e. for any observed signal $\boldsymbol{y}_\omega(t)$ in the set $\mathcal{K}_y$.

By known techniques in [2]–[16] and [19]–[22], an estimator (here, we choose the united term 'estimator' to denote an equalizer or a system) is specifically designed for *each* particular input–output pair represented by random vectors. That is, for different inputs (observed signals) $\boldsymbol{y}_\omega(t)$, known techniques require different specified estimators and the number of estimators should be equal to a number of processed signals. In the case of *large signal sets*, such approaches become inconvenient because the number of signals N can be very large as it is supposed in this paper. For example, in problems related to DNA analysis, $N = \mathcal{O}(10^4)$. Therefore, the inconvenient (burdened, difficult) restriction of using *a priori* information on only p reference signals, with $p \ll N$, is quite significant. At the same time, beside difficulties that this restriction imposes on the estimation procedure, we use it in a way that allows us to avoid the hard work associated with known techniques applied to large signal sets. To the best of our knowledge, the exception is the methodology in [17], [18] where for estimation of a set of signals, the single estimator is constructed. The estimation techniques in [17], [18] exploit information in the form of a vector obtained, in particular, from averaging over signals in $\mathcal{K}_x^{(p)}$.

Further, the semiblind techniques are not applicable to the considered problem because they require a knowledge of some 'parts' of each reference signal in $\mathcal{K}_x$ (e.g., see [3], [5], [19]) but it is not the case here. Although the blind techniques allow us to avoid this restriction, it is known that they have difficulties related to accuracy and computational load. In the problem under consideration, the advantage is a knowledge of some (small) part of the set of reference signals. It is natural to use this advantage in the estimator structure and we will do it in Section II.

Nonblind estimators [6]–[16] are not applicable here because they require *a priori* information on each reference signal in $\mathcal{K}_x$ (e.g., a knowledge of covariance matrix $E[\boldsymbol{x}_\omega \boldsymbol{y}_\omega^T]$ where E is the expectation operator). In particular, it is known that there are significant advantages in adaptive or recursive estimators (e.g., associated with Kalman filtering approach) and it may well be possible to embed our estimator into such an environment—but that is not our particular concern here. Further, we note that much of the literature on piecewise linear estimators [23]–[26] seems to be *not directly relevant* to the estimator proposed here. In the first instance papers such as [23]–[26] are mostly concerned with the theoretical problems of approximation by piecewise linear functions on multi-dimensional domains which is *not the case here*.

Also, unlike many known techniques, we consider the case of observations corrupted by an arbitrary noise (not by an additive noise only) and design the estimator in terms of the Moore-Penrose pseudo-inverse matrix [27]. Therefore it is always well defined.

II. The main results

In this section we outline the rationale for the proposed estimator and state the main results.

A. Some preliminaries

The proposed estimator $F^{(p-1)}$ is adaptive to a sparse set $\mathcal{K}_x^{(p)}$.

The conceptual device behind the proposed estimator is a linear interpolation of an optimal incremental estimation applied to random signal pairs $(\boldsymbol{x}_\omega(t_j), \boldsymbol{y}_\omega(t_j))$ and $(\boldsymbol{x}_\omega(t_{j+1}), \boldsymbol{y}_\omega(t_{j+1}))$, for $j = 1, \ldots, p-1$, interpreted an extension of the least squares linear (LSL) estimator (see, for example, [6], [11], [16]).
Although this idea may seem reasonable the detailed justification of the new estimator is not straightforward and requires careful analysis. We shall do this by establishing an upper bound for the associated error and by showing that this upper bound is directly related to the expected error for an incremental application of the optimal LSL estimator. In Section II-B below, we will show that such an estimator is possible under quite unrestrictive assumptions.

Since the estimator proposed below is based on an extension of the LSL estimator it is convenient to sketch known related results here. Consider a *single* random signal pair $(\boldsymbol{x}(\omega), \boldsymbol{y}(\omega))$ where $\boldsymbol{x} \in L^2(\Omega, \mathbb{R}^m)$ and $\boldsymbol{y} \in L^2(\Omega, \mathbb{R}^n)$ with zero mean $(E[\boldsymbol{x}], E[\boldsymbol{y}]) = (\boldsymbol{0}, \boldsymbol{0})$, where $\boldsymbol{0}$ is the zero vector. Note that here, $\boldsymbol{x}$ and $\boldsymbol{y}$ do not depend on t as above. The estimate $\widehat{\boldsymbol{x}}$ of the reference vector $\boldsymbol{x}$ by the optimal least squares linear estimator is given by

$$\widehat{\boldsymbol{x}}(\omega) = E_{\boldsymbol{xy}} E_{\boldsymbol{yy}}^{\dagger} \, \boldsymbol{y}(\omega) \tag{2}$$

where $E_{\boldsymbol{xy}} = E[\boldsymbol{x}\boldsymbol{y}^T]$ and $E_{\boldsymbol{yy}} = E[\boldsymbol{y}\boldsymbol{y}^T]$ are known covariance matrices and $E_{\boldsymbol{yy}}^{\dagger}$ is the Moore-Penrose pseudo-inverse of $E_{\boldsymbol{yy}}$. By the LSL estimator, matrices $E_{\boldsymbol{xy}}$ and $E_{\boldsymbol{yy}}^{\dagger}$ should be specified for each signal pair $(\boldsymbol{x}(\omega), \boldsymbol{y}(\omega))$.

Further, for a justification of our estimator, we need some more notation as follows. It is convenient to

denote $\boldsymbol{x}(t,\omega) = \boldsymbol{x}_\omega(t)$ and $\boldsymbol{y}(t,\omega) = \boldsymbol{y}_\omega(t)$ so that $\boldsymbol{x}(t,\omega) \in \mathbb{R}^m$ and $\boldsymbol{y}(t,\omega) \in \mathbb{R}^n$.

B. The piecewise LSL interpolation estimator

For each signal pair (or vector function pair) in the set $\mathcal{P}$, $(\boldsymbol{x}(t,\omega), \boldsymbol{y}(t,\omega))$, we assume that $(E[\boldsymbol{x}(t,\cdot)], E[\boldsymbol{y}(t,\cdot)]) = (\mathbf{0}, \mathbf{0})$. To begin the estimation process we need to find an initial estimate $\widehat{\boldsymbol{x}}(t_1,\omega)$. It is assumed this can be found by some known method. Further, let us consider a signal estimation procedure at $t_2, \cdots, t_p$. We use an inductive argument to define an incremental estimation procedure. Consider a typical interval $[t_j, t_{j+1}]$ and define incremental random vectors

$$\boldsymbol{v}_j(\omega) = \boldsymbol{x}(t_{j+1},\omega) - \boldsymbol{x}(t_j,\omega) \in \mathbb{R}^m, \quad (3)$$

$$\boldsymbol{w}_j(\omega) = \boldsymbol{y}(t_{j+1},\omega) - \boldsymbol{y}(t_j,\omega) \in \mathbb{R}^n \quad (4)$$

and construct the optimal linear estimate

$$\widehat{\boldsymbol{v}}_j(\omega) = E_{\boldsymbol{v}_j \boldsymbol{w}_j} E^{\dagger}_{\boldsymbol{w}_j \boldsymbol{w}_j}\, \boldsymbol{w}_j(\omega) \quad (5)$$

of the increment $\boldsymbol{v}_j(\omega)$ for each $j = 1, \ldots, p-1$. We will write

$$B_j = E_{\boldsymbol{v}_j \boldsymbol{w}_j} E^{\dagger}_{\boldsymbol{w}_j \boldsymbol{w}_j} \in \mathbb{R}^{m \times n}. \quad (6)$$

Define the estimate at point t_{j+1} by setting $\widehat{\boldsymbol{x}}(t_{j+1},\omega) = \widehat{\boldsymbol{x}}(t_j,\omega) + \widehat{\boldsymbol{v}}_j(\omega)$. Thus we have

$$\begin{aligned} \widehat{\boldsymbol{x}}(t_{j+1},\omega) &= \widehat{\boldsymbol{x}}(t_j,\omega) + B_j[\boldsymbol{y}(t_{j+1},\omega) - \boldsymbol{y}(t_j,\omega)] \\ &= \boldsymbol{\epsilon}_j(\omega) + B_j \boldsymbol{y}(t_{j+1},\omega) \end{aligned} \quad (7)$$

where we write

$$\boldsymbol{\epsilon}_j(\omega) = \widehat{\boldsymbol{x}}(t_j,\omega) - B_j \boldsymbol{y}(t_j,\omega). \quad (8)$$

Note that this definition can be rewritten more suggestively as

$$\widehat{\boldsymbol{x}}(t_j,\omega) = \boldsymbol{\epsilon}_j(\omega) + B_j \boldsymbol{y}(t_j,\omega) \quad (9)$$

for each $j = 1, \ldots, p-1$.

The formula (7) shows that on each subinterval $[t_j, t_{j+1}]$ the estimate of the reference signal at t_{j+1} is obtained from the initial estimate at t_j by adding the optimal LSL estimate of the increment.

Now, consider a signal estimation at any $t \in [a,b]$. The first step is simply to extend the formulæ (7) and (9) to all $t \in [t_j, t_{j+1}]$ by defining

$$\widehat{\boldsymbol{x}}(t,\omega) = \boldsymbol{\epsilon}_j(\omega) + B_j \boldsymbol{y}(t,\omega). \quad (10)$$

Thus the incremental estimation across each subinterval is extended to every point within the subinterval. Because of determining estimate $\widehat{\boldsymbol{x}}(t_{j+1},\omega)$ in the form (5)–(7) we interpret this procedure as the *LSL piecewise interpolation*.

The incremental estimations are collected together in the following way. For each $j = 1, 2, \ldots, p-1$, write

$$F_j[\boldsymbol{y}(t,\omega)] = \boldsymbol{\epsilon}_j(\omega) + B_j \boldsymbol{y}(t,\omega) \quad (11)$$

for all $t \in [t_j, t_{j+1}]$ and hence define the *piecewise LSL interpolation estimator* by setting

$$F^{(p-1)}[\boldsymbol{y}(t,\omega)] = \sum_{j=1}^{p-1} F_j[\boldsymbol{y}(t,\omega)][u(t-t_j) - u(t-t_{j+1})] \quad (12)$$

for all $t \in [a,b]$ where $u(t) = \begin{cases} 1 & \text{for } t > 0 \\ 0 & \text{otherwise.} \end{cases}$ is the unit step function. Thus we can now use the estimate

$$\widehat{\boldsymbol{x}}(t,\omega) = F^{(p-1)}[\boldsymbol{y}(t,\omega)] \quad (13)$$

for all $(t,\omega) \in T \times \Omega$. The idea of a piecewise LSL interpolation estimator on T seems intuitively reasonable for signals with a well defined gradient over T.

We note that by (6)-(13), the estimator $F^{(p-1)}$ is adaptive to a variation of signals in $\mathcal{K}_x^{(p)}$. A change of signals in $\mathcal{K}_x^{(p)}$ implies a change of determinations of sub-estimators B_j in (6) and keep the same structure of the $F^{(p-1)}$.

C. Justification of the LSL interpolation estimator

We wish to justify the proposed estimator by establishing an upper bound for the associated error.

To explain the technical details we introduce some further terminology.

Let us denote $\|\boldsymbol{x}(t,\cdot)\|_\Omega^2 = \int_\Omega \|\boldsymbol{x}(t,\omega)\|^2 d\mu(\omega)$. Assume that for all $t \in T$, we have

$$\|\boldsymbol{x}(t,\cdot)\|_\Omega^2 < \infty \quad \text{and} \quad \|\boldsymbol{y}(t,\cdot)\|_\Omega^2 < \infty, \quad (14)$$

where $\|\boldsymbol{x}(t,\omega)\|$ and $\|\boldsymbol{y}(t,\omega)\|$ are the Euclidean norms for $\boldsymbol{x}(t,\omega)$ and $\boldsymbol{y}(t,\omega)$ for each $(t,\omega) \in T \times \Omega$, respectively. Thus we will say that the signals are square integrable in ω and write $\boldsymbol{x}(t,\cdot) \in L^2(\Omega)$ and $\boldsymbol{y}(t,\cdot) \in L^2(\Omega)$ for each fixed $t \in T$.

For each $t \in T$, let $\mathcal{F} = \{\boldsymbol{f} : T \times \Omega \to \mathbb{R}^m \mid \boldsymbol{f}(t,\cdot) \in L^2(\Omega, \mathbb{R}^m)\}$ and define

$$\begin{aligned} \|\boldsymbol{f}\|_{T,\Omega} &= \frac{1}{b-a} \int_{T \times \Omega} \|\boldsymbol{f}(t,\omega)\|\, dt\, d\mu(\omega) \\ &= \frac{1}{b-a} \int_T E[\|\boldsymbol{f}(t,\cdot)\|]\, dt \end{aligned}$$

for each $\boldsymbol{f} \in \mathcal{F}$ where $\|\boldsymbol{f}(t,\omega)\|$ is the Euclidean norm of $\boldsymbol{f}(t,\omega)$ on $\mathbb{R}^m$ for all $(t,\omega) \in \mathbb{R}^m$. Suppose that for

all $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{P}$ there exist constants $\gamma_j, \delta_j > 0$ such that

$$\|\boldsymbol{x}(s,\omega) - \boldsymbol{x}(t,\omega)\| \leq \gamma_j |s-t|, \quad (15)$$

$$\|\boldsymbol{y}(s,\omega) - \boldsymbol{y}(t,\omega)\| \leq \delta_j |s-t| \quad (16)$$

for all $(s,\omega), (t,\omega) \in [t_j, t_{j+1}] \times \Omega$, i.e. we suppose that the Lipschitz constants in (15) are independent of ω.

The error bound for the piecewise LSL interpolation estimator is established in Theorem 1 below.

Theorem 1: If condition (15) is satisfied then the error $\epsilon_p = \|\boldsymbol{x} - F^{(p-1)}[\boldsymbol{y}]\|_{T,\Omega}$ associated with the piecewise LSL interpolation estimator satisfies the inequality

$$\epsilon_p \leq \max_{j=1,\ldots,p-1} \{(\gamma_j + \|B_j\|_2 \delta_j)|t_{j+1} - t_j| \quad (17)$$

$$+ \left[\|E^{1/2}_{\boldsymbol{v}_j,\boldsymbol{v}_j}\|_F^2 - \|E_{\boldsymbol{v}_j\boldsymbol{w}_j}(E^{1/2}_{\boldsymbol{w}_j\boldsymbol{w}_j})^\dagger\|_F^2\right]^{1/2}\} \quad (18)$$

where $\|B_j\|_2$ denotes the 2-norm given by the square root of the largest eigenvalue of $B_j^T B_j$ and $\|\cdot\|$ denotes the Frobenius norm.

III. Conclusion

The piecewise least squares linear (LSL) interpolation estimator was developed to estimate a large set of random signals of interest from the set of observed data. The distinctive feature is that *a priori* information can be obtained on only a *few* reference signals in the form of samples. Since no information of the major part of the set of reference signals is known, such a procedure is called *almost blind* estimation.

The proposed estimator mitigates to some extent the difficulties associated with existing estimation approaches such as the necessity to know information (in the form of a sample, for instance) on *each* random reference signal; invertibility of the matrices used to define the estimators; and demanding computational work.

References

[1] E. Zeidler, *Applied Functional Analysis, Applications to Mathematical Physics*, Applied Mathematical Sciences 108, Springer, 1997.

[2] Y. Hua, Fast maximum likelihood for blind identification of multiple FIR channels, *IEEE Trans. on Signal Processing,* 44, No. 3, pp. 661-672, 1996.

[3] K. Georgoulakis and S. Theodoridis, Blind and semi-blind equalization using hidden Markov models and clustering techniques, *Sifnal Processing,* 80, Issue 9, pp. 1795-1805, 2000.

[4] V. Zarzoso and P. Comon, Blind and Semi-Blind Equalization Based on the Constant Power Criterion, *IEEE Trans. on Signal Processing,* 53, 11, pp. 4363–4375, 2005.

[5] C.–Y. Chi, C.–H. Chen, C.–C. Feng, C.–Y. Chen, *Blind Equalization and System Identification*, Springer, 2006.

[6] Y. Hua, M. Nikpour, and P. Stoica, Optimal Reduced-Rank estimation and filtering, *IEEE Trans. on Signal Processing,* vol. 49, pp. 457-469, 2001.

[7] Y. Hua and W. Q. Liu, Generalized Karhunen-Loève transform, *IEEE Signal Processing Letters*, vol. 5, pp. 141-143, 1998.

[8] S. Haykin, *Adaptive Filter Theory*, Prentice Hall, ISBN 0-13-048434-2, 2002.

[9] J. Chen, J. Benesty, Y. Huang, and S. Doclo, New Insights Into the Noise Reduction Wiener Filter, *IEEE Trans. on Audio, Speech, and Language Processing,* **14**, No. 4, 1218 - 1234, 2006.

[10] M. Spurbeck and P. Schreier, Causal Wiener filter banks for periodically correlated time series, *Signal Processing,* 87, 6, pp. 1179-1187, 2007.

[11] J. S. Goldstein, I. Reed, and L. L. Scharf, A Multistage Representation of the Wiener Filter Based on Orthogonal Projections, *IEEE Trans. on Information Theory,* vol. 44, pp. 2943-2959, 1998.

[12] V. J. Mathews and G. L. Sicuranza, *Polynomial Signal Processing*, J. Wiley & Sons, 2001.

[13] A. P. Torokhti and P. G. Howlett, An Optimal Filter of the Second Order, *IEEE Trans. on Signal Processing*, **49**, No 5, 1044-1048, 2001.

[14] A. Torokhti and P. Howlett, Method of recurrent best estimators of second degree for optimal filtering of random signals, *Signal Processing*, **83**, 5, 1013-1024, 2003.

[15] A. Torokhti and P. Howlett, Optimal Transform Formed by a Combination of Nonlinear Operators: The Case of Data Dimensionality Reduction, *IEEE Trans. on Signal Processing*, **54**, No. 4, 1431-1444, 2006.

[16] A. Torokhti and P. Howlett, *Computational Methods for Modelling of Nonlinear Systems*, Mathematics in Science and Engineering, **212**, series editor C. K. Chui, Elsevier, 2007.

[17] A. Torokhti and P. Howlett, Filtering and Compression for Infinite Sets of Stochastic Signals, *Signal Processing*, 89, pp. 291-304, 2009.

[18] A. Torokhti and J. Manton, Generic Weighted Filtering of Stochastic Signals, *IEEE Trans. on Signal Processing,* 57, issue 12, pp. 4675-4685, 2009.

[19] H. A. Cirpan and M. K. Tsatsanis, Stochastic Maximum Likelihood Methods for Semi-Blind Channel Estimation, *IEEE Signal Processing Letters,* 5, No. 1, pp. 21–24,1998.

[20] G. Kutz and D. Raphaeli, Maximum-Likelihood Semiblind Equalization of Doubly Selective Channels Using the EM Algorithm, *EURASIP Journal on Advances in Signal Processing*, Springer Open J., 2010.

[21] D. He and H. Leung, Semi-Blind Identification of ARMA Systems Using a Dynamic-Based Approach, *IEEE Trans. on Circuits and Systems-I*, 52, 1, pp. 179–190, 2005.

[22] J. Even and K. Sugimoto, An ICA approach to semi-blind identification of strictly proper systems based on interactor polynomial matrix, *Int. J. Robust Nonlinear Control*, 17, pp. 752768, 2007.

[23] S. Kang and L. Chua, A global representation of multidimensional piecewise-linear functions with linear partitions, *IEEE Trans. on Circuits and Systems,* 25 Issue:11, pp. 938 - 940, 1978.

[24] L.O. Chua and A.-C. Deng, Canonical piecewise-linear representation, *IEEE Trans. on Circuits and Systems,* **35** Issue:1, pp. 101 - 111, 1988.

[25] P. Julian, A. Desages, B. D'Amico, Orthonormal high-level canonical PWL functions with applications to model reduction, *IEEE Trans. on Circuits and Systems I: Fundamental Theory and Applications,* 47 Issue:5, pp. 702 - 712, 2000.

[26] J.E. Cousseau, J.L. Figueroa, S. Werner, T.I. Laakso, Efficient Nonlinear Wiener Model Identification Using a Complex-Valued Simplicial Canonical Piecewise Linear Filter, *IEEE Trans. on Signal Processing*, 55 5, pp. 1780 - 1792, 2007.

[27] A. Ben-Israel and T. N. E. Greville, *Generalized Inverses: Theory and Applications,* John Wiley & Sons, New York, 1974.

[28] P. G. Howlett, P. Pudney, and X. Vu, Local energy minimization in optimal train control, *Automatica*, **45**(11), 2692–2698, 2009. DOI: 10.1016/j.automatica.2009.07.028.

ANALYSIS OF HIERARCHICAL IMAGE ALIGNMENT WITH DESCENT METHODS

Elif Vural and Pascal Frossard

Ecole Polytechnique Fédérale de Lausanne (EPFL)
Signal Processing Laboratory (LTS4)
Switzerland-1015 Lausanne

ABSTRACT

We present a performance analysis for image registration with gradient descent methods. We consider a multiscale registration setting where the global 2-D translation between a pair of images is estimated by smoothing the images and minimizing the distance between their intensity functions with gradient descent. We focus in particular on the effect of low-pass filtering on the alignment performance. We adopt an analytic representation for images and analyze the well-behavedness of the distance function by estimating the neighborhood of translations for which the distance function is free of undesired local minima. This corresponds to the set of translation vectors that are correctly computable with a simple gradient descent minimization. We show that the area of this neighborhood increases at least quadratically with the filter size, which justifies the use of smoothing in image registration with local optimizers. We finally use our results in the design of a regular multiscale grid in the translation parameter domain that has perfect alignment guarantees.

***Keywords*—** Image registration, image smoothing, gradient-descent, performance analysis.

1. INTRODUCTION

The estimation of the transformation that best aligns two images is one of the important problems of image processing. The necessity for registering images arises in many different applications; e.g., image analysis and classification [1], [2], stereo vision [3], motion estimation for video coding [4]. Many registration techniques adopt, or can be coupled with, a multiscale hierarchical search strategy. In hierarchical registration, reference and target images are aligned by applying a coarse-to-fine estimation of the transformation parameters, using a pyramid of low-pass filtered and downsampled versions of the images.

In this work, we analyze the effect of smoothing on the performance of registration. It is commonly admitted that smoothing an image pair is helpful for overcoming the undesired local minima of the distance function between images. In practice, filtering is commonly used in hierarchical registration and motion estimation methods [4]. However, to the best of our knowledge, the analytical relation between filtering and the well-behavedness of the image dissimilarity function has not been extensively studied. Most theoretical results in the image registration literature investigate how image noise affects the registration accuracy, e.g., [5], [6]. However, the analysis of the effect of smoothing on the registration performance has generally been given less attention in the literature. Some of the existing works examine how smoothing influences the bias on the registration with gradient-based methods [5], [7]. Also, there are some results in scale-space theory that examine the variation of the local minima of 1-D and 2-D functions with filtering [8], which however does not exactly have the same setting as in the image registration problem.

In this paper, we consider a setting where the geometric transformation between the reference and target patterns is a global 2-D translation. In particular, we examine the neighborhood of translation vectors in which the only local minimum of the distance function is also the global minimum in the alignment problem. This neighborhood defines the translations between a pair of images, which can be estimated correctly with a descent algorithm. We call this neighborhood the Single Distance Extremum Neighborhood (SIDEN) of the reference pattern. For the ease of derivations, we formulate the registration problem in the continuous domain of square-integrable functions $L^2(\mathbb{R}^2)$ and adopt an analytic and parametric model for the reference and target patterns. We derive an analytic estimation of the SIDEN in terms of the pattern parameters. Then, in order to study the effect of smoothing on the registration performance, we consider the alignment of low-pass filtered versions of the reference and target patterns and examine how the SIDEN varies with the filter size. Our main result is that the volume (area) of the SIDEN increases at a rate of at least $O(1+\rho^2)$ with respect to the filter size ρ. This formally shows that, when the patterns are low-pass filtered, a wider range of translation values can be recovered with descent-type methods; hence, smoothing improves the well-behavedness of the distance function. Finally, we demonstrate the usage of our SIDEN estimate in sampling the translation parameter domain to construct a grid such that any translation between the image pair can be exactly recovered by locating the closest solution on the grid and then locally refining this estimation with a descent method. This can be achieved by adjusting the grid units with respect to the SIDEN of the pattern.

2. IMAGE REGISTRATION ANALYSIS

2.1. Notation and Problem Formulation

Let $p \in L^2(\mathbb{R}^2)$ be a visual pattern. In order to study the image registration problem analytically, we adopt a representation of p in an analytic and parametric dictionary manifold

$$\mathcal{D} = \{\phi_\gamma : \gamma = (\psi, \tau_x, \tau_y, \sigma_x, \sigma_y) \in \Gamma\} \subset L^2(\mathbb{R}^2). \qquad (1)$$

Here, each atom ϕ_γ of the dictionary $\mathcal{D}$ is derived from an analytic mother function ϕ by a geometric transformation specified by the parameter vector γ, where ψ is a rotation parameter, τ_x and τ_y denote translations in x and y directions, and σ_x and σ_y represent an anisotropic scaling in x and y directions. Γ is the transformation parameter domain over which the dictionary is defined. Defining the spatial coordinate variable $X = [x\ y]^T \in \mathbb{R}^{2\times 1}$, we will refer to the mother function as $\phi(X)$. Then an atom ϕ_γ is given by $\phi_\gamma(X) = \phi(\sigma^{-1}\,\Psi^{-1}\,(X-\tau))$, where

$$\sigma = \begin{bmatrix} \sigma_x & 0 \\ 0 & \sigma_y \end{bmatrix},\ \Psi = \begin{bmatrix} \cos(\psi) & -\sin(\psi) \\ \sin(\psi) & \cos(\psi) \end{bmatrix},\ \tau = \begin{bmatrix} \tau_x \\ \tau_y \end{bmatrix}. \qquad (2)$$

This work has been partly funded by the Swiss National Science Foundation under Grant 200020_132772.

It is shown in [9] (in the proof of Proposition 2.1.2) that the linear span of a dictionary $\mathcal{D}$ generated with respect to the transformation model in (1) is dense in $L^2(\mathbb{R}^2)$ if the mother function ϕ has nontrivial support (unless $\phi(X) = 0$ almost everywhere). In our analysis, we choose ϕ to be the Gaussian function $\phi(X) = e^{-X^T X} = e^{-(x^2+y^2)}$ as it has good time-localization and it is easy to treat in derivations due to its well-studied properties. This choice also ensures that $Span(\mathcal{D})$ is dense in $L^2(\mathbb{R}^2)$; therefore, any pattern $p \in L^2(\mathbb{R}^2)$ can be approximated in $\mathcal{D}$ with arbitrary accuracy. We assume that a sufficiently accurate approximation of p with finitely many atoms in $\mathcal{D}$ is available

$$p(X) \approx \sum_{k=1}^{K} \lambda_k \, \phi_{\gamma_k}(X) \tag{3}$$

where K is the number of atoms used in the representation of p, γ_k are the atom parameters and λ_k are the atom coefficients.

Throughout the discussion, $T = [T_x \; T_y]^T \in S^1$ denotes a unit-norm vector and S^1 is the unit circle in $\mathbb{R}^2$. We use the notation tT for translation vectors, where $t \geq 0$ denotes the magnitude of the vector (amount of translation) and T defines the direction of translation. We consider the squared-distance between the reference pattern $p(X)$ and its translated version $p(X - tT)$, which is the continuous domain equivalent of the SSD measure that is widely used in registration methods. The squared-distance in the continuous domain is given by

$$f(tT) = \|p(X) - p(X - tT)\|^2 = \int_{\mathbb{R}^2} (p(X) - p(X - tT))^2 dX \tag{4}$$

where the notation[1] $\|.\|$ stands for the L^2-norm for vectors in $L^2(\mathbb{R}^2)$ and the ℓ^2-norm for vectors in $\mathbb{R}^2$.

The global minimum of f is at the origin $tT = 0$. Therefore, there exists an open neighborhood of 0 within which the restriction of f to a ray tT_a starting out from the origin along an arbitrary direction T_a is an increasing function of $t > 0$ for all T_a. This allows us to define the Single Distance Extremum Neighborhood (SIDEN) as follows.

Definition 1. *We call the set of translation vectors*

$$\mathcal{S} = \{0\} \cup \{\omega_T T : T \in S^1, \; \omega_T > 0, \; \textit{and} \; \frac{df(tT)}{dt} > 0 \; \textit{for all} \; 0 < t \leq \omega_T\} \tag{5}$$

the Single Distance Extremum Neighborhood (SIDEN) of p.

Note that the origin $\{0\}$ is included separately in the definition of SIDEN since the gradient of f vanishes at the origin and therefore $df(tT)/dt|_{t=0} = 0$ for all T. The SIDEN $\mathcal{S} \subset \mathbb{R}^2$ is an open neighborhood of the origin such that the only stationary point of f inside $\mathcal{S}$ is the origin. Therefore, when a translated version $p(X - tT)$ of the reference pattern is aligned with $p(X)$ with a local optimization method like a gradient descent algorithm, the local minimum achieved in $\mathcal{S}$ is necessarily also the global minimum.

Given a reference pattern p, we would like now to find an analytical estimation of $\mathcal{S}$. However, the exact derivation of $\mathcal{S}$ requires the calculation of the exact zero-crossings of $df(tT)/dt$, which is not easy to do analytically. Instead, one can characterize the SIDEN by computing a neighborhood $\mathcal{Q}$ of 0 that lies completely in $\mathcal{S}$; i.e., $\mathcal{Q} \subset \mathcal{S}$. $\mathcal{Q}$ can be derived by using a polynomial approximation of f and calculating, for all unit directions T, a lower bound δ_T for the supremum of ω_T such that $\omega_T T$ is in $\mathcal{S}$. This not only provides an analytic estimation of the SIDEN, but also defines a set that is known to be completely inside the SIDEN. The regions $\mathcal{S}$ and $\mathcal{Q}$ are illustrated in Figure 1.

[1] Since it is clear from the context which one of these norms is meant, we denote these two norms in the same way for simplicity of notation.

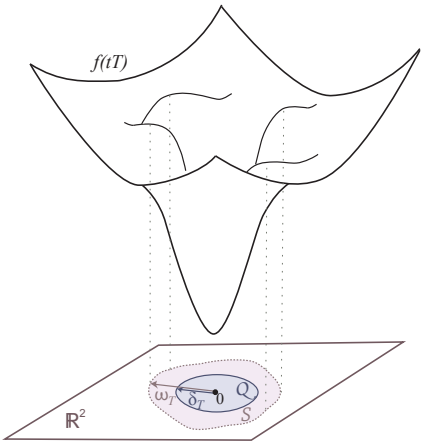


Fig. 1. SIDEN $\mathcal{S}$ is the largest open neighborhood around the origin within which the distance f is increasing along all rays starting out from the origin. Along each unit direction T, $\mathcal{S}$ covers points $\omega_T T$ such that $f(tT)$ is increasing between 0 and $\omega_T T$. The estimate $\mathcal{Q}$ of $\mathcal{S}$ is obtained by computing a lower bound δ_T for the first zero-crossing of $df(tT)/dt$.

2.2. Estimation of SIDEN

We now derive $\mathcal{Q}$ in an analytic and parametric form. In the following, we consider T to be a fixed unit direction in S^1. We derive $\mathcal{Q} \subset \mathcal{S}$ by computing a δ_T, which guarantees that $df(tT)/dt > 0$ for all $0 < t \leq \delta_T$. In the derivation of $\mathcal{Q}$, we need a closed-form expression for $df(tT)/dt$. Since f is the distance between two patterns represented in terms of Gaussian atoms, it involves the integration of the product of pairs of Gaussian atoms. These integrations yield the following terms, which are explained in more detail in [10]

$$\Sigma_{jk} := \frac{1}{2}\left(\Psi_j \, \sigma_j^2 \, \Psi_j^{-1} + \Psi_k \, \sigma_k^2 \, \Psi_k^{-1}\right)$$

$$a_{jk} := \frac{1}{2} \, T^T \, \Sigma_{jk}^{-1} \, T, \qquad b_{jk} := \frac{1}{2} \, T^T \, \Sigma_{jk}^{-1} \, (\tau_k - \tau_j)$$

$$c_{jk} := \frac{1}{2} \, (\tau_k - \tau_j)^T \, \Sigma_{jk}^{-1} \, (\tau_k - \tau_j), \qquad Q_{jk} := \frac{\pi \, |\sigma_j \sigma_k| e^{-c_{jk}}}{\sqrt{|\Sigma_{jk}|}}.$$

Notice that $a_{jk} > 0$ and $c_{jk} \geq 0$ since $\|T\| = 1$ and Σ_{jk}, Σ_{jk}^{-1} are positive definite matrices. By definition, $Q_{jk} > 0$ as well. We are now ready to state our result about the estimation of the SIDEN.

Theorem 1. *The region $\mathcal{Q} \subset \mathbb{R}^2$ is a subset of the SIDEN $\mathcal{S}$ of the pattern p if $\mathcal{Q} = \{tT : T \in S^1, \; 0 \leq t \leq \delta_T\}$, where δ_T is the only positive root of the polynomial $|\alpha_4| t^3 - \alpha_3 t^2 - \alpha_1$ and*

$$\alpha_1 = \sum_{j=1}^{K} \sum_{k=1}^{K} \lambda_j \lambda_k \, Q_{jk} \, (2 \, a_{jk} - 4 \, b_{jk}^2)$$

$$\alpha_3 = \sum_{j=1}^{K} \sum_{k=1}^{K} \lambda_j \lambda_k \, Q_{jk} \left(-\frac{8}{3} \, b_{jk}^4 + 8 \, b_{jk}^2 \, a_{jk} - 2 \, a_{jk}^2\right)$$

$$\alpha_4 = -1.37 \sum_{j=1}^{K} \sum_{k=1}^{K} |\lambda_j \lambda_k| \, Q_{jk} \, \exp\left(\frac{b_{jk}^2}{a_{jk}}\right) a_{jk}^{5/2}$$

are constants depending on T and on the parameters γ_k of the atoms of p.

The proof of Theorem 1 is given in Appendix A.1 of [10], which is an accompanying technical report. The proof applies a Taylor expansion of $df(tT)/dt$, and derives a δ_T such that $df(tT)/dt$ is positive for

all $t \leq \delta_T$. Therefore, along each direction T, δ_T constitutes a lower bound for the first zero-crossing of $df(tT)/dt$. Varying T over the unit circle, one obtains a closed neighborhood $\mathcal{Q}$ of 0 that is a subset of $\mathcal{S}$. This analytic estimate provides a guarantee for the range of translations tT over which $p(X)$ can be exactly aligned with $p(X - tT)$.

2.3. Variation of SIDEN with Smoothing

In this section, we examine how smoothing the reference pattern p with a low-pass filter influences its SIDEN. We assume a Gaussian kernel for the filter. As the reference pattern is sparsely represented in a parametric form in a Gaussian dictionary, its convolution with a Gaussian filtering function is also sparsely representable in the same dictionary. Therefore, the choice of the Gaussian kernel provides an immediate interpretation of our SIDEN estimation results for smoothed versions of the reference pattern. We assume that p is filtered with a Gaussian kernel of the form $\frac{1}{\pi\rho^2}\phi_\rho(X)$ with unit L^1-norm. The function $\phi_\rho(X) = \phi(\Lambda^{-1}(X))$ is an isotropic Gaussian atom with the diagonal scale matrix Λ having ρ on the diagonal entries. The scale parameter ρ controls the size of the Gaussian kernel. The smoothed version of the reference pattern $p(X)$ is given by

$$\hat{p}(X) = \frac{1}{\pi\rho^2}\,\phi_\rho(X) * p(X) = \sum_{k=1}^{K} \lambda_k \frac{1}{\pi\rho^2}\,\phi_\rho(X) * \phi_{\gamma_k}(X) \qquad (6)$$

by linearity of the convolution operator. As shown in [10], the filtered pattern is obtained as $\hat{p}(X) = \sum_{k=1}^{K} \hat{\lambda}_k \phi_{\hat{\gamma}_k}(X)$, where the smoothed atom $\phi_{\hat{\gamma}_k}(X)$ has parameters

$$\hat{\tau}_k = \tau_k, \quad \hat{\Psi}_k = \Psi_k, \quad \hat{\sigma}_k = \sqrt{\Lambda^2 + \sigma_k^2}, \quad \hat{\lambda}_k = \frac{|\sigma_k|}{|\hat{\sigma}_k|}\lambda_k. \qquad (7)$$

Therefore, the change in the pattern parameters due to the filtering can be captured by substituting the scale parameters σ_k of atoms with $\hat{\sigma}_k$ and replacing the coefficients λ_k with $\hat{\lambda}_k$. Now, considering the same setting as in Section 2.1, where the target pattern $p(X - tT)$ is exactly a translated version of the reference pattern $p(X)$, we examine how the volume of the SIDEN changes when the reference and target patterns are low-pass filtered as it is typically done in multiscale image registration algorithms. Hence, we analyze the variation of the smoothed SIDEN estimate $\hat{\mathcal{Q}}$ corresponding to the distance $\hat{f}(tT)$ between $\hat{p}(X)$ and $\hat{p}(X - tT)$ with respect to the filter size ρ. Since the smoothed pattern has the same parametric form as the original pattern, the variation of $\hat{\mathcal{Q}}$ with ρ can be analyzed easily by examining how the parameters involved in the derivation of the SIDEN, e.g., $\hat{a}_{jk}$, $\hat{b}_{jk}$, $\hat{\lambda}_k$, $\hat{\sigma}_k$, depend on ρ. We use the notation $\hat{(.)}$ for referring to the parameters corresponding to the filtered versions of the Gaussian atoms. We now give our main result, which summarizes the dependence of the smoothed SIDEN estimate on the filter size ρ.

Theorem 2. *Let $V(\hat{\mathcal{Q}})$ denote the volume (area) of the SIDEN estimate $\hat{\mathcal{Q}}$ for the smoothed pattern $\hat{p}$. Then, the order of dependence of the volume of $\hat{\mathcal{Q}}$ on ρ is given by $V(\hat{\mathcal{Q}}) = O(1 + \rho^2)$.*

Theorem 2 is proved in [10, Appendix A.2]. The proof is based on the examination of the order of variation of $\hat{a}_{jk}$, $\hat{b}_{jk}$, $\hat{c}_{jk}$, $\hat{Q}_{jk}$ with ρ, which is then used to derive the dependence of $\hat{\delta}_T$ on ρ. The theorem shows that the neighborhood of translation vectors inside which the reference pattern $\hat{p}(X)$ can be perfectly aligned with $\hat{p}(X - tT)$ using a descent method expands at the rate $O(1 + \rho^2)$ with respect to the increase in the filter size ρ. Here, the order of variation $O(1 + \rho^2)$ is obtained for the estimate $\hat{\mathcal{Q}}$ of the SIDEN. Since $\hat{\mathcal{Q}} \subset \hat{\mathcal{S}}$ for all ρ, one immediate observation is that the rate of expansion of the SIDEN $\hat{\mathcal{S}}$ must be at least of $O(1+\rho^2)$; i.e., $V(\hat{\mathcal{S}}) \geq V(\hat{\mathcal{Q}}) = O(1+\rho^2)$. Note that the dependence of $V(\hat{\mathcal{S}})$ on ρ may get above this rate for particular reference patterns. For instance, for patterns that consist only of atoms with coefficients of the same sign, there exists a threshold value ρ_0 of the filter size such that for all $\rho > \rho_0$, $\hat{\mathcal{S}} = \mathbb{R}^2$ and thus $V(\hat{\mathcal{S}}) = \infty$ [10, Proposition 4].

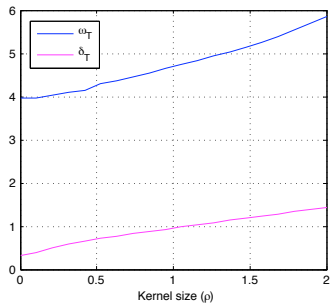


Fig. 2. The variations of the true distance $\hat{\omega}_T$ of the boundary of $\hat{\mathcal{S}}$ to the origin and its estimation $\hat{\delta}_T$ with respect to the filter size

2.4. Evaluation of SIDEN by experiments

We now evaluate our theoretical results about SIDEN estimation with an experiment that compares the estimated SIDEN to the true SIDEN. We generate a reference pattern consisting of 40 randomly selected Gaussian atoms with random coefficients, and choose a random unit direction T. Then, we determine the distance[2] $\hat{\omega}_T$ of the true SIDEN boundary from the origin along T, and compare it to its estimation $\hat{\delta}_T$ for a range of filter sizes ρ. The distance $\hat{\omega}_T$ is computed by searching the first zero-crossing of $d\hat{f}(tT)/dt$ numerically, while its estimate $\hat{\delta}_T$ is computed according to Theorem 1. We repeat the experiment 300 times with different random reference patterns p and directions T; and average the results. In 44% of the trials, $d\hat{f}(tT)/dt$ has been experimentally seen to have no zero-crossings when the pattern is filtered sufficiently. The distance $\hat{\omega}_T$ and its estimation $\hat{\delta}_T$ are plotted in Figure 2 for the remaining 56% of the patterns. The figure shows that both $\hat{\omega}_T$ and $\hat{\delta}_T$ have an approximately linear dependence on ρ. This is an expected behavior, since $\hat{\delta}_T = O\left((1+\rho^2)^{1/2}\right) \approx O(\rho)$ for large ρ. The estimate $\hat{\delta}_T$ is smaller than $\hat{\omega}_T$ since it is a lower bound for $\hat{\omega}_T$. Its variation with ρ is seen to capture well the variation of the true SIDEN boundary $\hat{\omega}_T$.

3. APPLICATION TO PARAMETER DOMAIN SAMPLING

We now demonstrate the usage of our SIDEN estimate in the construction of a grid in the translation parameter domain that is used for image registration. We have shown that small translations, i.e., vectors in $\mathcal{Q}$, can be perfectly recovered by minimizing the distance function with descent methods. However, the perfect alignment guarantee is lost for relatively large translations that are outside $\mathcal{Q}$. Hence, we propose to construct a grid in the translation parameter domain and estimate large translation vectors with the help of the grid. In particular, we describe a grid design procedure such that any translation vector tT lies inside the SIDEN of at least one grid point. Such a grid guarantees the recovery of the translation parameters if the distance function is minimized with

[2]With an abuse of notation, the parameter denoted as $\hat{\omega}_T$ in Section 2.4 corresponds in fact to $\sup \hat{\omega}_T$ in the definition of SIDEN in (5).

a gradient descent method that is initialized with the grid points. In order to have a perfect recovery guarantee, each one of the grid points must be tested. However, as this is computationally costly, we propose to use the following two-stage optimization instead, which offers a good compromise with respect to the accuracy-complexity tradeoff. First, we search for the grid vector that gives the smallest distance between the image pair, which results in a coarse alignment. Then, we refine the alignment with a gradient descent method initialized with this grid vector.

We now explain the grid construction. From Theorem 1, one can verify that the estimation δ_T of the SIDEN boundary along the direction T is symmetric and it satisfies $\delta_T = \delta_{-T}$. Therefore, one can easily determine a grid unit in the form of a parallelogram that lies completely inside $\mathcal{Q}$ and tile the (tT_x, tT_y)-plane with these grid units. This defines a regular grid in the (tT_x, tT_y)-plane such that each point of the plane lies inside the SIDEN of at least one grid point. As the SIDEN increases with the filter size, the area of the grid units expand at the rate $O(1+\rho^2)$ and the number of grid points decrease at the rate $O\left((1+\rho^2)^{-1}\right)$ with ρ.

The construction of a regular grid in this manner is demonstrated for a digit pattern. In Figure 3(a), the reference pattern and its translated versions corresponding to the neighboring grid points in the first and second directions of sampling are shown. In Figure 3(b), the reference pattern is shown when smoothed with a filter of size $\rho = 0.15$, as well as the neighboring patterns in the smoothed grid. The corresponding grids are displayed in Figures 3(c) and 3(d), where the SIDEN estimates $\mathcal{Q}$, $\hat{\mathcal{Q}}$ and the grid units are also plotted. One can observe that smoothing the pattern results in a coarser grid. In Figure 4, we plot the variation of the number of grid points with the filter size for the random patterns of the previous experiment and the digit pattern. The results confirm that the number of grid points decreases monotonically with the filter size, as stated by Theorem 2, which suggests that the number of grid points must be of $O\left((1+\rho^2)^{-1}\right)$. Finally, the experiments in [10] show that this registration method indeed has an optimal alignment performance.

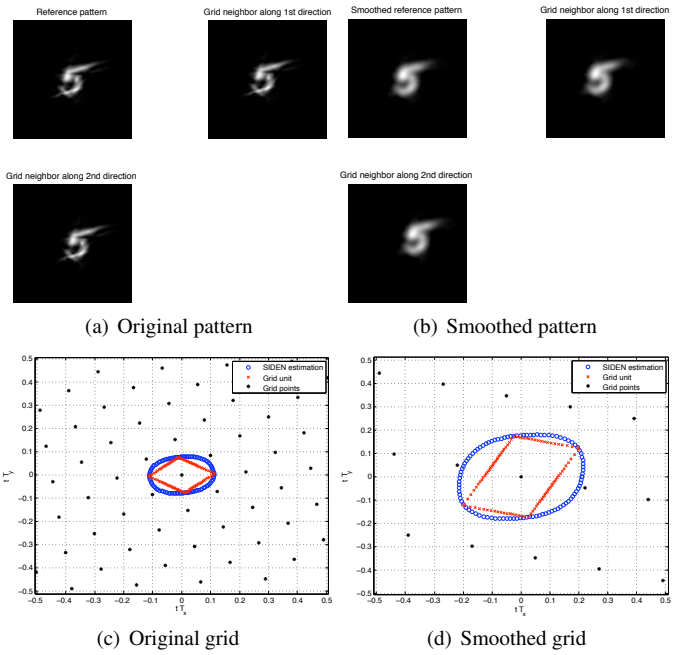
4. CONCLUSION

We have presented an analysis of hierarchical image registration with descent-type local minimizers. We have examined the problem of aligning a reference and a target pattern that differ by a two-dimensional translation. We have derived an estimation of the neighborhood of translations for which the image pair can be exactly aligned with a local optimizer. Then we have investigated how the area of this neighborhood varies with the size of the filter used in the coarse-to-fine registration process. Our finding is that the area of this neighborhood increases quadratically with the filter size, therefore, smoothing the patterns improves the well-behavedness of the distance function. We have used our results in the construction of a multiscale regular grid in the translation parameter domain that guarantees the exact alignment of a reference pattern with its translated versions. The fact that the number of grid points is inversely proportional to the square of the filter size shows that filtering is useful for decreasing the computational complexity of image alignment.

5. REFERENCES

[1] P. Simard, Y. LeCun, J. S. Denker, and B. Victorri, "Transformation invariance in pattern recognition-tangent distance and tangent propagation," in *Neural Networks: Tricks of the Trade*. 1998, New York: Springer-Verlag.

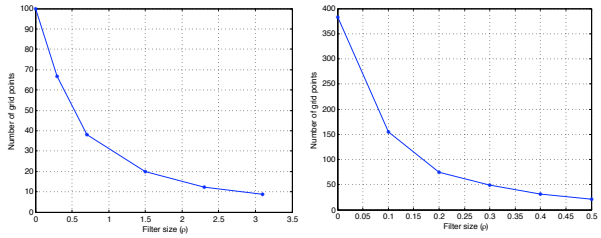
[2] A. W. Fitzgibbon and A. Zisserman, "Joint manifold distance: a new approach to appearance based clustering," *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 26, 2003.



(a) Original pattern (b) Smoothed pattern

(c) Original grid (d) Smoothed grid

Fig. 3. Construction of a regular grid in parameter domain with an exact alignment guarantee



(a) Random patterns (b) Digit pattern

Fig. 4. Number of grid points. The decay rate is of $O\left((1+\rho^2)^{-1}\right)$.

[3] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. 7th Intl. Joint Conf. on Artificial Intelligence*, 1981, pp. 674–679.

[4] G. Tziritas and C. Labit, *Motion Analysis for Image Sequence Coding*, Elsevier Science Inc., New York, NY, USA, 1994.

[5] D. Robinson and P. Milanfar, "Fundamental performance limits in image registration," *IEEE Trans. Img. Proc.*, vol. 13, no. 9, pp. 1185–1199, Sept. 2004.

[6] İ. Ş. Yetik and A. Nehorai, "Performance bounds on image registration," *IEEE Trans. Signal Proc.*, vol. 54, no. 5, pp. 1737 – 1749, May 2006.

[7] J. K. Kearney, W. B. Thompson, and D. L. Boley, "Optical flow estimation: An error analysis of gradient-based methods with local optimization," *IEEE Trans. Pattern Anal. Machine Intel.*, Mar. 1987.

[8] T. Lindeberg, *Scale-Space Theory in Computer Vision*, Kluwer Academic Publishers, 1994.

[9] J. Antoine, R. Murenzi, P. Vandergheynst, and S. Ali, *Two-Dimensional Wavelets and their Relatives*, Signal Processing. Cambridge University Press, 2004.

[10] E. Vural and P. Frossard, "Analysis of descent-based image registration," Available at: http://infoscience.epfl.ch/record/183845.

Spectrum Reconstruction from Sub-Nyquist Sampling of Stationary Wideband Signals

Deborah Cohen
Technion
Haifa, Israel
Email: debby@tx.technion.ac.il

Yonina C. Eldar
Technion
Haifa, Israel
Email: yonina@ee.technion.ac.il

***Abstract*—In light of the ever-increasing demand for new spectral bands and the underutilization of those already allocated, the new concept of Cognitive Radio (CR) has emerged. Opportunistic users could exploit temporarily vacant bands after detecting the absence of activity of their owners. One of the most crucial tasks in the CR cycle is therefore spectrum sensing and detection which has to be precise and efficient. Yet, CRs typically deal with wideband signals whose Nyquist rates are very high. In this paper, we propose to reconstruct the spectrum of such signals from sub-Nyquist samples in order to perform detection. We consider both sparse and non sparse signals as well as blind and non blind detection in the sparse case. For each one of those scenarii, we derive the minimal sampling rate allowing perfect reconstruction of the signal spectrum in a noise-free environment and provide recovery techniques. The simulations show spectrum recovery at the minimal rate in noise-free settings.**

I. Introduction

Spectral resources are traditionally allocated to primary users (PUs). As most are already licensed, new applications can hardly ever obtain access to free frequency bands. Paradoxally, the over-crowded spectrum is usually significantly underutilized as numerous studies have shown [1]–[3]. In order to respond to the increasing demand for spectrum usage from new users, the concept of Cognitive Radio (CR) [4], [5] has recently been considered. In this approach, secondary users opportunistically use temporarely vacant spectrum bands when their owners are inactive.

In this scheme, the CR has to constantly monitor the spectrum and detect the PUs' activity in order to select unoccupied bands, before and throughout its transmission. Obviously, the detection has to be extremely reliable and fast. On the other hand, it is worthwhile for the CR to sense a wide band of spectrum simultaneously, in order to increase the probability of finding a vacant spectral band. Nyquist rates of such wideband signals are very high and sometimes cannot even be met by today's best analog-to-digital converters (ADCs). Moreover, the tremendous amount of samples such high rates generate have to be processed by the CR, slowing down the digital detection process.

To overcome the rate bottleneck, several new sampling methods have recently been proposed [6]–[8] that reduce the sampling rate in multiband settings below the Nyquist rate. In [6]–[8], the authors derive the minimal sampling rate allowing for perfect signal reconstruction in noise-free settings and provide sampling and recovery techniques. However, when the final goal is spectrum sensing and detection, reconstructing the original signal is unnecessary. In this paper, we propose to only reconstruct the signal spectrum from sub-Nyquist samples, in order to perform signal detection. In [9], the authors propose a method to estimate finite resolution approximations to the true spectrum exploiting multicoset sampling. Spectrum reconstruction is also considered in [10] both in the time and frequency domains. However, no analysis on the minimal sampling rate ensuring perfect reconstruction of the spectrum was performed.

We consider the class of wide-sense stationary multiband signals, whose frequency support lies within several continuous intervals (bands). We will consider three different scenarii: (1) the signal is not assumed to be sparse, (2) the signal is assumed to be sparse and the carrier frequencies of the narrowband transmissions are known, (3) the signal is sparse but we do not assume carrier knowledge. We consider the sampling methods proposed in [6]–[8] and use a similar recovery technique to those derived in [9], [10] in order to reconstruct the signal spectrum from the sub-Nyquist samples. Our main contribution is deriving the minimal sampling rate allowing for perfect reconstruction of the spectrum in a noise-free environment, for each one of the above three cases. We show that the rate required for spectrum reconstruction is half the rate that allows for perfect signal reconstruction, for each one of the scenarii, namely the Nyquist rate, the Landau rate [11] and twice the Landau rate [7].

This paper is organized as follows. In Section II, we present the stationary multiband model and formulate the problem. Section III describes the sub-Nyquist sampling stage and the spectrum reconstruction. In Section IV, we derive the minimal sampling rate for each one of the three scenarii described above. Numerical experiments are presented in Section V.

II. System Model and Goal

A. System Model

Let $x(t)$ be a real-valued continuous-time signal, supported on $\mathcal{F} = [-T_{\text{Nyq}}/2, +T_{\text{Nyq}}/2]$. Formally, the Fourier transform of $x(t)$ defined by

$$X(f) = \int_{-\infty}^{\infty} x(t) e^{-j2\pi ft} \mathrm{d}t \tag{1}$$

is zero for every $f \notin \mathcal{F}$. We denote by $f_{\text{Nyq}} = 1/T_{\text{Nyq}}$ the Nyquist rate of $x(t)$. We assume that $x(t)$ is composed of up to N_{sig} uncorrelated stationary transmissions with disjoint frequency supports. The bandwidth of each signal does not exceed $2B$ (where we consider both positive and negative frequency bands together) [6]. We consider three different scenarii.

1) No sparsity assumption: In the first scenario, we assume no *a priori* knowledge on the signal and we do not suppose that $x(t)$ is sparse, namely $2N_{sig}B \leq f_{Nyq}$.

2) Sparsity assumption and non blind detection: Here, we assume that $x(t)$ is sparse, namely $2N_{sig}B \ll f_{Nyq}$. Moreover, the support of the potentially active transmissions is known and correspond to the frequency support of licensed users defined by the communication standard. However, since the PUs' activity can vary over time, we wish to develop a detection algorithm that is independent of a specific known signal support.

3) Sparsity assumption and blind detection: In the last scenario, we assume that $x(t)$ is sparse but we do not assume any *a priori* knowledge on the carrier frequencies.

B. Problem Formulation

In each one of the scenarii defined in the previous section, our goal is to assess which of the N_{sig} transmissions are active from sub-Nyquist samples of $x(t)$. For each signal, we define the hypothesis $\mathcal{H}_{i,0}$ and $\mathcal{H}_{i,1}$, namely the ith transmission is absent and active, respectively.

In order to assess which of the N_{sig} transmissions are active, we will first reconstruct the spectrum of $x(t)$. In our first and third scenarii, we fully reconstruct the spectrum. In the second one, we exploit our prior knowledge and reconstruct it only at the potentially occupied locations. We can then perform detection on the fully or partially reconstructed spectrum. Note that, to do so, we do not sample $x(t)$ at its Nyquist rate, nor compute its Nyquist rate samples. For each one of the scenarii, we derive the minimal sampling rate enabling perfect spectrum reconstruction in a noise-free environment.

III. Sub-Nyquist Sampling and Spectrum Reconstruction

We consider two different sampling schemes: multicoset sampling [7] and the modulated wideband converter (MWC) [6] which were previously proposed for sparse multiband signals. We show that the reconstruction stage is identical for both schemes. In this section, we reconstruct the whole spectrum. In Section IV-B, we show how we can reconstruct the spectrum only at potentially occupied locations when we have *a priori* knowledge on the carrier frequencies.

1) Multicoset sampling: Multicoset sampling [12] can be described as the selection of certain samples from the uniform grid. More precisely, the uniform grid is divided into blocks of N consecutive samples, from which only M are kept. The ith sampling sequence is defined as

$$x_{c_i}[n] = \begin{cases} x(nT_{\text{Nyq}}), & n = mN + c_i, m \in \mathbb{Z} \\ 0, & \text{otherwise}, \end{cases} \tag{2}$$

where $0 < c_1 < c_2 < \cdots < c_M < N-1$. Let $f_s = \frac{1}{NT_{\text{Nyq}}} \geq B$ be the sampling rate of each channel and $\mathcal{F}_s = [-f_s/2, f_s/2]$. Following the derivations from multicoset sampling [7], we obtain

$$\mathbf{z}(f) = \mathbf{A}\mathbf{x}(f), \qquad f \in \mathcal{F}_s, \tag{3}$$

where $\mathbf{z}_i(f) = X_{c_i}(e^{j2\pi f T_{\text{Nyq}}}), 0 \leq i \leq M-1$ is the DTFT of the multicoset samples and

$$\mathbf{x}_k(f) = X\left(f + \frac{K_k}{NT_{\text{Nyq}}}\right), \quad 1 \leq k \leq N, \tag{4}$$

where $K = \{(-\frac{N-1}{2}, \ldots, \frac{N-1}{2})\}$ for odd N (see [7] for even N). Each entry of $\mathbf{x}(f)$ is referred to as bin since it consists of a slice of the spectrum of $x(t)$. The ikth element of the $M \times N$ matrix $\mathbf{A}$ is given by

$$\mathbf{A}_{ik} = \frac{1}{NT_{\text{Nyq}}} e^{j\frac{2\pi}{N} c_i K_k}. \tag{5}$$

2) MWC sampling: The MWC [6] is composed of M parallel channels. In each channel, an analog mixing front-end, where $x(t)$ is multiplied by a mixing function $p_i(t)$, aliases the spectrum, such that each band appears in baseband. The mixing functions $p_i(t)$ are required to be periodic. We denote by T_p their period and we require $f_p = 1/T_p \geq B$. The function $p_i(t)$ has a Fourier expansion

$$p_i(t) = \sum_{l=-\infty}^{\infty} c_{il} e^{j\frac{2\pi}{T_p} lt}. \tag{6}$$

In each channel, the signal goes through a lowpass filter with cut-off frequency $f_s/2$ and is sampled at the rate $f_s \geq f_p$. For the sake of simplicity, we choose $f_s = f_p$. The overall sampling rate is Mf_s where $M \leq N = f_{\text{Nyq}}/f_s$. Repeating the calculations in [6], we derive the relation between the known DTFTs of the samples $z_i[n]$ and the unknown $X(f)$

$$\mathbf{z}(f) = \mathbf{A}\mathbf{x}(f), \qquad f \in \mathcal{F}_s, \tag{7}$$

where $\mathbf{z}(f)$ is a vector of length N with ith element $\mathbf{z}_i(f) = Z_i(e^{j2\pi f T_s})$. The unknown vector $\mathbf{x}(f)$ is given by (4). The $M \times N$ matrix $\mathbf{A}$ contains the coefficients c_{il}:

$$\mathbf{A}_{il} = c_{i,-l} = c_{il}^*. \tag{8}$$

For both sampling schemes, the overall sampling rate is

$$f_{tot} = Mf_s = \frac{M}{N} f_{\text{Nyq}}. \tag{9}$$

A. Spectrum Reconstruction

We note that the systems are identical for both sampling schemes. The only difference is the sampling matrix $\mathbf{A}$. We assume that $\mathbf{A}$ is full spark in both cases [6], [7]. We thus can derive a method for spectrum reconstruction for both sampling schemes together. We define the autocorrelation matrices $\mathbf{R_z} = \mathbb{E}[\mathbf{z}(f)\mathbf{z}^H(f)]$ and $\mathbf{R_x} = \mathbb{E}[\mathbf{x}(f)\mathbf{x}^H(f)]$. Then from (3), we have

$$\mathbf{R_z} = \mathbf{A}\mathbf{R_x}\mathbf{A^H}. \tag{10}$$

Here the exposant H denotes the Hermitian operation. Our goal is to recover $\mathbf{R_x}$ from $\mathbf{R_z}$.

Since $x(t)$ is a wide-sense stationary process, we have [13]

$$\mathbb{E}[X(\omega)X^*(\nu)] = 2\pi P_x(\omega)\delta(\omega-\nu) \tag{11}$$

where $P_x(\omega)$ denotes the spectrum of $x(t)$. Therefore $\mathbf{R_x}$ is a diagonal matrix with $\mathbf{R_x}(i,i) = P_x(f+\frac{i}{NT_s})$ [9]. It follows that

$$\mathbf{r_z} = (\mathbf{A}^* \otimes \mathbf{A})\text{vec}(\mathbf{R_x}) = (\mathbf{A}^* \otimes \mathbf{A})\mathbf{B}\mathbf{r_x} \triangleq \mathbf{\Phi}\mathbf{r_x}, \tag{12}$$

where $\mathbf{\Phi} = (\mathbf{A}^* \otimes \mathbf{A})\mathbf{B}$. Here the exposant $*$ denotes the conjugate operation. Here $\otimes$ is the Kronecker product, $\mathbf{r_z} = \text{vec}(\mathbf{R_z})$, $\mathbf{B}$ is a $N^2 \times N$ selection matrix that has a "1" at the jth column and the $[(j-1)N+j]$th row, $1 \le j \le N$ and zeros elsewhere.

We wish to recover $\mathbf{r_x}$ from $\mathbf{r_z}$. In the next section, we will derive the conditions on the sampling rate for (12) to have a unique solution.

IV. Minimal Sampling Rate

A. No sparsity Constraints

The system defined in (12) is overdetermined for $M^2 \ge N$, if $\mathbf{\Phi}$ is full column rank. The following proposition provides the condition for the system defined in (12) to have a unique solution. Due to lack of space, the proofs of the following two propositions are omitted here and will be found in a future paper.

Proposition 1. *Let $\mathbf{A}$ be a full spark $M \times N$ matrix ($M \le N$) and $\mathbf{B}$ be a $N^2 \times N$ selection matrix that has a "1" at the jth column and the $[(j-1)N+j]$th row, $1 \le j \le N$ and zeros elsewhere. The matrix $\mathbf{C} = (\mathbf{A}^* \otimes \mathbf{A})\mathbf{B}$ is full column rank if $M^2 \ge N$ and $2M > N$.*

From Proposition 1, (12) has a unique solution if $M^2 \ge N$ and $2M > N$. This can happen even for $M < N$ which is our basic assumption. If $M \ge 2$, we have $M^2 \ge 2M$. Thus, in this case, the values of M for which we obtain a unique solution are $N/2 < M < N$.

In this case, the minimal sampling rate is

$$f_{(1)} = Mf_s > \frac{N}{2}B = \frac{f_{\text{Nyq}}}{2}. \tag{13}$$

This means that even without any sparsity constraints on the signal, we can retrieve its spectrum by exploiting its stationary property, whereas the measurement vector $\mathbf{z}$ exhibits no stationary constraints in general.

B. Sparsity Constraints - Non-Blind Detection

We now consider the second scheme, where we have *a priori* knowledge on the frequency support of $x(t)$ and we assume that it is sparse. Instead of reconstructing the entire spectrum, we propose to exploit our knowledge of the signal's potential frequencies in order to further reduce the reconstruction problem and only reconstruct the potentially occupied bands.

In this scenario, the only non zero elements of $\mathbf{R_x}$ are $K_f \ll N$ diagonal elements. The reduced dimensionality spectrum is defined as

$$\hat{\mathbf{r}}_\mathbf{x} = \mathbf{M}_f \mathbf{r_x}. \tag{14}$$

Here $\mathbf{M}_f \in \mathbb{R}^{K_f \times N}$ is a matrix with elements equal to 1 at the indices of potential non-zero entries and $\hat{\mathbf{r}}_\mathbf{x} \in \mathbb{C}^{K_f \times 1}$. Furthermore, we also define $\mathbf{G}$ to be the $N \times K_f$ matrix that selects the corresponding K_f columns of $\mathbf{\Phi}$ and $\hat{\mathbf{\Phi}} = \mathbf{\Phi}\mathbf{G}$. The reduced problem can then be expressed as

$$\mathbf{r_z} = \hat{\mathbf{\Phi}}\hat{\mathbf{r}}_\mathbf{x}. \tag{15}$$

The following proposition provides the condition for the system defined in (12) to have a unique solution.

Proposition 2. *Let $\mathbf{A}$ be a full spark $M \times N$ matrix ($M \le N$) and $\mathbf{B}$ be defined as in Proposition 1. Let $\mathbf{C} = (\mathbf{A}^* \otimes \mathbf{A})\mathbf{B}$ and $\mathbf{G}$ be the $N \times K_f$ that selects any $K_f < N$ columns of $\mathbf{C}$. The matrix $\mathbf{D} = \mathbf{CG}$ is full column rank if $M^2 \ge K_f$ and $2M > K_f$.*

In this case, the minimal sampling rate is

$$f_{(2)} = Mf_s > \frac{K_f}{2}B = N_{sig}B. \tag{16}$$

Landau [11] developed a minimal rate requirement for perfect signal reconstruction in the non-blind setting, which corresponds to the actual band occupancy. Here, we find that the minimal sampling rate for perfect spectrum recovery is half the Landau rate.

C. Sparsity Constraints - Blind Detection

We now consider the second scheme, namely $x(t)$ is sparse, without any *a priori* knowledge on the support. In the previous section, we showed that $\hat{\mathbf{\Phi}}$ is full column rank, for any choice of K_f columns of $\mathbf{\Phi}$ (that correspond to K_f columns of $\mathbf{A}$), if $M^2 \ge K_f$ and $2M > K_f$. Therefore, for $M \ge 2$, if $\mathbf{r_x}$ is M-sparse, it is the unique sparsest solution of (12).

In this case, the minimal sampling rate is

$$f_{(3)} = Mf_s > K_f B = 2N_{sig}B. \tag{17}$$

As expected, this is twice the rate obtained in the previous case. As in signal recovery, the minimal rate for blind reconstruction is twice the minimal rate for non-blind reconstruction [7].

V. Simulation Results

We now demonstrate spectrum reconstruction from sub-Nyquist samples obtained close to the minimal sampling rate for the first and third scenarii, respectively. We use the MWC analog front-end [6] for the sampling stage.

It is interesting to notice that (12), which is written in the frequency domain, is valid in the time domain as well. We can therefore estimate $\mathbf{r_z}(f)$ and reconstruct $\mathbf{r_x}(f)$ in the frequency domain, or alternatively, we can estimate $\mathbf{r_z}[n]$ and reconstruct $\mathbf{r_x}[n]$ in the time domain. In order to estimate the autocorrelation matrix $\mathbf{R_z}(f)$, we first compute the estimates

of $\mathbf{z}_i(f), 1 \leq i \leq M$, $\hat{\mathbf{z}}_i(f)$, using FFT on the samples $z_i[n]$ over a finite time window. We then estimate the elements of $\mathbf{R}_{\mathbf{z}}(f)$ as

$$\hat{\mathbf{R}}_{\mathbf{z}}(i,j,f) = \frac{1}{P}\sum_{p=1}^{P} \hat{\mathbf{z}}^p(i,f)\hat{\mathbf{z}}^p(j,f), \quad f \in \mathcal{F}_s, \qquad (18)$$

where P is the number of frames for the averaging of the spectrum and $\hat{\mathbf{z}}^p(i,f)$ is the value of the FFT of the samples $\mathbf{z}_i[n]$ at the frequency f and the pth frame. In order to estimate the autocorrelation matrix $\mathbf{R}_{\mathbf{z}}[n]$ in the time domain, we perform a convolution between the samples $z_i[n]$ over a finite time window as

$$\hat{\mathbf{R}}_{\mathbf{z}}[i,j,n] = \frac{1}{P}\sum_{p=1}^{P} z_i^p[n] * z_j^p[n], \quad n \in [0, T/T_{\text{Nyq}}]. \qquad (19)$$

We first consider the spectrum reconstruction of a non sparse signal. Let $x(t)$ be white Gaussian noise with variance 100, and Nyquist rate $f_{\text{Nyq}} = 10GHz$ with two stop bands. We consider $N = 65$ spectral bands and $M = 33$ analog channels, each with sampling rate $f_s = 154MHz$ and with $N_s = 131$ samples each. The overall sampling rate is therefore equal to 50.77% of the Nyquist rate. Figure 1 shows the original and the reconstructed spectrum at half the Nyquist rate (both with averaging over $P = 1000$).

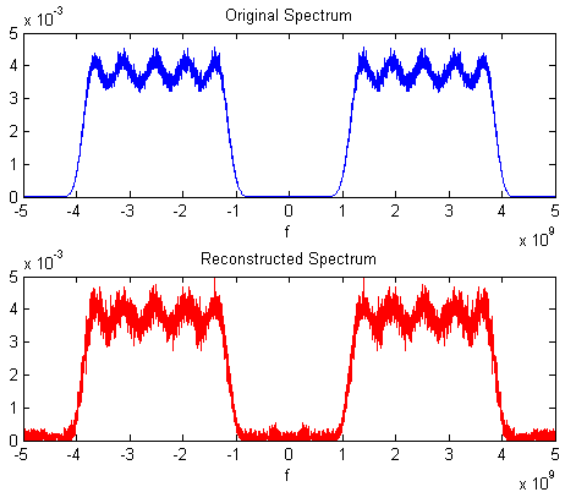


Fig. 1. Original and reconstructed spectrum of a non sparse signal at half the Nyquist rate.

We now consider the blind reconstruction of a sparse signal. Let the number of potentially active transmissions $N_{\text{sig}} = 6$ and the actual number of active transmissions be 3. Each transmission is white Gaussian noise with variance 100 and Nyquist rate $f_{\text{Nyq}} = 10GHz$, filtered by a bandpass filter whose central frequency is drawn uniformly at random and whose bandwidth is $B = 120Mhz$. We consider $N = 75$ spectral bands and $M = 13$ analog channels, each with sampling rate $f_s = 133MHz$ and with $N_s = 131$ samples each. The overall sampling rate is equal to 110% of the minimal rate (17). Figure 2 shows the original and the reconstructed spectrum at 17.3% of the Nyquist rate (both with averaging over $P = 1000$ frames).

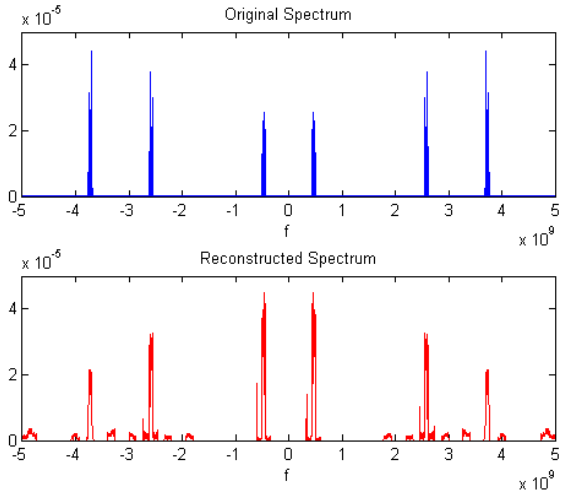


Fig. 2. Original and reconstructed spectrum of a non sparse signal at 17.3% of the Nyquist rate.

We note that the difference between the original and the reconstructed spectrum comes from the fact that the matrix $\mathbf{R}_{\mathbf{x}}(f)$ is not perfectly diagonal.

REFERENCES

[1] FCC, "Spectrum policy task force report: Federal communications commission, tech. rep. 02-135. [online]," *http://www.gov/edocs_public/attachmatch/DOC228542A1.pdf*, Nov. 2002.

[2] M. McHenry, "NSF spectrum occupancy measurements project summary. shared spectrum co., tech. rep. [online]," *http://www.sharedspectrum.com*, Aug. 2005.

[3] R. I. C. Chiang, G. B. Rowe, and K. W. Sowerby, "A quantitative analysis of spectral occupancy measurements for cognitive radio," *Proc. of IEEE Vehicular Technology Conference*, pp. 3016–3020, Apr. 2007.

[4] J. Mitola, "Software radios: Survey, critical evaluation and future directions," *IEEE Aerosp. Electron. Syst. Mag*, vol. 8, pp. 25–36, Apr. 1993.

[5] J. Mitola and C. Q. Maguire Jr., "Cognitive radio: Making software radios more personal," *IEEE Personal Commun.*, vol. 6, pp. 13–18, Aug. 1999.

[6] M. Mishali and Y. C. Eldar, "From theory to practice: Sub-Nyquist sampling of sparse wideband analog signals," *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 2, pp. 375–391, Apr. 2010.

[7] ——, "Blind multi-band signal reconstruction: Compressed sensing for analog signals," *IEEE Trans. on Signal Processing*, vol. 57, no. 3, pp. 993–1009, Mar. 2009.

[8] ——, "Sub-Nyquist sampling: Bridging theory and practice," *IEEE Signal Proc. Magazine*, vol. 28, no. 6, pp. 98–124, Nov. 2011.

[9] M. A. Lexa, M. E. Davies, J. S. Thompson, and J. Nikolic, "Compressive power spectral density estimation," *IEEE ICASSP*, 2011.

[10] D. D. Arianandа and G. Leus, "Compressive wideband power spectrum estimation," *IEEE Trans. on Signal Processing*, vol. 60, pp. 4775–4789, Sept. 2012.

[11] H. Landau, "Necessary density conditions for sampling and interpolation of certain entire functions," *Acta Math*, vol. 117, pp. 37–52, Jul. 1967.

[12] P. Feng and Y. Bresler, "Spectrum-blind minimum-rate sampling and reconstruction of multiband signals," *IEEE ICASSP*, vol. 3, pp. 1688 – 1691, May 1996.

[13] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*. McGraw Hill, 1991.

Energy-aware adaptive bi-Lipschitz embeddings

Ali Sadeghian
LIONS, EPFL, Switzerland

Bubacarr Bah
LIONS, EPFL, Switzerland

Volkan Cevher
LIONS, EPFL, Switzerland

***Abstract*—We propose a dimensionality reducing matrix design based on training data with constraints on its Frobenius norm and number of rows. Our design criteria is aimed at preserving the distances between the data points in the dimensionality reduced space as much as possible relative to their distances in original data space. This approach can be considered as a deterministic Bi-Lipschitz embedding of the data points. We introduce a scalable learning algorithm, dubbed AMUSE, and provide a rigorous estimation guarantee by leveraging game theoretic tools. We also provide a generalization characterization of our matrix based on our sample data. We use compressive sensing problems as an example application of our problem, where the Frobenius norm design constraint translates into the sensing energy.**

I. Introduction

Embedding of high dimensional data into lower dimensions is almost a classical subject. Random projections is one way of doing such embeddings and this method rely on the famous Johnson-Lindenstrauss (JL) lemma [1]. Recently, JL mappings have also found use in compressed sensing (CS), which is a promising alternative to Nyquist sampling [2]. The current CS theory uses random, non-adaptive matrices and provide recovery guarantees for highly under sampled signals. An key component in the analysis of CS recovery is the restricted isometry property (RIP), [3], [4].

Definition 1 ([3]): A matrix $\mathbf{\Phi}$ satisfies the RIP of order k if the following holds for all vectors $\mathbf{z}$, which has at most k nonzero entries (i.e., k-sparse):

$$(1-\delta_k)\|\mathbf{z}\|_2^2 \leq \|\mathbf{\Phi z}\|_2^2 \leq (1+\delta_k)\|\mathbf{z}\|_2^2. \tag{1}$$

The RIP constant (RIC) of $\mathbf{\Phi}$ of order k is the smallest δ_k for which (1) holds. In the sequel, we use δ without explicit reference to k for the RIC.

In this paper, we consider *adaptivity* in matrix design. Our setting is as follows: we are given a representative data set which can well-approximate an unknown signal. Using this data set, we would like to design a CS matrix that incorporates time and energy constraints while trying to approximate the best RIP matrix. We provide that the embedding we learn is also generalizable to some extent, that is, if a signal is drawn within ϵ of the data set, then the matrix will have good RIC. We formulate the matrix learning problem into a semidefinite program (SDP) and propose an algorithm leveraging tools from game theory.

This work was supported in part by the European Commission under Grant MIRG-268398, ERC Future Proof, SNF 200021-132548.

The main contribution of this work is that, to the best of our knowledge, it is the first deterministic design that is adaptive to data, uses RIP and gives provable approximation guarantees. A salient feature of our approach is that the design has the *digital fountain* property, which makes it nested, that is, if the measurements are not enough, we can still increase the measurements without changing the previous rows of the matrix. In addition, our approach incorporates an important criteria: the *energy constraint*, which may also be important for applications beyond CS. The algorithm we propose is also highly *scalable*, that is, it works in linear space in the matrix size because it only keep the matrix factors. Experimentally, using the matrices we design for CS seems promising as our matrices outperform those of random projections.

Notation: We define the set of k-sparse vectors as $\Sigma_k := \{\mathbf{z} \in \mathbb{R}^n : \|\mathbf{z}\|_0 \leq k\}$; and the set $\Xi_r := \left\{\mathbf{X} \in \mathbb{S}_+^{n\times n} : \text{rank}(\mathbf{X}) \leq r \text{ and } \|\mathbf{X}\|_{\text{tr}} \leq \lambda\right\}$ for scalars $r > 0$ and $\lambda > 0$, where $\mathbb{S}_+^{n\times n}$ is the set of positive semidefinite (PSD) matrices. We denote the n-dimensional simplex by Δ^n.

Definition 2 ([5]): Given $\mathbf{x}_l \in \mathcal{X} \subset \Sigma_k$ we define the set of normalized secants vectors of $\mathcal{X}$ as:

$$\mathcal{S}(\mathcal{X}) := \left\{\mathbf{v}_{ij} = \frac{\mathbf{x}_i - \mathbf{x}_j}{\|\mathbf{x}_i - \mathbf{x}_j\|_2} \text{ for } i \neq j\right\}. \tag{2}$$

Ouline: Section II is problem statement with a bit of background; while Section III formulates the problem and presents the algorithm. We analyse the algorithm and give generalization bounds in Section IV, followed by empirical results from simulations and conclusions in Sections V and VI respectively.

II. Background and problem description

The CS literature heavily relies on random matrices in establishing recovery guarantees. There has also been also progress in obtaining structured matrices via randomization. However, for CS to live up to its promise, real applications must be able to use data adaptive matrices. Attempts have been made in this direction that include what is referred to as optimizing projection matrices which entails reducing the correlation between normalised data points (dictionary) of the given data set, see [6], [7]. Our work is in this direction as is [5]. Precisely, this work build on what was done in [5] by learning a projection (embedding) matrix from a given data set via the RIP. However, in sharp contrast to [5], our solution provides rigorous approximation guarantees.

To set up the problem, let us assume that we are given a set of $p \gg n$ sample points (training set) $\mathcal{X} = \{\mathbf{x}_j\}_{j=1}^{p}$. Then we impose that the embedding matrix we are learning $\mathbf{\Phi}$ satisfy RIP on the pairwise distances of the points in $\mathcal{X}$, that is $\mathbf{\Phi}$ satisfies (1) with $\mathbf{z}$ replaced by $\mathbf{x}_i - \mathbf{x}_j$ for all $\mathbf{x}_i, \mathbf{x}_j \in \mathcal{X}$ where $i \neq j$. $\mathbf{\Phi}$ is bi-Lipschitz due to the RIP construct. Theoretical guarantees for this approach relies on results from differential geometry, see [5] and the references therein.

If we replaced $\mathbf{z}$ in (1) by $\mathbf{x}_i - \mathbf{x}_j$ and normalized the pairwise distances, then the RIP condition (1) on $\mathcal{S}(\mathcal{X})$ becomes $(1-\delta) \leq \mathbf{v}_{ij}^T \mathbf{\Phi}^T \mathbf{\Phi} \mathbf{v}_{ij} \leq (1-\delta)$. This expression simplifies to $|\mathbf{v}_{ij}^T \mathbf{\Phi}^T \mathbf{\Phi} \mathbf{v}_{ij} - 1| \leq \delta$ for each $i \neq j$. Re-indexing the $\mathbf{v}_{ij}$ to $\mathbf{v}_l$ for $l = 1, \ldots, M$, where $M = \binom{p}{2}$, we form the M secant vectors $\mathcal{S}(\mathcal{X}) = \{\mathbf{v}_1, \ldots, \mathbf{v}_M\}$ into an $n \times M$ matrix $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_M]$ and let $\mathbf{B} = \mathbf{\Phi}^T \mathbf{\Phi}$. Then we define a linear transform $\boldsymbol{\mathcal{A}} : \mathbb{S}_+^{n \times n} \to \mathbb{R}^M$ as:

$$\boldsymbol{\mathcal{A}}(\mathbf{B}) := \operatorname{diag}\left(\mathbf{V}^T \mathbf{B} \mathbf{V}\right), \tag{3}$$

where $\operatorname{diag}(\mathbf{H})$ denotes a vector of the entries of the principal diagonal of the matrix $\mathbf{H}$. Note that the rank of $\mathbf{B}$ is the same as that of $\mathbf{\Phi}$ and $\mathbf{B}$ is a PSD self-adjoint matrix. In addition, we place a constraint on the energy of $\mathbf{B}$ to be a fixed budget, say b, which adds a trace constraint to our problem and in practice may translates for example to having the entries of $\mathbf{\Phi}$ to all have a certain magnitude range. So our problem of adaptively learning an energy-aware RIP matrix $\mathbf{\Phi}$ and an RIC δ is equivalent to the following trace constrained affine rank minimization (ARM) problem:

$$\begin{aligned} \min_{\mathbf{B}} \quad & \|\boldsymbol{\mathcal{A}}(\mathbf{B}) - \mathbf{1}_M\|_\infty \\ \text{s.t.} \quad & \mathbf{B} \succeq 0, \quad \operatorname{rank}(\mathbf{B}) = r, \quad \operatorname{trace}(\mathbf{B}) = b. \end{aligned} \tag{4}$$

In [5], they solve a different problem by constraining the value of δ. Then they use eigen-decomposition to reach a number of samples. We directly take the constraints, design the matrix and give approximation guarantees. In our case, our algorithm returns the factors directly, which reduces the post processing costs such as taking eigendecompositions.

III. Proposed design and our algorithm

Problem (4), as is common practice for ARM problems, can be relaxed as follows:

$$\begin{aligned} \min_{\mathbf{B}} \quad & \|\mathbf{y} - \boldsymbol{\mathcal{A}}(\mathbf{B})\|_\infty \\ \text{s.t.} \quad & \operatorname{rank}(\mathbf{B}) \leq r \quad \text{and} \quad \|\mathbf{B}\|_{\text{tr}} \leq b. \end{aligned} \tag{5}$$

where $\mathbf{y} = \mathbf{1}_M$ and $\|\mathbf{B}\|_{\text{tr}} \leq b$ captures the PSD and the trace constraints. Based on the work in [8], we reformulate (5) as a minimax game next.

A. Reformulation of (5)

We first define a linear map $\boldsymbol{\mathcal{A}}_+ : \mathbb{S}^{n \times n} \to \mathbb{R}^{2M}$ where $\boldsymbol{\mathcal{A}}_+(\mathbf{B})$ is a concatenation of $\boldsymbol{\mathcal{A}}(\mathbf{B})$ and $-\boldsymbol{\mathcal{A}}(\mathbf{B})$ that is: $\boldsymbol{\mathcal{A}}_+(\mathbf{B}) = [\boldsymbol{\mathcal{A}}(\mathbf{B})^T, -\boldsymbol{\mathcal{A}}(\mathbf{B})^T]^T$, and correspondingly set $\mathbf{f} = [\mathbf{y}^T, -\mathbf{y}^T]^T$. Therefore, we have

$$\begin{aligned} \|\mathbf{y} - \boldsymbol{\mathcal{A}}(\mathbf{B})\|_\infty = \max_{i \in [2M]} |\, [\boldsymbol{\mathcal{A}}_+(\mathbf{B}) - \mathbf{f}]_i \,| = \\ \max_{i \in [2M]} \mathbf{e}_i^T \left(\boldsymbol{\mathcal{A}}_+(\mathbf{B}) - \mathbf{f}\right) = \max_{\mathbf{N} \in \Delta^{2M}} \mathcal{L}(\mathbf{N}, \mathbf{B}), \end{aligned} \tag{6}$$

where $\mathcal{L}(\mathbf{N}, \mathbf{B}) := \langle \mathbf{N}, (\boldsymbol{\mathcal{A}}_+(\mathbf{B}) - \mathbf{f}) \rangle$ and $\mathbf{e}_i$ is the canonical basis vector. The last equality in (6) is due to the fact that the maximum of a linear program occurs at a boundary point of the simplex Δ^{2M}. This reduces problem (5) to a minimax problem:

$$\min_{\mathbf{B} \in \Xi_r} \quad \max_{\mathbf{N} \in \Delta^{2M}} \quad \mathcal{L}(\mathbf{N}, \mathbf{B}) \tag{7}$$

where Ξ_r is the primal set, Δ^{2M} is the dual set and the mapping $\mathcal{L} : \Xi_r \times \Delta^{2M} \to \mathbb{R}$ is referred to as the *loss function* in game theory. We would need the following $\mathcal{L}_{\max} := \max_{\mathbf{N},\mathbf{B}} |\mathcal{L}(\mathbf{N}, \mathbf{B})| = \|\boldsymbol{\mathcal{A}}(\mathbf{B})\|_\infty + \|\mathbf{y}\|_\infty$. Note that $\boldsymbol{\mathcal{A}}_+^* : \mathbb{R}^{2M} \to \mathbb{S}^{n \times n}$, which is the adjoint of $\boldsymbol{\mathcal{A}}_+$, can be expressed in terms of the adjoint of $\boldsymbol{\mathcal{A}}$, denoted by $\boldsymbol{\mathcal{A}}^*$, precisely $\boldsymbol{\mathcal{A}}_+^*(\mathbf{w}) = \boldsymbol{\mathcal{A}}^*(\mathbf{w}_1 - \mathbf{w}_2)$ for $\mathbf{w} = [\mathbf{w}_1, \mathbf{w}_2]^T$ where $\mathbf{w}_1, \mathbf{w}_2 \in \mathbb{R}^M$.

B. AMUSE algorithm

We now propose an algorithm that solves the minimax game (7) with provable theoretical guarantees: see **Algorithm 1**. It is important to note that the algorithm works with rank-1 updates B^t (in a matter similar to the conditional gradient descent algorithms [9]). As a result, after r iterations, our algorithm returns an estimator $\widehat{\mathbf{B}} = \frac{1}{r} \sum_{t=1}^{r} \mathbf{B}^t$. As we do not explicitly compute the product of the factors, the algorithm is scalable since each factor corresponds to 1 measurement. Moreover, we bound the recovery error as thus:

$$\|\boldsymbol{\mathcal{A}}(\widehat{\mathbf{B}}) - \mathbf{y}\|_\infty \leq \min_{\mathbf{B} \in \chi_r^*} \|\boldsymbol{\mathcal{A}}(\mathbf{B}) - \mathbf{y}\|_\infty + \mathcal{O}\left(\frac{1}{\sqrt{r}}\right).$$

This is the first approximation bound for obtaining such sensing matrices.

Essentially, the MUSE for ARM (AMUSE) algorithm we propose is a modification of the Multiplicative Update Selector and Estimator (MUSE) algorithm for learning to play repeated games proposed in [8]. The MUSE itself can be thought of as a restatement of the Multiplicative Weights Algorithm (MWA), which in turn uses the Weighted Majority Algorithm, see [8] and references therein. We also point out also that the multiplicative updating has connections to Frank-Wolfe and related algorithms [10].

Steps 2 and 3 of the loop of AMUSE performs the multiplicative update of the dual variable $\mathbf{N}$ and the update is exactly the same as in MUSE for a given primal variable at iteration t, $\mathbf{B}^t$. Therefore the step size η remains the same as in the MUSE algorithm, [8]; that is $\eta = \ln\left(1 + \sqrt{2\ln(2M)/r}\right)$. As a result, the theoretical guarantees given in [8] for MUSE also holds for AMUSE. Basically, for a fixed matrix at iteration t, $\mathbf{B}^t$, the proof for the multiplicative update in [8] for the vector case remains the same.

Algorithm 1 MUSE for ARM (AMUSE)

Input: $\mathbf{y}$, η

Output: $\widehat{\mathbf{B}} \approx \mathbf{B}^*$ with $\text{rank}(\widehat{\mathbf{B}}) \leq r$

Initialize $\mathbf{N}^1 = \frac{1}{2M}\mathbf{1}_{2M}$

For $t = 1, \ldots, r$ **do**

1. Find $\mathbf{B}^t = \text{argmin}_{\|\mathbf{B}\|_{\text{tr}} \leq 1} \ \mathcal{L}(\mathbf{N}^t, \mathbf{B})$
2. Set $\mathbf{Q}_j^{t+1} = \mathbf{N}_j^t \cdot e^{\frac{\eta \cdot \mathcal{L}(\mathbf{e}_j, \mathbf{B}^t)}{\mathcal{L}_{\max}}}$ for $j \in [2M]$
3. Update $\mathbf{N}^{t+1} = \frac{\mathbf{Q}^{t+1}}{\sum_{j=1}^{2M} \mathbf{Q}_j^{t+1}}$

End for

Return $\widehat{\mathbf{B}} = \frac{1}{r}\sum_{t=1}^{r} \mathbf{B}^t$

Note that the main and crucial difference between AMUSE and MUSE is the first step of the loop where we update our primal variable $\mathbf{B}$ given our dual variable at iteration t, $\mathbf{N}^t$, by $\mathbf{B}^t = \text{argmin}_{\|\mathbf{B}\|_{\text{tr}} \leq 1} \ \mathcal{L}(\mathbf{N}^t, \mathbf{B})$. These updates have rank 1 and hence their linear combination, $\widehat{\mathbf{B}}$, has rank at most r, since rank is sub-additive.

AMUSE is used to approximate problem (4) by rescaling to meet the trace constraint. The parameter η remain the same and $\mathcal{L}_{\max} = 1 + \max_i \max_j v_{ij}^2$ where v_{ij} is the (i,j) entry of $\mathbf{V}$.

IV. Analysis

A. AMUSE guarantees

The following theorem formalizes our claim that the AMUSE algorithm outputs an approximate solution $\widehat{\mathbf{B}}$ with $\text{rank}(\widehat{\mathbf{B}}) \leq r$ with a bounded ℓ_∞ loss in the measurement domain after r iterations. The proof of this theorem use Lemma 4.1 of [8].

Theorem 1: Let AMUSE return $\widehat{\mathbf{B}}$ after r iterations. Then $\text{rank}(\widehat{\mathbf{B}}) \leq r$ and $\|\boldsymbol{\mathcal{A}}(\widehat{\mathbf{B}}) - \mathbf{y}\|_\infty$ is at most

$$\|\mathbf{e}\|_\infty + \left(1 + \sqrt{2}\right) \cdot \left(2\|\boldsymbol{\mathcal{A}}(\widehat{\mathbf{B}})\|_\infty + \|\mathbf{e}\|_\infty\right) \sqrt{\frac{\ln(2M)}{r}},$$

where $\mathbf{e}$ measures the perturbation of the linear model.

Proof: We sketch the proof as follows, for details see [8]. By the definition of $\boldsymbol{\mathcal{A}}$, $\mathbf{y}$ and $\mathcal{L}$, $\|\boldsymbol{\mathcal{A}}(\widehat{\mathbf{B}}) - \mathbf{y}\|_\infty = \max_{\mathbf{N}} \mathcal{L}(\mathbf{N}, \widehat{\mathbf{B}})$. Then we first show that $\min_{\mathbf{B}} \max_{\mathbf{N}} \mathcal{L}(\mathbf{N}, \mathbf{B}) + (1+\sqrt{2})\mathcal{L}_{\max}\sqrt{\frac{\ln(2M)}{r}}$ upper bounds $\max_{\mathbf{N}} \mathcal{L}(\mathbf{N}, \widehat{\mathbf{B}})$, a key ingredient of which is the min-max theorem. Next we deduce that $\min_{\mathbf{B}} \max_{\mathbf{N}} \mathcal{L}(\mathbf{N}, \mathbf{B}) = \min_{\mathbf{B}} \|\boldsymbol{\mathcal{A}}(\mathbf{B}) - \mathbf{y}\|_\infty \leq \|\mathbf{e}\|_\infty$. Then, using the triangle inequality we bound $\mathcal{L}_{\max}$ by bounding $\|\mathbf{y}\|_\infty$ as thus: $\mathcal{L}_{\max} = \|\boldsymbol{\mathcal{A}}(\mathbf{B})\|_\infty + \|\mathbf{y}\|_\infty$ which is upper bounded by $2\|\boldsymbol{\mathcal{A}}(\mathbf{B})\|_\infty + \|\mathbf{e}\|_\infty$. ■

Furthermore, we bound the error of the output of AMUSE for the RIP matrix learning problem in Corollary 1 which follows from Theorem 1.

Corollary 1: Let AMUSE learn an RIP matrix $\widehat{\mathbf{B}}$ from a given data set $\mathcal{X}$ after r iterations with RIC $\hat{\delta}$. Assume that the optimal RIP matrix for that $\mathcal{X}$ has RIC δ^*. Then $\widehat{\mathbf{B}}$ has $\text{rank}(\widehat{\mathbf{B}}) \leq r$ and

$$\|\boldsymbol{\mathcal{A}}(\widehat{\mathbf{B}}) - \mathbf{1}_M\|_\infty \leq \delta^* + 2(1+\sqrt{2})\sqrt{\frac{\ln(2M)}{r}}.$$

This implies that if the optimal solution $\boldsymbol{\Phi}^*$ has RIC δ^* on the training set, then our approximation, $\widehat{\boldsymbol{\Phi}}$, of $\boldsymbol{\Phi}^*$ also satisfies RIP on these data points but with a slightly larger constant $\hat{\delta} \leq \delta^* + \mathcal{O}\left(1/\sqrt{r}\right)$. As the dimensions increase, we approximate the best RIP constant for the given dataset.

B. Generalization bounds

Interestingly, we can provably approximate the optimal RIC even for points that are outside our sample points as stated in the following proposition.

Proposition 1: Given the pair δ and $\boldsymbol{\Phi}$ as the optimal solution to (4), $\boldsymbol{\Phi}$ applied to any $\boldsymbol{z}$ with $\|\boldsymbol{z} - \mathbf{x}\|_2 \leq \epsilon$ for all $\mathbf{x} \in \mathcal{X}$ and $\epsilon \in [0, 1)$ gives an RIC, $\bar{\delta}$, bounded as follows:

$$\bar{\delta} \leq (\delta + \epsilon)/(1 - \epsilon). \tag{8}$$

Proof: Since $\boldsymbol{\Phi}$ is linear w.l.o.g let $\|\mathbf{x}\|_2 = 1$. For any $\boldsymbol{z}$ such that $\|\boldsymbol{z} - \mathbf{x}\|_2 \leq \epsilon$ and $\|\boldsymbol{z}\|_2 = 1$ then $\|\boldsymbol{\Phi}\boldsymbol{z}\|_2$ can be written as:

$$\|\boldsymbol{\Phi}\left(\mathbf{x} - (\boldsymbol{z} - \mathbf{x})\right)\|_2 \leq \|\boldsymbol{\Phi}\mathbf{x}\|_2 + \|\boldsymbol{\Phi}(\boldsymbol{z} - \mathbf{x})\|_2$$

using the triangle inequality. Let α_1 be the smallest constant such that $\|\boldsymbol{\Phi}\boldsymbol{z}\|_2 \leq (1 + \alpha_1)\|\boldsymbol{z}\|_2$ then with the definition of δ from the above inequality we have

$$\|\boldsymbol{\Phi}\boldsymbol{z}\|_2 \leq (1 + \delta)\|\mathbf{x}\|_2 + (1 + \alpha_1)\|\boldsymbol{z} - \mathbf{x}\|_2.$$

Evaluating and upper bounding the norms and using the definition of α_1 gives

$$(1 + \alpha_1) \leq (1 + \delta) + (1 + \alpha_1)\epsilon.$$

This simplifies to $\alpha_1 \leq (\delta + \epsilon)/(1 - \epsilon)$. Similarly, we lower bound $\|\boldsymbol{\Phi}\boldsymbol{z}\|_2$ and have an α_2 to be the largest constant such that $\|\boldsymbol{\Phi}\boldsymbol{z}\|_2 \geq (1 - \alpha_1)\|\boldsymbol{z}\|_2$, this leads to a bound on α_2 as thus: $\alpha_2 \leq (\delta + \epsilon)/(1 + \epsilon)$. The RIC, $\bar{\delta}$, is therefore given by $\max(\alpha_1, \alpha_2)$ and for the values of ϵ considered this is α_1, hence (8). ■

V. Empirical results

We use the synthetic data set of images of translations of white squares in a black background from [5]. In the first experiment we investigate the dependence of RIC we learn on the number of rows (or rank) of the $\boldsymbol{\Phi}$ we learn. Here, we use $M = 2000$ number of secants vectors. We use the same for PCA projected to meet the trace constraint of our problem (4) and also generate a random Gaussian matrix also constrained to have trace as our problem. Figure 1 displays this comparison, where our method clearly outperforms PCA and random designs.

In the second experiment we learn a $\boldsymbol{\Phi}$ from the data and use it to encode a randomly selected subset of $\mathcal{X}$ corrupted with Gaussian noise of varying signal-to-noise ratio (SNR).

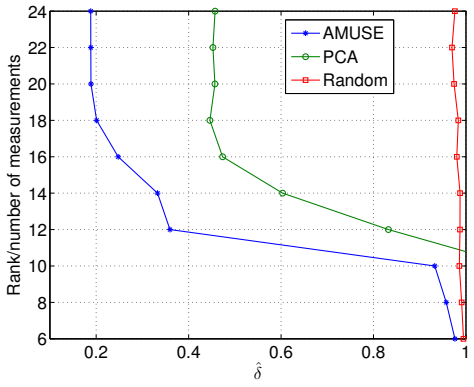


Fig. 1. A plot of the number of measurements (or rank of the $\widehat{\mathbf{\Phi}}$) as a function of the RIC $\widehat{\delta}$ for data points with an ambient dimension $n = 256$.

We then do Basis Pursuit denoising to decode these points. For comparison we use a Gaussian matrix with the trace-constrained and compute the mean-square error (MSE) over the subset. The results are displayed in Figure 2, which show that our approach outperforms the random projections due to its adaptivity to the underlying data manifold. Note that in this experiment, we simply searched over Frobenius norm constraint to approximate the RIC without any energy constraint.

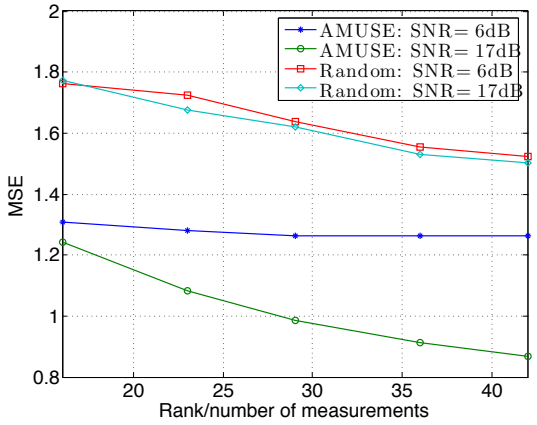


Fig. 2. CS recovery performance of our adaptive approach compared to energy constrained random projections.

VI. Conclusions

We reformulate the adaptive learning of a data embedding into an optimization problem and propose an algorithm that approximately solves this problem with provable guarantees. We show generalizability of our embedding to a test data set ϵ away from the training set in terms of the RIC of the embedding matrix learnt. Our experiments show better performance of our derived matrices as compared to random designs with regard to the empirical RIC and CS recovery.

References

[1] W. B. Johnson and J. Lindenstrauss, "Extensions of lipschitz mappings into a hilbert space," *Contemporary mathematics*, vol. 26, no. 189-206, p. 1, 1984.

[2] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inform. Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[3] E. Candes and T. Tao, "Decoding by linear programming," *Information Theory, IEEE Transactions on*, vol. 51, no. 12, pp. 4203–4215, 2005.

[4] H. Rauhut, "Compressive sensing and structured random matrices," *Theoretical foundations and numerical methods for sparse recovery*, vol. 9, pp. 1–92, 2010.

[5] C. Hegde, A. Sankaranarayanan, W. Yin, and R. Baraniuk, "A convex approach for learning near-isometric linear embeddings," *preparation, August*, 2012.

[6] M. Elad, "Optimized projections for compressed sensing," *Signal Processing, IEEE Transactions on*, vol. 55, no. 12, pp. 5695–5702, 2007.

[7] J. M. Duarte-Carvajalino and G. Sapiro, "Learning to sense sparse signals: Simultaneous sensing matrix and sparsifying dictionary optimization," *Image Processing, IEEE Transactions on*, vol. 18, no. 7, pp. 1395–1408, 2009.

[8] S. Jafarpour, R. Schapire, and V. Cevher, "Compressive sensing meets game theory," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 3660–3663.

[9] D. P. Bertsekas, "Nonlinear programming," 1999.

[10] K. L. Clarkson, "Coresets, sparse greedy approximation, and the frank-wolfe algorithm," *ACM Transactions on Algorithms (TALG)*, vol. 6, no. 4, p. 63, 2010.

Randomized Singular Value Projection

Stephen Becker
UPMC Paris 6
Email: stephen.becker@upmc.fr

Volkan Cevher
Ecole Polytechnique Fédérale de Lausanne
Email: volkan.cevher@epfl.ch

Anastasios Kyrillidis
Ecole Polytechnique Fédérale de Lausanne
Email: anastasios.kyrillidis@epfl.ch

***Abstract*—Affine rank minimization algorithms typically rely on calculating the gradient of a data error followed by a singular value decomposition at every iteration. Because these two steps are expensive, heuristic approximations are often used to reduce computational burden. To this end, we propose a recovery scheme that merges the two steps with randomized approximations, and as a result, operates on space proportional to the degrees of freedom in the problem. We theoretically establish the estimation guarantees of the algorithm as a function of approximation tolerance. While the theoretical approximation requirements are overly pessimistic, we demonstrate that in practice the algorithm performs well on the quantum tomography recovery problem.**

I. Introduction

In many signal processing and machine learning applications, we are given a set of observations $\mathbf{y} \in \mathbb{R}^p$ of a rank-r matrix $\mathbf{X}^\star \in \mathbb{R}^{m\times n}$ as $\mathbf{y} = \boldsymbol{\mathcal{A}}\mathbf{X}^\star + \boldsymbol{\varepsilon}$ via the linear operator $\boldsymbol{\mathcal{A}} : \mathbb{R}^{m\times n} \to \mathbb{R}^p$, where $r \ll \min\{m,n\}$ and $\boldsymbol{\varepsilon} \in \mathbb{R}^p$ is additive noise. As a result, we are interested in the solution of

$$\begin{aligned} \underset{\mathbf{X}\in\mathbb{R}^{m\times n}}{\text{minimize}} \quad & f(\mathbf{X}) \\ \text{subject to} \quad & \operatorname{rank}(\mathbf{X}) \le r, \end{aligned} \tag{1}$$

where $f(\mathbf{X}) := \|\mathbf{y} - \boldsymbol{\mathcal{A}}\mathbf{X}\|_2^2$ is the data error. While the optimization problem in (1) is non-convex, it is possible to obtain robust recovery with provable guarantees via iterative greedy algorithms (SVP) [1], [2] or convex relaxations [3], [4] from measurements as few as $p = \mathcal{O}(r(m+n-r))$.

Currently, there is a great interest in designing algorithms to handle large scale versions of (1) and its variants. As a concrete example, consider quantum tomography (QT), where we need to recover low-rank density matrices from dimensionality reducing Pauli measurements [5]. In this problem, the size of these density matrices grows exponentially with the number of quantum bits. Other collaborative filtering problems, such as the Netflix challenge, also require huge dimensional optimization. Without careful implementations or non-conventional algorithmic designs, existing algorithms quickly run into time and memory bottlenecks.

These computational difficulties typically revolve around two critical issues. First, virtually all recovery algorithms require calculating the gradient $\nabla f(\mathbf{X}) = 2\boldsymbol{\mathcal{A}}^*(\boldsymbol{\mathcal{A}}(\mathbf{X}) - \mathbf{y})$ at an intermediate iterate $\mathbf{X}$, where $\boldsymbol{\mathcal{A}}^*$ is the adjoint of $\boldsymbol{\mathcal{A}}$. When the range of $\boldsymbol{\mathcal{A}}^*$ contains dense matrices, this forces algorithms to use memory proportional to $\mathcal{O}(mn)$. Second, after the iterate is updated with the gradient, projecting onto the low-rank space requires a partial singular value decomposition (SVD). This is usually problematic for the initial iterations of convex algorithms, where they may have to perform full SVD's. In contrast, greedy algorithms [2] fend off the complexity of full SVD's, since they need fixed rank projections, which can be approximated via Lanczos or randomized SVD's [6].

Algorithms that avoid these two issues do exist, such as [7]–[10], and are typically based on the Burer-Monteiro splitting [11]. The main idea in Burer-Monteiro splitting is to remove the non-convex rank constraint by directly embedding it into the objective: as opposed to optimizing $\mathbf{X}$, splitting algorithms directly work with its fixed factors $\mathbf{U}\mathbf{V}^T = \mathbf{X}$ in an alternating fashion, where $\mathbf{U} \in \mathbb{R}^{m\times\hat{r}}$ and $\mathbf{V} \in \mathbb{R}^{n\times\hat{r}}$ for some $\hat{r} \ge r$. Unfortunately, rigorous guarantees are difficult.[1] The work [12] has shown approximation guarantees if $\boldsymbol{\mathcal{A}}$ satisfies a restricted isometry property with constant $\delta_{2r} \le \kappa^2/(100r)$ (noiseless), where $\kappa = \sigma_1(\mathbf{X}^\star)/\sigma_r(\mathbf{X}^\star)$, or $\delta_{2r} \le 1/(3200r^2)$ for a bound independent of κ. The authors suggest that these bounds may be tightened based on the good empirical performance of the algorithm.

In this paper, we merge the gradient calculation and the singular value projection steps into one and show that this not only removes a huge computational burden, but suffers only a minor convergence speed drawback in practice. Our contribution is a natural but non-trivial fusion of the Singular Value Projection (SVP) algorithm in [1] and the approximate projection ideas in [2]. The SVP algorithm is a hard-thresholding algorithm that has been considered in [1], [13]. Inexact steps in SVP have been considered as a heuristic [13] but have not been incorporated into an overall convergence result. A non-convex framework for affine rank minimization (including variants of the SVP algorithm) that utilizes inexact projection operations with provable signal approximation and convergence guarantees is proposed in [2]. Both [1], [2] do not consider splitting techniques in the proposed schemes.

In this work, departing from [1], [2], we engineer the SVP algorithm to operate like splitting algorithms that *directly work with the factors*; this added twist decreases the per iteration requirements in terms of storage and computational complexity. Using this new formulation, each iteration is nearly as fast as in the splitting method, hence removing a

[1] If $\hat{r} \gtrsim \sqrt{p}$, then [11] shows their method obtains a global solution, but this is impractical for large p. Moreover, it is shown that the explicit rank $\hat{r}$ splitting method solves a non-convex problem that has the same local minima as (1) (if $\hat{r} = r$). However, the non-convex problems are not *equivalent* (e.g. $\mathbf{U} = \mathbf{0}$, $\mathbf{V} = \mathbf{0}$ is a stationary point for the splitting problem whereas $\mathbf{X} = \mathbf{0}$ is generally not a stationary point for (1)).

drawback to SVP in relation to splitting methods. Furthermore, we prove that, under some conditions, it is still possible to obtain perfect recovery even if the projections are inexact. Such characterizations have been used for convex [3] and non-convex [1], [2] algorithms to obtain approximation guarantees.

II. Preliminary material

Notation: $\mathcal{P}_\Omega$ is an orthogonal projection onto the closed set Ω when it exists, and $\mathcal{P}_r$ stands for $\mathcal{P}_{\{\mathbf{X}:\mathrm{rank}(\mathbf{X})\leq r\}}$ (which does exist by the Eckart-Young theorem). Computer routine names are typeset with a `typewriter font`.

R-RIP: The Rank Restricted Isometry Property (R-RIP) is a common tool used in matrix recovery [1]–[3]:

Definition 1 (R-RIP for linear operators on matrices [3])**.** *A linear operator $\boldsymbol{\mathcal{A}} : \mathbb{R}^{m\times n} \to \mathbb{R}^p$ satisfies the R-RIP with constant $\delta_r(\boldsymbol{\mathcal{A}}) \in (0,1)$ if, $\forall \mathbf{X} \in \mathbb{R}^{m\times n}$ with $\mathrm{rank}(\mathbf{X}) \leq r$,*

$$(1-\delta_r(\boldsymbol{\mathcal{A}}))\|\mathbf{X}\|_F^2 \leq \|\boldsymbol{\mathcal{A}}\mathbf{X}\|_2^2 \leq (1+\delta_r(\boldsymbol{\mathcal{A}}))\|\mathbf{X}\|_F^2. \quad (2)$$

We use the short notation δ_r to mean $\delta_r(\boldsymbol{\mathcal{A}})$.

Additional convex constraints: Consider the variant

$$\begin{aligned} \underset{\mathbf{X}\in\mathbb{R}^{m\times n}}{\text{minimize}} \quad & f(\mathbf{X}) \\ \text{subject to} \quad & \mathrm{rank}(\mathbf{X}) \leq r,\ \mathbf{X} \in \mathcal{C}, \end{aligned} \quad (3)$$

for a convex set $\mathcal{C}$. Our main interests are $\mathcal{C}_+ = \{\mathbf{X} : \mathbf{X} \succeq 0\}$ and the matrix simplex $\mathcal{C}_\Delta = \{\mathbf{X} : \mathbf{X} \succeq 0,\ \mathrm{trace}(\mathbf{X}) = 1\}$. In both cases the constraints are unitarily invariant and the projection onto these sets can be done by taking the eigenvalue decomposition and projecting the eigenvalues. Furthermore, for these specific $\mathcal{C}$, $\mathcal{P}_{\{\mathbf{X}:\mathrm{rank}(\mathbf{X})\leq r\}\cap\mathcal{C}} = \mathcal{P}_\mathcal{C} \circ \mathcal{P}_r$ (see [14]).

Approximate singular value computations: The standard method to compute a partial SVD is the Lanczos method. However, the method is somewhat hard to parallelize and it lacks theoretical bounds of the form used in Theorem 1.

Algorithm 1 `RandomizedSVD`

Finds Q such that $X \approx \mathcal{P}_Q X = QQ^*X$

Require: Function $\mathtt{h} : \widetilde{Z} \mapsto X\widetilde{Z}$
Require: Function $\mathtt{h}^* : \widetilde{Q} \mapsto X^*\widetilde{Q}$
Require: $r \in \mathbb{N}$ // Rank of output
Require: $q \in \mathbb{N}$ // Number of power iterations to perform
$\ell = r + \rho$ // Typical value of ρ is 5
Ω a $n \times \ell$ standard Gaussian matrix
$W \leftarrow \mathtt{h}(\Omega)$
$Q \leftarrow \mathtt{QR}(W)$ // The QR algorithm to orthogonalize W
for $j = 1, 2, \ldots, q$ **do**
  $Z \leftarrow \mathtt{QR}(\mathtt{h}^*(Q))$
  $Q \leftarrow \mathtt{QR}(\mathtt{h}(Z))$
end for
$Z \leftarrow \mathtt{h}^*(Q)$
$(U, \Sigma, V) \leftarrow \mathtt{FactoredSVD}(Q, I_\ell, Z)$
Let Σ_r be the best rank r approximation of Σ
return (U, Σ_r, V)

Algorithm 2 `FactoredSVD`$(\widetilde{U}, \widetilde{D}, \widetilde{V})$

Computes the SVD $U\Sigma V^*$ of the matrix X implicitly given by $X = \widetilde{U}\widetilde{D}\widetilde{V}^*$

$(U, R_U) \leftarrow \mathtt{QR}(\widetilde{U})$
$(V, R_V) \leftarrow \mathtt{QR}(\widetilde{V})$
$(u, \Sigma, v) \leftarrow \mathtt{DenseSVD}(R_U\widetilde{D}R_V^*)$
return $(U, \Sigma, V) \leftarrow (Uu, \Sigma, Vv)$

As an alternative, we turn to randomized linear algebra. On this front, we restrict ourselves to algorithms that require only multiplications, as opposed to sub-sampling entries/rows/columns, as sub-sampling is not efficient for the application we present. The randomized approach presented in Algorithm 1 has been rediscovered many times, but has seen a recent resurgence of interest due to theoretical analysis [6]:

Theorem 1 (Average Frobenius error)**.** *Suppose $\mathbf{X} \in \mathbb{R}^{m\times n}$, and choose a target rank r and oversampling parameter $\rho \geq 2$ where $\ell := r + \rho \leq \min\{m, n\}$. Calculate Q and $\mathcal{P}_Q$ via `RandomizedSVD` using $q = 0$ and set $\widetilde{\mathbf{X}} = \mathcal{P}_Q\mathbf{X}$. Then*

$$\mathbb{E}\|\mathbf{X} - \widetilde{\mathbf{X}}\|_F^2 \leq (1+\epsilon)\,\|\mathbf{X} - \mathbf{X}_r\|_F^2$$

where $\mathbf{X}_r$ is the best rank r approximation in the Frobenius norm of $\mathbf{X}$, $\widetilde{\mathbf{X}}$ has rank ℓ, and $\epsilon = \frac{r}{\rho-1}$.

The theorem follows from the proof of Thm. 10.5 in [6]. The expectation is with respect to the Gaussian r.v. in `RandomizedSVD`. For the sake of our analysis, we cannot immediately truncate $\widetilde{\mathbf{X}}$ to rank r since then the error bound in [6] is not tight enough. Thus, since $\widetilde{X}$ is rank ℓ, in practice we even observe that $\|\mathbf{X} - \widetilde{\mathbf{X}}\|_F^2 < \|\mathbf{X} - \mathbf{X}_r\|_F^2$, especially for small r, as shown in Figure 3. The figure also shows that using $q > 0$ power iterations is extremely helpful, though this is not taken into account in our analysis since there are no useful theoretical bounds. Note that variants for eigenvalues also exist; we refer to the equivalent of `RandomizedSVD` as `RandomizedEIG`, which has the property that $U = V$ and Σ need not be positive (cf., [6])

III. A projected gradient descent algorithm

Our approach is based on the projected gradient descent:

$$\mathbf{X}_{i+1} = \mathcal{P}_r^\epsilon(\mathbf{X}_{i+1} - \mu_i\nabla f(\mathbf{X}_i)), \quad (4)$$

where $\mathbf{X}_i$ is the i-th iterate, $\nabla f(\cdot)$ is the gradient of the loss function, μ_i is a step-size, and $\mathcal{P}_r^\epsilon(\cdot)$ is the approximate projector onto rank r matrices given by `RandomizedSVD`. If we include a convex constraint $\mathcal{C}$, then the iteration is

$$\mathbf{X}_{i+1} = \mathcal{P}_\mathcal{C}(\mathcal{P}_r^\epsilon(\mathbf{X}_{i+1} - \mu_i\nabla f(\mathbf{X}_i))). \quad (5)$$

In practice, Nesterov acceleration improves performance:

$$\mathbf{Y}_{i+1} = (1+\beta_i)\mathbf{X}_i - \beta_i\mathbf{X}_{i-1} \quad (6)$$
$$\mathbf{X}_{i+1} = \mathcal{P}(\mathbf{Y}_i - \mu_i\nabla f(\mathbf{Y}_i)), \quad (7)$$

Algorithm 3 Efficient implementation of SVP, $\mathcal{K} = \{\mathbb{R}, \mathbb{C}\}$

Require: step-size $\mu > 0$, measurements $\mathbf{y}$, initial points $u_0 \in \mathcal{K}^{m\times r}$, $v_0 \in \mathcal{K}^{n\times r}$, $d_0 \in \mathcal{K}^r$, (opt.) unitarily invariant convex set $\mathcal{C}$

Require: Function $\mathtt{A} : (u, d, v) \mapsto \boldsymbol{\mathcal{A}}(u\,\mathrm{diag}(d)v^*)$

Require: Function $\mathtt{At} : (\mathbf{z}, w) \mapsto \boldsymbol{\mathcal{A}}^*(\mathbf{z})w$

Require: Function $\mathtt{At}^* : (\mathbf{z}, w) \mapsto (\boldsymbol{\mathcal{A}}^*(\mathbf{z}))^*w$

$v_{-1} \leftarrow 0$, $u_{-1} \leftarrow 0$, $d_{-1} \leftarrow 0$

for $i = 0, 1, \ldots$ **do**

  Compute β_i // See text

  $u_y \leftarrow [u_i, u_{i-1}]$, $v_y \leftarrow [v_i, v_{i-1}]$

  $d_y \leftarrow [(1+\beta_i)d_i, -\beta_i d_{i-1}]$

  $\mathbf{z} \leftarrow A(u_y, d_y, v_y) - \mathbf{y}$ // Compute the residual

  Define the functions

  $\mathtt{h} : w \mapsto u_y\,\mathrm{diag}(d_y)v_y^*w - \mu\mathtt{At}(\mathbf{z}, w)$

  $\mathtt{h}^* : w \mapsto v_y\,\mathrm{diag}(d_y)u_y^*w - \mu\mathtt{At}^*(\mathbf{z}, w)$

  $(u_{i+1}, d_{i+1}, v_{i+1}) \leftarrow \mathtt{RandomizedSVD}(\mathtt{h}, \mathtt{h}^*, r)$

  $d_{i+1} \leftarrow \mathcal{P}_{\mathcal{C}}(d_{i+1})$ // Optional

end for

return $X \leftarrow u_i d_i v_i^*$ // If desired

where β_i is chosen $\beta_i = (\alpha_{i-1} - 1)/\alpha_i$ and $\alpha_0 = 1$, $2\alpha_{i+1} = 1 + \sqrt{4\alpha_i^2 + 1}$ [15] (see [2]). Algorithm 3 shows details for low-memory implementation. The implementation of `A` and `At` depends on the structure of $\boldsymbol{\mathcal{A}}$ in the specific problem.

IV. Convergence

We assume the observations are generated by $\mathbf{y} = \boldsymbol{\mathcal{A}}\mathbf{X}^\star + \boldsymbol{\varepsilon}$ where $\boldsymbol{\varepsilon}$ is a noise term (not to be confused with ϵ). In the following theorem, we will assume that $\|\boldsymbol{\mathcal{A}}\|^2 \le mn/p$, which is true for quantum tomography [16]; if $\boldsymbol{\mathcal{A}}$ is a normalized Gaussian, then this assumption holds in expectation.

Theorem 2. *(Iteration invariant) Pick an accuracy $\epsilon = \frac{r}{\rho - 1}$, where ρ is defined as in Theorem 1. Define $\ell = r + \rho$ and let c be an integer such that $\ell = (c-1)r$. Let $\mu_i = \frac{1}{2(1+\delta_{cr})}$ in* (4) *and assume $\|\boldsymbol{\mathcal{A}}\|^2 \le mn/p$ and $f(\mathbf{X}_i) > C^2\|\boldsymbol{\varepsilon}\|^2$, where $C \ge 4$ is a constant. Then the descent scheme* (4) *or* (5) *has the following iteration invariant*

$$\mathbb{E}f(\mathbf{X}_{i+1}) \le \theta f(\mathbf{X}_i) + \tau\|\boldsymbol{\varepsilon}\|^2, \tag{8}$$

in expectation, where

$$\theta \le 12 \cdot \frac{1+\delta_{2r}}{1-\delta_{cr}} \cdot \left(\frac{\epsilon}{1+\delta_{cr}} \cdot \frac{mn}{p} + (1+\epsilon)\frac{3\delta_{cr}}{1-\delta_{2r}} \right),$$

and

$$\tau \le \frac{1+\delta_{2r}}{1-\delta_{cr}} \cdot \left(12 \cdot (1+\epsilon)\left(1 + \frac{2\delta_{cr}}{1-\delta_{2r}}\right) + 8 \right).$$

The expectation is taken with respect to Gaussian random designs in `RandomizedSVD`. If $\theta \le \theta_\infty < 1$ for all iterations, then $\lim_{i\to\infty} \mathbb{E}f(\mathbf{X}_i) \le \max\{C^2, \frac{\tau}{1-\theta_\infty}\}\|\boldsymbol{\varepsilon}\|^2$.

Each call to `RandomizedSVD` draws a new Gaussian r.v., so the expected value does not depend on previous iterations.

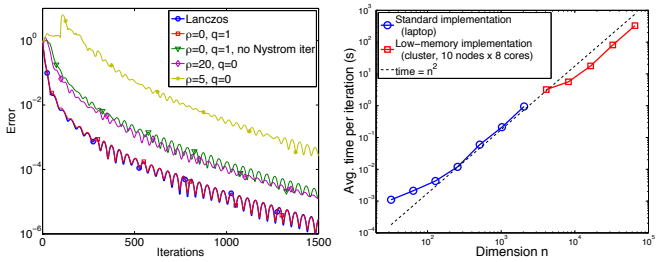


Fig. 1. (Left) Convergence rate as a function of parameters to `RandomizedSVD`/`RandomizedEIG`. (Right) Scaling plot for computation time of `RandomizedEIG`.

The expected value of the function converges linearly at rate θ to within a constant of the noise level, and in particular, it converges to zero when there is no noise.

Unfortunately, the theorem imposes overly pessimistic values for ϵ. The bound on θ should be less than 1 in order to have a contraction. This gives the requirement that $\delta_{cr} \lesssim 1/200$, which is reasonable (cf. [12]). However, it also imposes[2] $\frac{12}{1-\delta_{cr}^2} \cdot \frac{\epsilon\, mn}{p} < \frac{1}{2}$, which means that we need $\epsilon \lesssim \frac{p}{24mn}$. For quantum tomography, $m = n$ and $p = \mathcal{O}(rn)$, so we require $\varepsilon \lesssim r/n$. From Theorem 1, our bound on ϵ is $r/(\rho - 1)$, so we require $\rho \simeq n$, which defeats the purpose of the randomized algorithm (in this case, one would just perform a dense SVD). Surprisingly, numerical examples in the next section show that ρ can be nearly a small constant, so the theory is not sharp.

V. Numerical experiments

We apply the algorithm to the quantum tomography problem, which is a particular instance of (1). For details, we refer to [5]. The salient features are that the variable $\mathbf{X} \in \mathbb{C}^{n\times n}$ is constrained to be Hermitian positive-definite, and that, unlike many low-rank recovery problems, the linear operator $\boldsymbol{\mathcal{A}}$ satisfies the R-RIP: [16] establishes that Pauli measurements (which comprise $\boldsymbol{\mathcal{A}}$) have R-RIP with overwhelming probability when $p = \mathcal{O}(rn\log^6 n)$. In the ideal case, $\mathbf{X}^\star$ is exactly rank 1, but it may have larger rank due to some (non-Gaussian) noise processes, in addition to AWGN $\boldsymbol{\varepsilon}$. Furthermore, it is known that the true solution $\mathbf{X}^\star$ has trace 1, which is also possible to exploit in our algorithmic framework. Each component of the linear operator $\boldsymbol{\mathcal{A}}$ has a special Kronecker product structure, which we exploit in order to keep memory low, using custom parallel code.

Figure 1 (left) plots convergence and accuracy results for a quantum tomography problem with 8 qubits and $p = 4rn$ with $r = 1$. The SVP algorithm works well on noisy problems but we focus here on a noiseless (and truly low-rank) problem in order to examine the effects of approximate SVD/eigenvalue computations. The figure shows that the power method with $q \ge 1$ is extremely effective (if $q = 0$, then $\rho \simeq 20$ still leads to convergence). When p is smaller and the R-RIP is not satisfied, taking ρ or q too small can lead to divergence.

[2]For the details, see the extended version at http://arxiv.org/abs/1303.0167.

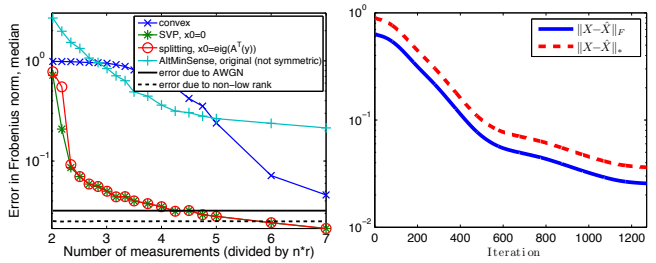


Fig. 2. (Left) Accuracy comparison of several algorithms, as a function of number of samples. (Right) Convergence for the 16-qubit simulation

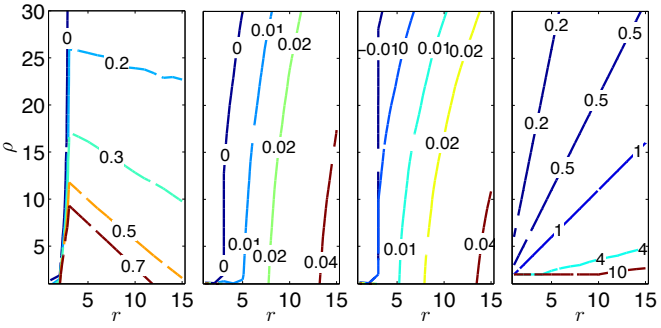


Fig. 3. Left 3 plots are the empirical estimates $\widetilde{\epsilon}$ for $q = 0, 1, 2$ power iterations. Rightmost plot is the theoretical values ϵ from Thm. 1

The right subfigure of Figure 1 shows that the low-memory implementation has time complexity $\mathcal{O}(n^2)$ up to $n = 2^{16}$.

The left subfigure of Figure 2 reports the median error on 10 test problems across a range of p. Here, $\mathbf{X}^\star$ is only approximately low rank and y is contaminated with noise. We compare the convex approach [5], the "AltMinSense" approach [12], and a standard splitting approach. AltMinSense and the convex approach have poor accuracy; the accuracy of AltMinSense can be improved by incorporating symmetry, but this changes the algorithm fundamentally and the theoretical guarantees are lost. The splitting approach, if initialized correctly, is accurate, but lacks guarantees. Furthermore, it is slower in practice due to slower convergence, though for some simple problems it is possible to accelerate using L-BFGS [10].

Figure 3 tests Theorem 1 by plotting the value of

$$\widetilde{\epsilon} = \|\mathbf{X} - \widetilde{\mathbf{X}}\|_F^2 / \|\mathbf{X} - \mathbf{X}_r\|_F^2 - 1$$

(which is bounded by ϵ) for matrices $\mathbf{X}$ that are generated by the iterates of the algorithm. The algorithm is set for $r = 1$ (so $\mathbf{X}$ is the sum of a rank 2 term, which includes the Nesterov term, and the full rank gradient), but the plots consider a range of r and a range of oversampling parameters ρ. Because $\widetilde{\mathbf{X}}$ has rank $\ell = r + \rho$, it is possible for $\widetilde{\epsilon} < 0$, as we observe in the plots when r is small and ρ is large. For two power iterations, the error is excellent. In all cases, the observed error $\widetilde{\epsilon}$ is much better than the bound ϵ from Theorem 1.

Finally, to test scaling to very large data, we compute a 16 qubit state ($n = 65536$), using a known quantum state as input, with realistic quantum mechanical perturbations (global depolarizing noise of level $\gamma = 0.01$; see [5]) as well as AWGN to give a SNR of 30 dB, and $p = 5n = 327680$ measurements. The first iteration uses Lanczos and all subsequent iterations use `RandomizedEIG` using $\rho = 5$ and $q = 3$ power iterations. On a cluster with 10 computers, the mean time per iteration is 401 seconds. After 1270 iterations, $\|\mathbf{X} - \mathbf{X}^\star\|_F = 0.0256$; see Figure 2 (right).

VI. Conclusion

Randomization is a powerful tool to accelerate and scale optimization algorithms, and it can be rigorously included in algorithms that are robust to small errors. In this paper, we leverage randomized approximations to remove memory bottlenecks by merging the two-key steps of most recovery algorithms in affine rank minimization problems: gradient calculation and low-rank projection. Unfortunately, the current black-box approximation guarantees, such as Theorem 1, are too pessimistic to be directly used in theoretical characterizations of our approach. For future work, motivated by the overwhelming empirical evidence of the good performance of our approach, we plan to directly analyze the impact of randomization in characterizing the algorithmic performance.

Acknowledgment: VC and AK's work was supported by MIRG-268398, ERC Future Proof, and SNF 200021-132548. SRB is supported by the Fondation Sciences Mathématiques de Paris. The authors thank Alex Gittens for his insightful comments.

References

[1] R. Meka, P. Jain, and I. S. Dhillon, "Guaranteed rank minimization via singular value projection," in *NIPS*, 2010.
[2] A. Kyrillidis and V. Cevher, "Matrix recipes for hard thresholding methods," *arXiv preprint arXiv:1203.4481*, 2012.
[3] B. Recht, M. Fazel, and P. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM review*, vol. 52, no. 3, pp. 471–501, 2010.
[4] E. Candes and B. Recht, "Exact matrix completion via convex optimization," *Found. Comput. Math.*, vol. 9, pp. 717–772, 2009.
[5] S. Flammia, D. Gross, Y. Liu, and J. Eisert, "Quantum tomography via compressed sensing: error bounds, sample complexity, and efficient estimators," *New J. Phys.*, vol. 14, no. 9, p. 095022, 2012.
[6] N. Halko, P. G. Martinsson, and J. A. Tropp, "Finding structure with randomness: Stochastic algorithms for constructing approximate matrix decompositions," *SIAM Rev.*, vol. 53, no. 2, pp. 217–288, 2011.
[7] Z. Wen, W. Yin, and Y. Zhang, "Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm," *Mathematical Programming Computation*, pp. 1–29, 2010.
[8] B. Recht and C. Ré, "Parallel stochastic gradient algorithms for large-scale matrix completion," *Math. Prog. Comput.*, to appear, 2013.
[9] J. Lee, B. Recht, R. Salakhutdinov, N. Srebro, and J. A. Tropp, "Practical large-scale optimization for max-norm regularization," in *NIPS*, 2011.
[10] S. Laue, "A hybrid algorithm for convex semidefinite optimization," in *ICML*, 2012.
[11] S. Burer and R. Monteiro, "A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization," *Math. Prog. (series B)*, vol. 95, no. 2, pp. 329–357, 2003.
[12] P. Jain, P. Netrapalli, and S. Sanghavi, "Low-rank matrix completion using alternating minimization," in *ACM Symp. Theory Comput.*, 2012.
[13] D. Goldfarb and S. Ma, "Convergence of fixed-point continuation algorithms for matrix rank minimization," *Foundations of Computational Mathematics*, vol. 11, no. 2, pp. 183–210, 2011.
[14] S. Becker, V. Cevher, C. Koch, and A. Kyrillidis, "Sparse projections onto the simplex," in *ICML*, to appear, 2013.
[15] Y. Nesterov, "A method for unconstrained convex minimization problem with the rate of convergence $\mathcal{O}(1/k^2)$," *Doklady AN SSSR*, translated as *Soviet Math. Docl.*, vol. 269, pp. 543–547, 1983.
[16] Y. K. Liu, "Universal low-rank matrix recovery from Pauli measurements," in *NIPS*, 2011, pp. 1638–1646.

On Sparsity Averaging

Rafael E. Carrillo*, Jason D. McEwen†, and Yves Wiaux*‡§
* Institute of Electrical Engineering, Ecole Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland.
† Department of Physics and Astronomy, University College London, London WC1E 6BT, UK.
‡ Department of Radiology and Medical Informatics, University of Geneva (UniGE), CH-1211 Geneva, Switzerland.
§ Department of Radiology, Lausanne University Hospital (CHUV), CH-1011 Lausanne, Switzerland.

***Abstract*—Recent developments in [1] and [2] introduced a novel regularization method for compressive imaging in the context of compressed sensing with coherent redundant dictionaries. The approach relies on the observation that natural images exhibit strong *average sparsity* over multiple coherent frames. The associated reconstruction algorithm, based on an *analysis* prior and a *reweighted* ℓ_1 scheme, is dubbed Sparsity Averaging Reweighted Analysis (SARA). We review these advances and extend associated simulations establishing the superiority of SARA to regularization methods based on sparsity in a single frame, for a generic spread spectrum acquisition and for a Fourier acquisition of particular interest in radio astronomy.**

I. Introduction

Consider a complex-valued signal $\boldsymbol{x} \in \mathbb{C}^N$, assumed to be sparse in some orthonormal basis $\Psi \in \mathbb{C}^{N\times N}$, and also consider the measurement model $\boldsymbol{y} = \Phi\boldsymbol{x}+\boldsymbol{n}$, where $\boldsymbol{y} \in \mathbb{C}^M$ denotes the measurement vector, $\Phi \in \mathbb{C}^{M\times N}$ with $M < N$ is the sensing matrix and $\boldsymbol{n} \in \mathbb{C}^M$ represents the observation noise. The most common approach in compressed sensing (CS) is to recover $\boldsymbol{x}$ from $\boldsymbol{y}$ solving the following convex problem [3]:

$$\min_{\bar{\boldsymbol{\alpha}}\in\mathbb{C}^N} \|\bar{\boldsymbol{\alpha}}\|_1 \text{ subject to } \|\boldsymbol{y} - \Phi\Psi\bar{\boldsymbol{\alpha}}\|_2 \leq \epsilon, \tag{1}$$

where ϵ is an upper bound on the ℓ_2 norm of the noise and $\|\cdot\|_1$ denotes the ℓ_1 norm of a complex-valued vector. The signal is recovered as $\hat{\boldsymbol{x}} = \Psi\hat{\boldsymbol{\alpha}}$, where $\hat{\boldsymbol{\alpha}}$ denotes the solution to (1). Such problems that solve for the representation of the signal in a sparsity basis are known as synthesis-based problems. The standard CS theory provides results for the recovery of $\boldsymbol{x}$ from $\boldsymbol{y}$ if Φ obeys a Restricted Isometry Property (RIP) and Ψ is orthonormal [3]. However, signals often exhibit better sparsity in an overcomplete dictionary [4]–[6].

Recent works have begun to address the case of CS with redundant dictionaries. In this setting the signal $\boldsymbol{x}$ is expressed in terms of a dictionary $\Psi \in \mathbb{C}^{N\times D}$, $N < D$, as $\boldsymbol{x} = \Psi\boldsymbol{\alpha}$, $\boldsymbol{\alpha} \in \mathbb{C}^D$. Rauhut et al. [7] find conditions on the dictionary Ψ such that the compound matrix $\Phi\Psi$ obeys the RIP to accurately recover $\boldsymbol{\alpha}$ by solving a synthesis-based problem. Candès et al. [8] provide a theoretical analysis of the ℓ_1 analysis-based problem. As opposed to synthesis-based problems, analysis-based problems recover the signal itself solving:

$$\min_{\bar{\boldsymbol{x}}\in\mathbb{C}^N} \|\Psi^\dagger\bar{\boldsymbol{x}}\|_1 \text{ subject to } \|\boldsymbol{y} - \Phi\bar{\boldsymbol{x}}\|_2 \leq \epsilon, \tag{2}$$

where $\Psi^\dagger$ denotes the adjoint operator of Ψ. The aforementioned work [8] extends the standard CS theory to coherent and redundant dictionaries, providing theoretical stability guarantees based on a general condition of the sensing matrix Φ, coined the Dictionary Restricted Isometry Property (D-RIP).

In [1] and [2], we proposed a novel sparsity analysis prior for compressive imaging in the context of CS with coherent and redundant dictionaries, relying on the observation that natural images are simultaneously sparse in various frames, in particular wavelet frames, or in their gradient. Promoting *average sparsity* over multiple frames, as opposed to single frame sparsity, is an extremely powerful prior. The associated reconstruction algorithm, based on an *analysis* prior and a *reweighted* ℓ_1 scheme, is dubbed Sparsity Averaging Reweighted Analysis (SARA)[1].

In this work, we review and further discuss these recent advances. The superiority of SARA to regularization methods based on sparsity in a single frame, as established through simulations for a generic spread spectrum acquisition, is described with an additional extensive visual support. Moreover, we bring a novel illustration for a realistic continuous Fourier sampling strategy of particular interest for radio interferometry in astronomy. We finally discuss possible avenues to establish explicit theoretical stability results for the algorithm.

II. Sparsity Averaging Reweighted Analysis

Natural images are often complicated and encompass several types of structures admitting sparse representations in different frames. For example, piecewise smooth structures exhibit gradient sparsity, while extended structures are better encapsulated in wavelet frames. Observing that natural images actually exhibit sparsity in multiple frames, we hypothesise in [1] and [2] that average sparsity over multiple coherent frames represents a strong prior. We thus proposed the use of a dictionary composed of a concatenation of q frames, i.e.

$$\Psi = \frac{1}{\sqrt{q}}[\Psi_1, \Psi_2, \ldots, \Psi_q], \tag{3}$$

with $\Psi \in \mathbb{C}^{N\times D}$, $N < D$, and an analysis ℓ_0 prior,

$$\|\Psi^\dagger\bar{\boldsymbol{x}}\|_0 \sim \frac{1}{q}\sum_{i=1}^{q} \|\Psi_i^\dagger\bar{\boldsymbol{x}}\|_0, \tag{4}$$

to promote this average sparsity. Note that in this setting each frame contains all the signal information as opposed to

[1] In [9], similar ideas were applied to the reverberant audio source separation problem exploiting sparsity in a redundant short time Fourier transform.

component separation approaches such as [4] and [5]. Also note on a theoretical level that a single signal cannot be arbitrarily sparse simultaneously in a set of incoherent frames. For example, a signal extremely sparse in the Dirac basis is completely spread in the Fourier basis. As discussed in [2], each frame, Ψ_i, should be highly coherent with the other frames in order for the signal to have a sparse representation in Ψ. Concatenation of the first eight orthonormal Daubechies wavelet bases (Db1-Db8) is an example of interest. The first Daubechies wavelet basis, Db1, is the Haar wavelet basis. It can be used as an alternative to gradient sparsity, usually imposed by a total variation (TV) prior, to promote piecewise smooth signals. The Db2-Db8 bases provide smoother decompositions. Coherence between the bases is ensured by the compact support of the Daubechies wavelets.

A reweighted ℓ_1 minimization scheme [10] promotes average sparsity through the prior (4). The algorithm replaces the ℓ_0 norm by a weighted ℓ_1 norm and solves a sequence of weighted ℓ_1 problems with weights essentially the inverse of the values of the solution of the previous problem:

$$\min_{\bar{\boldsymbol{x}} \in \mathbb{C}^N} \|\mathsf{W}\Psi^\dagger \bar{\boldsymbol{x}}\|_1 \text{ subject to } \|\boldsymbol{y} - \Phi\bar{\boldsymbol{x}}\|_2 \leq \epsilon, \qquad (5)$$

where $\mathsf{W} \in \mathbb{R}^{D\times D}$ is a diagonal matrix with positive weights. The solution to (5) is denoted as $\Delta(\boldsymbol{y}, \Phi, \mathsf{W}, \epsilon)$. We update the weights at each iteration, i.e. after solving a complete weighted ℓ_1 problem, by the function $f(\gamma, a) \propto (\gamma + |a|)^{-1}$, where a denotes the coefficient value estimated at the previous iteration and γ plays the role of a stabilization parameter, avoiding undefined weights when the signal value is zero. Note that as $\gamma \to 0$ the solution of the weighted ℓ_1 problem approaches the solution of the ℓ_0 problem. We use a homotopy strategy and solve a sequence of weighted ℓ_1 problems with a decreasing sequence $\{\gamma^{(t)}\}$, with t denoting the iteration time variable.

The sparsity averaging reweighted analysis (SARA) algorithm is defined in Algorithm 1, with Ψ defined as in (3). A rate parameter $\beta \in (0, 1)$ controls the decrease of the sequence through $\gamma^{(t)} = \beta\gamma^{(t-1)}$. However, the noise standard deviation σ_α in the representation domain, rough estimate for a baseline above which significant signal components could be identified, serves as a lower bound: $\gamma^{(t)} \geq \sigma_\alpha = \sqrt{M/D}\sigma_n$, with σ_n the noise standard deviation in measurement space. As a starting point we set $\hat{\boldsymbol{x}}^{(0)}$ as the solution of the ℓ_1 problem and $\gamma^{(0)} = \sigma_s\left(\Psi^\dagger\hat{\boldsymbol{x}}^{(0)}\right)$, where $\sigma_s(\cdot)$ takes the empirical standard deviation of a signal. The re-weighting process ideally stops when the relative variation between successive solutions is smaller than some bound $\eta \in (0, 1)$, or after the maximum number of iterations allowed, $N_{\max}$, is reached. We fix $\eta = 10^{-3}$ and $\beta = 10^{-1}$.

Algorithm 1 SARA algorithm

Input: $\boldsymbol{y}$, Φ, ϵ, σ_α, β, η and $N_{\max}$.
Output: Reconstructed image $\hat{\boldsymbol{x}}$.

Initialize $t = 1$, $\mathsf{W}^{(0)} = \mathsf{I}$ and $\rho = 1$.
Compute $\hat{\boldsymbol{x}}^{(0)} = \Delta(\boldsymbol{y}, \Phi, \mathsf{W}^{(0)}, \epsilon)$, $\gamma^{(0)} = \sigma_s\left(\Psi^\dagger\hat{\boldsymbol{x}}^{(0)}\right)$.
while $\rho > \eta$ and $t < N_{\max}$ **do**
  Update $\mathsf{W}_{ij}^{(t)} = f\left(\gamma^{(t-1)}, \hat{\alpha}_i^{(t-1)}\right)\delta_{ij}$, for $i, j = 1, \ldots, D$ with $\hat{\boldsymbol{\alpha}}^{(t-1)} = \Psi^\dagger\hat{\boldsymbol{x}}^{(t-1)}$.
  Compute a solution $\hat{\boldsymbol{x}}^{(t)} = \Delta(\boldsymbol{y}, \Phi, \mathsf{W}^{(t)}, \epsilon)$.
  Update $\gamma^{(t)} = \max\{\beta\gamma^{(t-1)}, \sigma_\alpha\}$.
  Update $\rho = \|\hat{\boldsymbol{x}}^{(t)} - \hat{\boldsymbol{x}}^{(t-1)}\|_2 / \|\hat{\boldsymbol{x}}^{(t-1)}\|_2$.
  $t \leftarrow t + 1$
end while

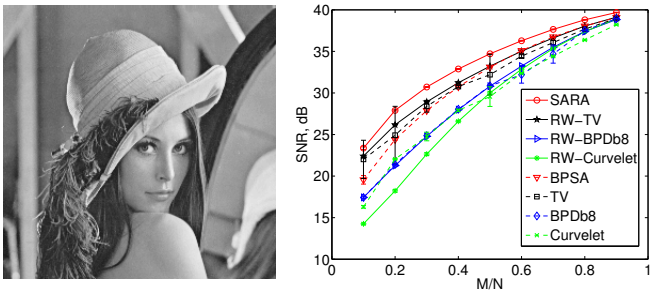


Figure 1. Reconstruction quality results for Lena in the context of a spread spectrum acquisition. Left: original image. Right: SNR results against the undersampling ratio for an input SNR of 30 dB (average values over 100 simulations are shown with corresponding standard deviations).

III. Simulations

In this section, the superiority of SARA to regularization methods based on sparsity in a single frame, as established through simulations in the context of a generic spread spectrum acquisition, is described with a new extensive visual support. Moreover, we bring a novel illustration for a realistic continuous Fourier sampling strategy of particular interest for radio interferometry.

For the first experiment we recover a 256×256 version of Lena from compressive measurements. The spread spectrum technique described in [11] is used as measurement operator. We compare SARA to analogous analysis algorithms, and their reweighted versions, changing the sparsity dictionary Ψ in (2) and (5) respectively. Three different dictionaries are considered: the Daubechies 8 wavelet basis, the redundant curvelet frame [6] and the concatenation of the first eight Daubechies bases described above for SARA. The associated algorithms are respectively denoted BPDb8, Curvelet and BPSA for the non reweighted case. The reweighted versions are respectively denoted RW-BPDb8, RW-Curvelet and SARA. Additionally, we also compare to the TV prior, where the TV minimization problem is formulated as a constrained problem like (2), but replacing the ℓ_1 norm by the image TV norm. The reweighted version of TV is denoted as RW-TV. Since the image of interest is positive, we impose the additional constraint that $\bar{\boldsymbol{x}} \in \mathbb{R}_+^N$ for all problems. The reconstruction quality of SARA is evaluated as a function of the undersampling ratio M/N, for M/N in the range $[0.1, 0.9]$. The input SNR is set to 30 dB. The SNR results comparing SARA against all the other benchmark methods are shown in the right panel of Figure 1. The results demonstrate that SARA outperforms state-of-the-

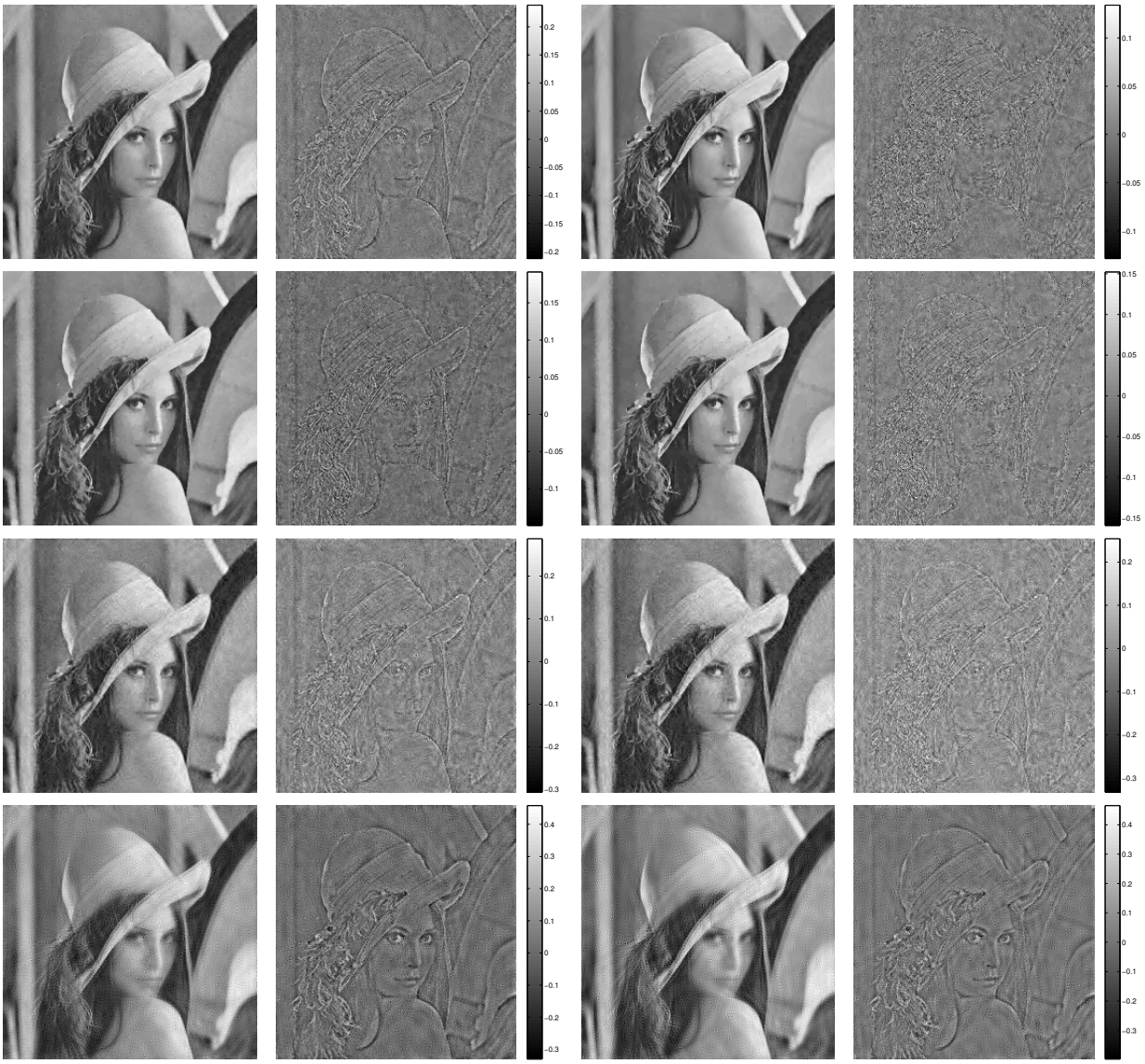


Figure 2. Reconstruction example of Lena for spread spectrum acquisition, with $M = 0.2N$ and input SNR set to 30 dB. First and third columns show the reconstructed images and the second and fourth columns show the error images. First row: BPSA(24.4 dB) and SARA (27.9 dB). Second row: TV(26.3 dB) and RW-TV (26.6 dB). Third row: BPDb8 (21.4 dB) and RW-BPDb8 (21.2 dB). Fourth row: Curvelet (18.7 dB) and RW-Curvelet (18.3 dB).

art methods for all undersampling ratios. RW-TV provides the second best results. BPSA achieves better SNRs than BPDb8, curvelet and their reweighted versions for all undersampling ratios. It also achieves similar SNRs to TV in the range 0.4-0.9. Figure 2 presents a visual assessment for $M = 0.2N$, showing both reconstructed and error images. SARA provides an impressive reduction of visual artifacts relative to the other methods in this high undersampling regime. In particular RW-TV exhibits expected cartoon-like artifacts. Other methods do not yield results of comparable quality, either in SNR or visually, with associated reconstructions full of visual artifacts.

The second experiment illustrates the performance of SARA in the context of radio interferometric imaging by recovering a 256×256 version of the well known M31 galaxy from simulated continuous Fourier samples associated with a realistic radio telescope sampling pattern (superposition of arcs of ellipses). The number of measurements is $M = 9413$, affected by 30 dB of input noise. The dictionary for SARA is the concatenation of the first eight Daubechies bases *and* the Dirac basis. The Dirac basis is added given the sparsity in image space due to the large field of view. For comparison, we use two different methods: BP, constrained ℓ_1-minimization in the Dirac basis (used as benchmark in the field), and BPDb8, constrained analysis-based ℓ_1-minimization in the Db8 basis. Figure 3 shows the original test image, the sampling pattern and the corresponding dirty image, i.e. the inverse Fourier transform of the measurements, with non-measured points set to zero. The reconstructed images for BP, BPDb8 and SARA are also reported. Once more, SARA provides not only a drastic SNR increase but also a significant reduction of visual

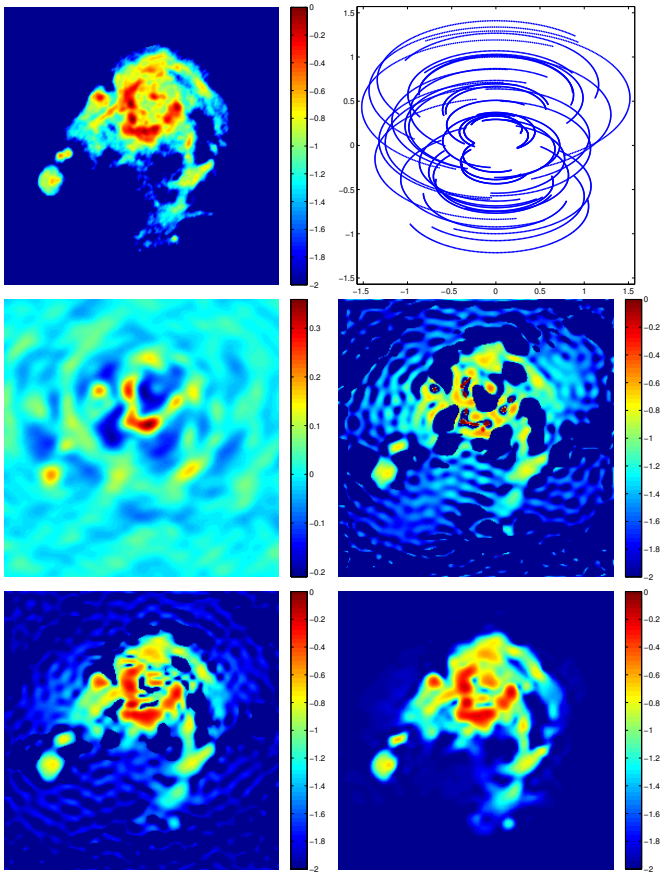


Figure 3. Radio Astronomy example. From left to right. Top row: original test image in $\log_{10}$ scale and Fourier sampling pattern. Middle row: corresponding dirty image in linear scale and reconstruction results for BP (3.9 dB) in $\log_{10}$ scale. Bottom row: reconstruction results for BPDb8 (10.3 dB) and SARA (14.1 dB) in $\log_{10}$ scale.

artifacts relative to the other methods.

IV. Conclusion and Discussion

In this paper we have reviewed recent advances in the average sparsity model and the associated algorithm SARA. Extended simulations demonstrating the superiority of SARA for compressive imaging reconstruction were described. Novel results on the application of SARA to a realistic radio interferometric imaging scenario were also described.

Future work will concentrate on finding a theoretical framework for the average sparsity model. In [2] we have put average sparsity in the context of theory developed in [8]. However, specialized results for the particular case of concatenation of frames (or orthogonal bases) are needed. The co-sparsity analysis model [12] proposes a general framework for general analysis operators. Similar properties to the D-RIP coined Ω-RIP are introduced in [13] to analyze greedy algorithms in the context of the co-sparsity analysis model. It would be interesting to explore the connections between average sparsity and the co-sparsity model to have an estimate on the number of measurements needed for reconstruction compared to single frame representations.

The proposed approach relies on the observation that natural images exhibit strong average sparsity, i.e. the signals of interest have so-called simultaneous structured models. Recently, it was shown in [14] that combinations of convex relaxations of the individual structured models do not yield better results than an algorithm that exploits only one of the structured models, while *non-convex* approaches that approximate the simultaneous model can exploit the multiple structured models. Those results suggest that the *re-weighting* approach in SARA to approximate the ℓ_0 norm is fundamental to exploit average sparsity, as observed in the simulation results (see the gap between SARA and BPSA in Fig. 1 and Fig. 2).

Acknowledgment

REC is supported by the Swiss National Science Foundation (SNSF) under grant 200021-130359. JDM is supported by a Newton International Fellowship from the Royal Society and the British Academy. YW is supported by the Center for Biomedical Imaging (CIBM) of the Geneva and Lausanne Universities and EPFL.

References

[1] R. E. Carrillo, J. D. McEwen, and Y. Wiaux, "Sparsity averaging reweighted analysis (SARA): a novel algorithm for radio-interferometric imaging," *Monthly Notices of the Royal Astronomical Society*, vol. 426, no. 2, pp. 1223–1234, 2012.

[2] R. E. Carrillo, J. D. McEwen, D. V. D. Ville, J.-P. Thiran, and Y. Wiaux, "Sparsity averaging for compressive imaging," *IEEE Signal Processing Letters*, vol. 20, no. 6, pp. 591–594, 2013.

[3] M. Fornasier and H. Rauhut, *Handbook of Mathematical Methods in Imaging*. Springer, 2011, ch. Compressed sensing.

[4] R. Gribonval and M. Nielsen, "Sparse representations in unions of bases," *IEEE Transactions on Information Theory*, vol. 49, no. 12, pp. 3320–3325, Dec. 2003.

[5] J. Bobin, J.-L. Starck, J. Fadili, Y. Moudden, and D. Donoho, "Morphological component analysis: an adaptative thresholding strategy," *IEEE Transactions on Image Processing*, vol. 16, no. 11, pp. 2675–2681, 2007.

[6] J. Starck, F. Murtagh, and J. Fadili, *Sparse Image and Signal Processing: Wavelets, Curvelets, Morphological Diversity*. Cambridge University Press, Cambridge, GB, 2010.

[7] H. Rauhut, K. Schnass, and P. Vandergheynst, "Compressed sensing and redundant dictionaries," *IEEE Transactions on Information Theory*, vol. 54, no. 5, pp. 2210–2219, May 2008.

[8] E. J. Candès, Y. Eldar, D. Needell, and P. Randall, "Compressed sensing with coherent and redundant dictionaries," *Applied and Computational Harmonic Analysis*, vol. 31, no. 1, pp. 59–73, 2010.

[9] S. Arberet, P. Vandergheynst, R. E. Carrillo, J.-P. Thiran, and Y. Wiaux, "Sparse reverberant audio source separation via reweighted analysis," *IEEE Transactions on Audio Speech and Language Processing*, vol. 21, no. 7, pp. 1391–1402, 2013.

[10] E. J. Candès, M. Wakin, and S. Boyd, "Enhacing sparsity by reweighted ℓ_1 minimization," *Journal of Fourier Analysis and Applications*, vol. 14, no. 5, pp. 877–905, 2008.

[11] G. Puy, P. Vandergheynst, R. Gribonval, and Y. Wiaux, "Universal and efficient compressed sensing by spread spectrum and application to realistic fourier imaging techniques," *EURASIP Journal on Applied signal Processing*, vol. 2012, no. 3, 2012.

[12] S. Nam, M. Davies, R.Gribonval, and M. Elad, "The cosparse analysis model and algorithms," *Applied and Computational Harmonic Analysis*, vol. 34, no. 1, pp. 30–56, 2013.

[13] R. Giryes, S. Nam, M. Elad, R.Gribonval, and M. Davies, "Greedy-like algorithms for the cosparse analysis model," 2013, preprint, arXiv:1207.2456v2.

[14] S. Oymak, A. Jalali, M. Fazel, Y. Eldar, and B. Hassibi, "Simultaneously structured models with application to sparse and low-rank matrices," 2013, preprint, arXiv:1212.3753v2.

Conditions for Dual Certificate Existence in Semidefinite Rank-1 Matrix Recovery

Paul Hand
Department of Mathematics
Massachusetts Institute of Technology
Cambridge, MA 02139
Email: hand@mit.edu

***Abstract*—We study the existence of dual certificates in convex minimization problems where a rank-1 matrix X_0 is to be recovered under semidefinite and linear constraints. We provide an example where such a dual certificate does not exist. We prove that dual certificates are guaranteed to exist if the linear measurement matrices can not be recombined to form something positive and orthogonal to X_0. If the measurements can be recombined in this way, the problem is equivalent to one with additional linear constraints. That augmented problem is guaranteed to have a dual certificate at the minimizer, providing the form of an optimality certificate for the original problem.**

I. Introduction

We consider the problem of showing that $X_0 = x_0 x_0^*$ is a minimizer to the semidefinite program

$$\min f(X) \text{ subject to } X \succeq 0, \mathcal{A}(X) = b. \tag{1}$$

for $x_0 \in \mathbb{R}^n$, $X \in \mathcal{S}_n$ is a symmetric real $n \times n$ matrix, f is convex and continuous everywhere, and $\mathcal{A}$ is linear, and $\mathcal{A}(X_0) = b \in \mathbb{R}^n$. Let $\langle X, Y \rangle = \text{tr}(Y^* X)$ be the Hilbert-Schmidt inner product. Matrix orthogonality is understood to be with respect to this inner product. The linear measurements $\mathcal{A}(X) = b$ can be written as

$$\mathcal{A}(X)_i = \langle X, A_i \rangle = b_i \text{ for } i = 1, \ldots, m$$

for certain symmetric matrices A_i. Note that the adjoint of $\mathcal{A}$ is given by $\mathcal{A}^* \lambda = \sum_i \lambda_i A_i$.

One problem of this form is phase retrieval via PhaseLift, where $f(X) = \text{tr}(X)$ and $A_i = z_i z_i^*$ for vectors z_i [3]. Another example is the corresponding sparse recovery problem with $f(X) = \|X\|_1 + c \,\text{tr}(X)$, where the first term is the entry-wise ℓ^1 norm of X [6].

In these matrix recovery problems, a recovery result that X_0 is a minimizer is often proved by constructing a dual certificate (or approximation thereof) at X_0. Similar to [5] and [2], we call $Y \in \mathcal{S}_n$ a dual certificate at X_0 if

$$\begin{cases} Y = \mathcal{A}^* \lambda + Q \in -\partial f(X_0) \\ Q \preceq 0 \\ Q \perp X_0. \end{cases} \tag{2}$$

If a dual certificate exists at X_0 then X_0 is a minimizer of (1). Further, it is straightforward to prove that existence of a dual certificate at X_0 is equivalent to (1) satisfying strong duality with dual attainment by (λ, Q).

In the development of convex programs for matrix recovery, it is desirable to know if strong duality holds. Without guarantees of existence, attempting to analytically construct dual certificates in particular problems may be futile. Under strong duality, negative results guaranteeing that X_0 is not a minimizer could be proven by showing no dual certificate exists, as done in [6].

The perspective of this note is to ease proofs of new semidefinite relaxations, rather than easing their computation. In particular, we are concerned with conditions on A_i under which problem (1) has a dual certificate at the minimizer X_0 or can be augmented into an equivalent problem that does.

A. Counterexample

Though sufficient, existence of a dual certificate (2) is not necessary for X_0 to minimize (1). Consider the following problem:

$$\min \frac{1}{2}\|X\|^2 \text{ subject to } X \succeq 0, \left\langle X, \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \right\rangle = 0, \quad \left\langle X, \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \right\rangle = 1, \tag{3}$$

where $\|\cdot\|$ is the Frobenius norm. To analyze this problem, we recall the fact that

$$X \succeq 0 \text{ and } \langle X, qq^* \rangle = 0 \text{ for } q \in \mathbb{R}^n \Rightarrow Xq = 0 \Rightarrow \langle X, y \otimes q \rangle = 0 \text{ for any } y, \tag{4}$$

where $y \otimes q = yq^* + qy^*$ is the symmetric tensor product. Using (4), we can see that any feasible X satisfies

$$\left\langle X, \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \right\rangle = 0. \tag{5}$$

Hence, the minimizer and only feasible point of (3) is

$$X_0 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}.$$

In this example, the subdifferential of $f(X) = \frac{1}{2}\|X\|^2$ contains only the single element $\partial f(X_0) = \{X_0\}$. Again using (4), we note that the dual certificate conditions (2) can not be satisfied because there is no (Q, λ) such that

$$-\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} = \lambda_1 \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} + \lambda_2 \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} + Q.$$

for $Q \preceq 0, Q \perp X_0$. If we were to supplement (3) with the constraint (5), the conditions (2) could be satisfied for some (Q, λ).

B. Constraint Qualifications

It is well known that semidefinite programs of form (1) can have a nonzero duality gap or can have a Lagrangian dual problem for which the dual optimum is not attained [10], [12]. A constraint qualification (CQ) is a condition such that strong duality and dual attainment is ensured. For example, the presence of a strictly feasible $X \succ 0$ such that $\mathcal{A}(X) = b$, is a constraint qualification and is known as Slater's condition [1].

Slater's condition can be insufficient for low-rank matrix recovery problems. As in the counterexample, if a linear combination of the A_i are nonnegative and orthogonal to X_0, then there is no strictly feasible point. Additional constraint qualifications can be found in [11], [12].

The work in this paper will be based of the following constraint qualification. The Rockafellar-Pshenichnyi condition [4], [7], [12], [13] in the present context is that X_0 minimizes (1) if and only if there exists a $Y \in (-\partial f(X_0)) \cap \partial I_{X \succeq 0, \mathcal{A}(X)=b}(X_0)$, where $I_{X \succeq 0, \mathcal{A}(X)=b}$ is the indicator function of the feasible set. Let the cone of candidate dual certificates be

$$S := \left\{ \sum_i \lambda_i A_i + Q \mid Q \preceq 0, Q \perp X_0 \right\}, \tag{6}$$

$$= \partial I_{X \succeq 0}(X_0) + \partial I_{\mathcal{A}(X)=b}(X_0). \tag{7}$$

A constraint qualification is thus that

$$\partial I_{X \succeq 0}(X_0) + \partial I_{\mathcal{A}(X)=b}(X_0) = \partial I_{X \succeq 0, \mathcal{A}(X)=b}(X_0). \tag{8}$$

This constraint qualification is a weakest constraint qualification because it is independent of f [12].

One way to interpret this CQ is in terms of extremal directions. We say that A is an extremal direction of X_0 relative to the feasible set if $\langle A, X \rangle \leq \langle A, X_0 \rangle$ for all feasible X. Any element of S is an extremal direction of X_0, but S does not necessarily contain all directions in which X_0 is extreme. The set of all such directions is the subdifferential $\partial I_{X \succeq 0, \mathcal{A}(X)=b}(X_0)$. The CQ (8) is that S contains all directions in which X_0 is extreme. Note that $\partial I_{X \succeq 0, \mathcal{A}(X)=b}(X_0)$ is the negative polar cone of the tangent cone of the feasible set at X_0.

C. Sufficient condition for dual certificate existence

Avoiding the pathology of the counterexample, we present a condition for which dual certificates are guaranteed to exist.

Theorem 1. *Let X_0 minimize (1). If $\nexists A \in \text{span}\{A_i\}$ such that $A \succeq 0$ and $A \perp X_0$, then strong duality holds and a dual certificate exists at X_0.*

That is, the pathology of the counterexample arrives because there is a linear combination of A_i that is positive semi-definite and orthogonal to X_0. If this case is excluded, a dual certificate necessarily exists at the rank-one solution X_0.

D. Weaker sufficient condition for dual certificate existence

If there is a positive semi-definite measurement matrix A that is orthogonal to X_0, then (4) provides additional constraints on X that may or may not be implied by the linear constraints $\mathcal{A}(X) = b$ alone. For any $q \in \text{Range}(A)$, and for any y, all feasible X satisfy $\langle X, y \otimes q \rangle = 0$. Hence $y \otimes q$ is an extremal direction of X_0, and must be in S in order for strong duality to hold. We say that S is complete at X_0 if the following condition holds:

$$\begin{aligned} &\text{If } A = \mathcal{A}^* \lambda \succeq 0, A \perp X_0, \text{ then} \\ &\quad y \otimes q \in S \text{ for all } y \text{ and for all } q \in \text{Range}(A). \end{aligned} \tag{9}$$

Theorem 2. *Let X_0 minimize (1). If S satisfies the completeness condition (9) then strong duality holds and a dual certificate exists at X_0.*

E. General certificate form

As the counterexample illustrates, the problem (1) may not contain the linear equations $\langle X, y \otimes q \rangle = 0$ for the q described in section I-D. In this case, the optimality certificate for (1) can be expressed as a dual certificate for the problem augmented with linear constraints implied by $X \succeq 0$ and $\mathcal{A}(X) = b$. This augmented problem is equivalent to (1) and satisfies the conditions of Theorem 2. Hence, its dual contains the form of the optimality certificate for (1).

The following procedure outlines a process for augmenting the measurement matrices $\{A_i\}$ in such a way that there exists a dual certificate of the form $\sum_i \lambda_i A_i + Q$ for $Q \preceq 0, Q \perp X_0$:

1) Consider all $A \succeq 0, A \in \text{span}\{A_i\}, \langle A, X_0 \rangle = 0$.
2) Write each $A = \sum_k c_k q_k q_k^*$ with $c_k > 0$.
3) For each coordinate basis element e_j, if $e_j \otimes q_k \notin \text{span}\{A_i\}$, append $\langle X, e_j \otimes q_k \rangle = 0$ to $\mathcal{A}(X) = b$.
4) Repeat until $\mathcal{A}$ remains unchanged.

This process will produce a set S satisfying (9), and it will terminate after finitely many repetitions because $\text{rank}(\text{span}\{A_i\})$ increases each time. Each added measurement is implied by the constraints of (1) and does not change the underlying problem.

This process can be viewed as a regularization of the convex problem (1). It differs from a minimal cone regularization because the positive semidefinite cone constraint is kept [10], [12]. Another regularization approach in the literature is the extended Lagrange-Slater Dual (ELSD), which is an alternative to the Langrangian dual that guarantees strong duality at the expensive of polynomially many additional variables [9], [10]. The regularization procedure above is different from ELSD because it get strong duality while keeping the standard Lagrangian dual. The dual variables can hence be viewed as Lagrange multipliers for direct or implied measurements of the matrix X_0. Unfortunately, the procedure above can not be written down mechanically, whereas the ELSD can. Hence, it is less useful for performing the semidefinite optimization than it is as a theoretical process for ensuring that a dual certificate exists.

II. Proofs

A. Notation

For a subspace $V \subset \mathbb{R}^n$, let $V^\perp$ be the orthogonal complement with respect to the ordinary inner product. Let $I_{V^\perp}$ be the matrix corresponding to orthogonal projection of vectors onto $V^\perp$. Let $\mathcal{P}_{V^\perp} X = I_{V^\perp} X I_{V^\perp}$ be the projection of symmetric matrices onto symmetric matrices with row and column spans in $V^\perp$. Let $\mathcal{P}_{x_0^\perp}$ be the special case in the instance where $V = \text{span}\{x_0\}$. In the special case where x_0 is the coordinate basis element e_1, $\mathcal{P}_{x_0^\perp} X$ is the restriction of X to the lower-right $n-1 \times n-1$ block. Let the indicator function for the set Ω be $I_\Omega(X)$, which is zero on Ω and infinity otherwise.

B. Proof of Theorems

Under the assumptions of Theorem 1, the set S trivially satisfies the completeness condition (9). The theorem is thus a special case of Theorem 2, and we will prove them together. As per the constraint qualification (8), it suffices to prove the following technical lemma. This main technical lemma establishes additivity of subgradients of a class of indicator functions. The primary direction uses a separating hyperplane argument to build an item in the subgradient. That argument requires S be closed, as proven in Lemma 2. It also hinges on Lemma 4 which classifies when perturbations from X_0 remain positive semidefinite.

Lemma 1. *Let $X_0 = x_0 x_0^*$ and $\mathcal{A}(X_0) = b$. S satisfies the completeness condition* (9) *if and only if*

$$\partial I_{X \succeq 0, \mathcal{A}(X)=b}(X_0) = \partial I_{X \succeq 0}(X_0) + \partial I_{\mathcal{A}(X)=b}(X_0). \quad (10)$$

Proof of Lemma 1: We omit the proof that $\neg(9) \Rightarrow \neg(10)$. Now, we show (9) $\Rightarrow$ (10). One inclusion in (10) is automatic:

$$\partial I_{X \succeq 0, \mathcal{A}(X)=b}(X_0) = \partial (I_{X \succeq 0} + I_{\mathcal{A}(X)=b})(X_0) \quad (11)$$
$$\supset \partial I_{X \succeq 0}(X_0) + \partial I_{\mathcal{A}(X)=b}(X_0). \quad (12)$$

To prove the other inclusion, we let $Y \notin S = \partial I_{X \succeq 0}(X_0) + \partial I_{\mathcal{A}(X)=b}(X_0)$. We will show that $Y \notin \partial I_{X \succeq 0, \mathcal{A}(X)=b}(X_0)$ by exhibiting a feasible X such that $\langle Y, X - X_0 \rangle > 0$.

As we will prove in Lemma 2, (9) implies that S is closed. By the separating hyperplane theorem, for any $Z \notin S$, there exists a Λ_Z such that

$$\mathcal{A}(\Lambda_Z) = 0, \quad (13)$$
$$\langle \Lambda_Z, Q \rangle \leq 0 \text{ for all } Q \preceq 0, Q \perp X_0, \quad (14)$$
$$\langle \Lambda_Z, M \rangle = 0 \text{ if } \pm M \in S, \quad (15)$$
$$\langle \Lambda_Z, Z \rangle > 0. \quad (16)$$

We observe that (14) implies $\mathcal{P}_{x_0^\perp} \Lambda_Z \succeq 0$.

Let $B = \{qq^* \mid qq^* \perp X_0, qq^* \notin S\}$. We will build a $\tilde{\Lambda}$ satisfying the following properties:

$$\mathcal{A}(\tilde{\Lambda}) = 0, \quad (17)$$
$$\langle \tilde{\Lambda}, Q \rangle \leq 0 \text{ for all } Q \preceq 0, Q \perp X_0, \quad (18)$$
$$\langle \tilde{\Lambda}, M \rangle = 0 \text{ if } \pm M \in S, \quad (19)$$
$$\langle \tilde{\Lambda}, qq^* \rangle > 0 \text{ for all } qq^* \in B. \quad (20)$$

We build $\tilde{\Lambda}$ through the following process. Choose a $q_1 q_1^* \in B$ and find a corresponding $\Lambda_{q_1 q_1^*}$. Restrict B to a set $\tilde{B}$ containing only the elements that are orthogonal to $\Lambda_{q_1 q_1^*}$. All elements in $B \setminus \tilde{B}$ have a positive inner product with $\Lambda_{q_1 q_1^*}$. Choose $q_2 q_2^* \in \tilde{B}$ and find $\Lambda_{q_2 q_2^*}$. Further restrict $\tilde{B}$ to only the elements that are orthogonal to $\Lambda_{q_2 q_2^*}$. Now, all elements in $B \setminus \tilde{B}$ have a positive inner product with $\Lambda_{q_1 q_1^*}$ or $\Lambda_{q_2 q_2^*}$. Repeat this process until B is empty. The process will complete after a finite number of repetitions because the set $\tilde{B}$ is restricted to a space of strictly decreasing dimension at each step. Let $\tilde{\Lambda} = \sum_i \Lambda_{q_i q_i^*}$. We observe (17)–(19) hold due to (13)–(15). Every element of B has a positive inner product with $\Lambda_{q_i q_i^*}$ for some i. Hence, we have (20).

Let $\Lambda = \Lambda_Y + \varepsilon \tilde{\Lambda}$, where ε is small enough that $\langle \Lambda, Y \rangle > 0$. By Lemma 4, if (a) $\mathcal{P}_{x_0^\perp} \Lambda \succeq 0$ and (b) $\Lambda \perp qq^*$ and $qq^* \perp X_0 \Rightarrow \Lambda \perp x_0 \otimes q$, then there exists $\delta > 0$ such that $X_0 + \delta \Lambda \succeq 0$. By (14) and (18), (a) holds. To show (b) holds, we consider a $qq^* \perp \Lambda$, $qq^* \perp X_0$. By (20) and the definition of Λ, qq^* must be in S. By (9), $\pm x_0 \otimes q \in S$. Hence, by (15) and (19), $\Lambda \perp x_0 \otimes q$, and (b) holds.

As given by Lemma 4, let $X = X_0 + \delta \Lambda$. Because $X \succeq 0$ and $\mathcal{A}(\Lambda) = 0$, X is feasible. Additionally, $\langle Y, X - X_0 \rangle > 0$ because $\langle \Lambda, Y \rangle > 0$. Hence, $Y \notin \partial I_{X \succeq 0, \mathcal{A}(X)=b}(X_0)$. ∎

The hyperplane separation argument above requires that S be closed. The following lemma reduces the closedness of $S \subset \mathcal{S}_n$ to an $n-1 \times n-1$ case without the orthogonality constraint, which is proved in Lemma 3.

Lemma 2. *If $S = \{\sum_i \lambda_i A_i + Q \mid Q \preceq 0, Q \perp X_0\}$ satisfies the completeness condition* (9) *then S is closed.*

Proof of Lemma 2: Without loss of generality let $X_0 = e_1 e_1^*$. This can be seen by letting V be an orthogonal matrix with $x_0/\|x_0\|$ in the first column, and by considering the set $V^* S V$. If necessary, linearly recombine the A_i such that the first columns of $A_1, \ldots, A_\ell$ are independent and the first columns of the remaining $A_{\ell+1}, \ldots, A_m$ are zero.

Consider a Cauchy sequence $A^{(k)} + Q^{(k)} \to X$, where $A^{(k)} = \sum_{i=1}^m \lambda_i^{(k)} A_i$. We will establish that $X \in S$. Because $Q^{(k)} \preceq 0$ and $Q^{(k)} \perp e_1 e_1^*$, it is zero in the first row and column. Hence the first column of $\sum_{i=1}^\ell \lambda_i^{(k)} A_i$ converges to the first column of X. By independence, we obtain that $\lambda_i^{(k)}$ converges to some $\lambda_i^{(\infty)}$ for each $1 \leq i \leq \ell$. As a result,

$$\sum_{i=\ell+1}^m \lambda_i^{(k)} A_i + Q^{(k)} \to \overline{X},$$

where $\overline{X} = X - \sum_{i=1}^{\ell} \lambda_i^{(\infty)} A_i$, and $\overline{X}$ is zero in the first row and column.

The problem has now been reduced to one of size $n-1 \times n-1$ without an orthogonality constraint, and Lemma 3 completes the proof. Let $\tilde{A}_i$ be the lower-right $n-1 \times n-1$ sub matrix of A_i. Let $\tilde{S} = \{\sum_{i=\ell+1}^{m} \lambda_i \tilde{A}_i + \tilde{Q} \mid \tilde{Q} \preceq 0\} \in \mathcal{S}_{n-1}$. If $\tilde{q}\tilde{q}^* \in \tilde{S}$ then $\begin{pmatrix} 0 \\ \tilde{q} \end{pmatrix} \begin{pmatrix} 0 \\ \tilde{q} \end{pmatrix}^* \in S$. By (9), $\begin{pmatrix} 0 \\ \tilde{y} \end{pmatrix} \otimes \begin{pmatrix} 0 \\ \tilde{q} \end{pmatrix} \in S \; \forall y \in \mathbb{R}^{n-1}$. By independence of the first columns of $A_1, \ldots, A_\ell$, $\tilde{y} \otimes \tilde{q} \in \tilde{S}$. The conditions of Lemma 3 are met. Hence, $\overline{X} = \sum_{i=\ell+1}^{m} \lambda_i^{(\infty)} A_i + Q^{(\infty)}$ with $Q^{(\infty)} \preceq 0, Q^{(\infty)} \perp e_1 e_1^*$. We conclude $X \in S$ and S is closed. ∎

The closedness of S above relies on the closedness of a lower dimensional $\tilde{S}$ without the orthogonality constraint.

Lemma 3. *The set* $\tilde{S} = \{\sum_i \lambda_i A_i + Q \mid Q \preceq 0\} \subset \mathcal{S}_n$ *is closed if*

$$qq^* \in \tilde{S} \Rightarrow y \otimes q \in \tilde{S} \; \forall y. \tag{21}$$

Proof of Lemma 3: Consider a Cauchy sequence $A^{(k)} + Q^{(k)} \to X$, where $A^{(k)} = \sum_i \lambda_i^{(k)} A_i$. Let $V = \text{span}\{q \mid qq^* \in \tilde{S}\}$. For each $q \in V$, (21) gives that $y \otimes q \in \tilde{S} \; \forall y$. Because $\mathcal{P}_{V^\perp}$ is the projection of matrices onto matrices with row and column spaces living in $V^\perp$,

$$\pm(X - \mathcal{P}_{V^\perp} X) \in \tilde{S} \text{ for any } X. \tag{22}$$

The Cauchy sequence satisfies

$$\mathcal{P}_{V^\perp} A^{(k)} + \mathcal{P}_{V^\perp} Q^{(k)} \to \mathcal{P}_{V^\perp} X. \tag{23}$$

If $\|\mathcal{P}_{V^\perp} A^{(k)}\|_{\text{F}} \to \infty$, then $\frac{\|\mathcal{P}_{V^\perp} A^{(k)}\|_{\text{F}}}{\|\mathcal{P}_{V^\perp} Q^{(k)}\|_{\text{F}}} \to 1$ and $\left\langle \frac{\mathcal{P}_{V^\perp} A^{(k)}}{\|\mathcal{P}_{V^\perp} A^{(k)}\|_{\text{F}}}, \frac{\mathcal{P}_{V^\perp} Q^{(k)}}{\|\mathcal{P}_{V^\perp} Q^{(k)}\|_{\text{F}}} \right\rangle \to -1$ as $k \to \infty$. The sets $\{A \in \mathcal{P}_{V^\perp} \text{span } A_i\} \cap \{\|A\|_{\text{F}} = 1\}$ and $\{Q \preceq 0\} \cap \{\|Q\|_{\text{F}} = 1\}$ are compact. Hence $\langle A, Q \rangle$ achieves its minimum. That minimum value must be -1, which implies that there exists a nonzero, positive semidefinite matrix $-Q \in \mathcal{P}_{V^\perp} \text{span } A_i$. This is is impossible by the construction of V. Suppose $\mathcal{P}_{V^\perp} \mathcal{A}^* \lambda \succeq 0$. By (22), we see $\mathcal{P}_{V^\perp} \mathcal{A}^* \lambda \in \tilde{S}$. Hence every rank-1 component qq^* of $\mathcal{P}_{V^\perp} \mathcal{A}^* \lambda \succeq 0$ belongs to $\tilde{S}$. We reach a contradiction because q would belong to V and can not be in the column space of $\mathcal{P}_{V^\perp} \mathcal{A}^* \lambda$.

Hence, $\mathcal{P}_{V^\perp} A^{(k)}$ has a bounded subsequence. Thus, there is a further subsequence that converges and $\mathcal{P}_{V^\perp} X$ is of the form $\mathcal{P}_{V^\perp}(\sum_i \lambda_i^{(\infty)} A_i + Q^{(\infty)})$. By (22), we conclude $X = \sum_{i=1}^{m} \lambda_i^{(\infty)} A_i + Q^{(\infty)}$ with $Q^{(\infty)} \preceq 0$. ∎

The following lemma establishes a necessary and sufficient condition for when a symmetric perturbation from a positive rank 1 matrix remains positive.

Lemma 4. *Let* $X_0 = x_0 x_0^* \in \mathbb{R}^{n \times n}$. $X_0 + \delta\Lambda \succeq 0$ *for some* $\delta > 0$ *if and only if (a)* $\mathcal{P}_{x_0^\perp} \Lambda \succeq 0$ *and (b)* $\Lambda \perp qq^*$ *and* $q \perp x_0 \Rightarrow \Lambda \perp x_0 \otimes q$.

Proof: Without loss of generality, assume $X_0 = e_1 e_1^*$. In this case $\mathcal{P}_{x_0^\perp}$ is the restriction to the lower-right $n-1 \times n-1$ block. Let $\Lambda_{x_0^\perp} \in \mathcal{S}_{n-1}$ be that lower-right block of Λ. Write the block form

$$\Lambda = \begin{pmatrix} \Lambda_{11} & \rho^* \\ \rho & \Lambda_{x_0^\perp} \end{pmatrix}.$$

First we prove $X_0 + \delta\Lambda \succeq 0 \Rightarrow$ (a) and (b). We immediately have (a) because X_0 is zero on the lower-right subblock. Using a Schur complement, if $1 + \delta\Lambda_{11} > 0$, then

$$X_0 + \delta\Lambda \succeq 0 \Leftrightarrow \Lambda_{x_0^\perp} - \frac{\delta}{1 + \delta\Lambda_{11}} \rho\rho^* \succeq 0. \tag{24}$$

If necessary, δ can be reduced to enforce $1 + \delta\Lambda_{11} > 0$. If (b) does not hold, then there is $\xi \in \mathbb{R}^{n-1}$ such that $\Lambda_{x_0^\perp} \perp \xi\xi^*$ and $\rho \not\perp \xi$. By testing against ξ, we see $\Lambda_{x_0^\perp} - \frac{\delta}{1+\delta\Lambda_{11}} \rho\rho^* \not\succeq 0$

Second, we prove (a) and (b) $\Rightarrow X_0 + \delta\Lambda$ for some $\delta > 0$. Assume (a) and (b) hold. Using the property (24) about Schur complements, it suffices to show

$$\Lambda_{x_0^\perp} - \frac{\delta}{1 + \delta\Lambda_{11}} \rho\rho^* \succeq 0. \tag{25}$$

Let $V = \text{span}\{q \mid \Lambda_{x_0^\perp} \perp qq^*\} \subset \mathcal{S}_{n-1}$. There is some ϵ such that $\Lambda_{x_0^\perp} \succeq \epsilon I_{V^\perp}$. If not, there would be a sequence of $x^{(\varepsilon)} \in V^\perp$ such that $\|x^{(\varepsilon)}\| = 1$ and $0 < x^{(\varepsilon)} \Lambda_{x_0^\perp} x^{(\varepsilon)*} < \varepsilon$. Such $x^{(\varepsilon)}$ would have a convergent subsequence to some $x^{(0)} \in V^\perp$ such that $x^{(0)} \Lambda_{x_0^\perp} x^{(0)*} = 0$, which is impossible.

We note that for any $q \in V$, (b) guarantees $\rho \perp q$. Hence $\rho \in V^\perp$ and there is a sufficiently small δ such that $\frac{\delta}{1+\delta\Lambda_{11}} \rho\rho^* \preceq \varepsilon I_{V^\perp}$. We conclude that (25) holds, and hence $\exists \delta > 0$ such that $X_0 + \delta\Lambda \succeq 0$. ∎

ACKNOWLEDGMENT

The author thanks Laurent Demanet for many useful discussions. This work was partially funded by a NSF Mathematical Sciences Postdoctoral Research Fellowship.

REFERENCES

[1] S. Boyd, L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[2] E. J. Candes, Y. Eldar, T. Strohmer,V. Voroninski. *Phase retrieval via matrix completion.* SIAM J. on Imaging Sciences 6(1), 2011. 199–225.

[3] E. J. Candes, T. Strohmer, V. Voroninski. *PhaseLift: Exact and Stable Signal Recovery from Magnitude Measurements via Convex Programming* To appear in Comm. Pure Appl. Math., 2012.

[4] M. Guignard. *Generalized Kuhn-Tucker conditions for mathematical programming problems in a Banach space.* SIAM J. Control, 7(2):232-241, 1969.

[5] D. Gross. *Recovering Low-Rank Matrices From Few Coefficients In Any Basis* IEEE Trans. Inf. Theory., 57(3), 2011.

[6] X. Li, V. Voroninski. *Sparse Signal Recovery from Quadratic Measurements via Convex Programming.* arXiv preprint 1209.4785., 2012.

[7] B.N. Pshenichniyi. *Necessary conditions for an extremum*. Pure and Applied Mathematics, vol. 4. Dekker, New York (1971).

[8] R. Rockafellar. *Convex Analysis* Princeton University Press, 1970.

[9] M. V. Ramana, *An Exact Duality Theory for Semidefinite Programming and its Complexity Implications*, DIMACS Technical report 95-02R, RUTCOR, Rutgers University, New Brunswick, NJ, 1995.

[10] M. V. Ramana, L. Tuncel, H. Wolkowicz. *Strong Duality for Semidefinite Programming* SIAM J. Optim. 7(3), 641–662, 1997.

[11] A. Shapiro, K. Scheinberg. *Duality and optimality conditions*. In: *Handbook of Semidefinite Programming*. Internat. Ser. Oper. Res. Management Sci., 27, 67–110. Kluwer Academic Boston, 2000.

[12] L. Tuncel, H. Wolkowicz. *Strong duality and minimal representations for cone optimization.* Comput. Optim. Appl 53, 619-648, 2012.

[13] H. Wolkowicz. *Geometry of optimality conditions and constraint qualifications: the convex case.* Math. Programming, 19(1):32-60, 1980.

The restricted isometry property for random convolutions

Felix Krahmer
Institute for Numerical and Applied Mathematics
University of Göttingen
Lotzestraße 16-18
37085 Göttingen, Germany
Email: f.krahmer@math.uni-goettingen.de

Shahar Mendelson
Department of Mathematics
Technion
Haifa 32000, Israel
Email: shahar@tx.technion.ac.il

Holger Rauhut
RWTH Aachen University
Lehrstuhl C für Mathematik (Analysis)
Templergraben 55
52056 Aachen, Germany
Email: rauhut@mathc.rwth-aachen.de

***Abstract*—We present significantly improved estimates for the restricted isometry constants of partial random circulant matrices as they arise in the matrix formulation of subsampled convolution with a random pulse. We show that the required condition on the number m of rows in terms of the sparsity s and the vector length n is $m \gtrsim s \log^2 s \log^2 n$.**

I. Introduction

The theory of *compressed sensing* is based on the observation that many natural signals are approximately sparse in appropriate representation systems, that is, only few entries are significant. The goal of the theory is to devise methods to recover such a signal $\boldsymbol{x}$ from linear measurements

$$\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{x}.$$

For example, it has been shown [1] that under the assumption of a small *restricted isometry constant* on the matrix $\boldsymbol{\Phi}$, approximate recovery via ℓ_1-minimization

$$\min_{\boldsymbol{z}} \|\boldsymbol{z}\|_1 \quad \text{subject to } \boldsymbol{\Phi}\boldsymbol{z} = \boldsymbol{y},$$

(where $\|\boldsymbol{z}\|_p$ denotes the usual ℓ_p-norm) is guaranteed even in the presence of noise.

Here, for a matrix $\boldsymbol{\Phi} \in \mathbb{R}^{m\times n}$ and $s < n$, the restricted isometry constant $\delta_s = \delta_s(\boldsymbol{\Phi})$ is defined as the smallest number such that

$$(1-\delta_s)\|\boldsymbol{x}\|_2^2 \le \|\boldsymbol{\Phi}\boldsymbol{x}\|_2^2 \le (1+\delta_s)\|\boldsymbol{x}\|_2^2 \quad \text{for all } s\text{-sparse } \boldsymbol{x}.$$

If a matrix has a small restricted isometry constant, we also say that the matrix has the *restricted isometry property (RIP)*.

A class of measurement models that is of particular relevance for sensing applications is that of subsampled convolution with a random pulse. In such a model, the convolution of a signal $\boldsymbol{x} \in \mathbb{R}^n$ with a random vector $\boldsymbol{\epsilon} \in \mathbb{R}^n$ given by

$$\boldsymbol{x} \mapsto \boldsymbol{\epsilon} * \boldsymbol{x}, (\boldsymbol{\epsilon} * \boldsymbol{x})_k = \sum_{j=1}^{n} \epsilon_{(k-j) \bmod n} x_j.$$

is followed by a restriction $\boldsymbol{P}_\Omega$ to a deterministic subset of the coefficients $\Omega \subset \{1,\dots,n\}$ and normalization of the columns. The resulting measurement map is linear; its matrix representation $\boldsymbol{\Phi}$ given by

$$\boldsymbol{\Phi}\boldsymbol{x} = \frac{1}{\sqrt{m}}\boldsymbol{\epsilon} * \boldsymbol{x}$$

is called a *partial random circulant matrix*. In this paper, we will focus on the case that the random vector $\boldsymbol{\epsilon}$ is a Rademacher random vector, that is, its entries are independent random variables with distribution $\mathbb{P}(\epsilon_i = \pm 1) = 1/2$. Note, however, that the corresponding results in [2] consider more general random vectors.

The problem of proving the RIP for subsampled convolutions has first been considered in [3]; these results have later been improved in [4]. In [5], a similar problem is considered. Both the sampling sets and the generators, however, are chosen at random. In contrast, our result below holds for an arbitrary fixed sampling sets $\Omega \subset \{1,\dots,n\}$, which is important in applications since in many practical problems, it is natural or desired to consider structured sampling sets such as $\Omega = \{L, 2L, 3L, \dots, mL\}$ for some $L \in \mathbb{N}$; these sets are clearly far from being random.

This paper is structured as follows. In Section II, we present our main result and compare it to the previously best known results. Section IV formulates the problem in terms of chaos processes and presents bounds for such processes in terms of complexity parameters, which are introduced before that in Section III. These bounds are then used to prove the main result in Section V.

II. Main result

Theorem II.1. *([2]) Let $\boldsymbol{\Phi} \in \mathbb{R}^{m\times n}$ be a draw of a partial random circulant matrix generated by a Rademacher vector $\boldsymbol{\epsilon}$. If*

$$m \ge c\delta^{-2} s\,(\log^2 s)(\log^2 n), \tag{1}$$

then with probability at least $1 - n^{-(\log n)(\log^2 s)}$, the restricted isometry constant of $\boldsymbol{\Phi}$ satisfies $\delta_s \le \delta$. The constant $c > 0$ is universal.

This result improves the best previously known estimates for a partial random circulant matrix [4], namely that $m \ge C_\delta (s \log n)^{3/2}$ is a sufficient condition for achieving $\delta_s \le \delta$ with high probability. In particular, Theorem II.1 removes the exponent $3/2$ of the sparsity s, which was already conjectured in [4] to be an artefact of the proof.

Remark II.2. *In certain application scenarios, the ambient dimension n as well as the number of measurements m may*

be given, while one is interested on the sparsity level that still guarantees recovery. To obtain such a bound, we estimate the logarithmic factors in s by $\log(n)$, so we obtain the condition $s \leq \frac{m}{\log^4(n)}$. Again, the dependence is linear up to logarithmic factors, which cannot be guaranteed using previous bounds.

III. Important concepts and definitions

In the proof, two types of complexity parameters of a set of matrices $\mathcal{A}$ will play an important role. The first one, denoted by $d_F(\mathcal{A})$ and $d_{2\to 2}(\mathcal{A})$, is the radius of $\mathcal{A}$ in the Frobenius norm $\|\boldsymbol{A}\|_F = \sqrt{\operatorname{tr}(\boldsymbol{A}^*\boldsymbol{A})}$ and the operator norm $\|\boldsymbol{A}\|_{2\to 2} = \sup_{\|\boldsymbol{x}\|_2 \leq 1} \|\boldsymbol{A}\boldsymbol{x}\|_2$, respectively. That is, $d_F(\mathcal{A}) = \sup_{\boldsymbol{A}\in\mathcal{A}} \|\boldsymbol{A}\|_F$ and $d_{2\to 2}(\mathcal{A}) = \sup_{\boldsymbol{A}\in\mathcal{A}} \|\boldsymbol{A}\|_{2\to 2}$. The second one, Talagrand's functional $\gamma_2(\mathcal{A}, \|\cdot\|_{2\to 2})$, is given by the following definition.

Definition III.1 ([6]). *For a metric space (T, d), an admissible sequence of T is a collection of subsets of T, $\{T_s : s \geq 0\}$, such that for every $s \geq 1$, $|T_s| \leq 2^{2^s}$ and $|T_0| = 1$. Then the γ_2 functional is given by*

$$\gamma_2(T, d) = \inf \sup_{t\in T} \sum_{s=0}^{\infty} 2^{s/2} d(t, T_s),$$

where the infimum is taken with respect to all admissible sequences of T.

Recall that for a metric space (T, d) and $u > 0$, the covering number $N(T, d, u)$ is the minimal number of open balls of radius u in (T, d) needed to cover T. The γ_2-functionals can be bounded in terms of such covering numbers by the well-known Dudley integral (see, e.g., [6]). A formulation specific to a set of matrices $\mathcal{A}$ endowed with the operator norm is

$$\gamma_2(\mathcal{A}, \|\cdot\|_{2\to 2}) \leq C \int_0^{d_{2\to 2}(\mathcal{A})} \sqrt{\log N(\mathcal{A}, \|\cdot\|_{2\to 2}, u)}\, du \qquad (2)$$

for some absolute constant C.

IV. Reformulation as a chaos process

Let $\boldsymbol{\Phi}$ be a partial circulant matrix based on a Rademacher vector, then

$$\begin{aligned}\delta_s(\boldsymbol{\Phi}) &= \sup_{\substack{\boldsymbol{x}\in S^{n-1} \\ |\operatorname{supp}\boldsymbol{x}|\leq s}} \left| \|\boldsymbol{\Phi}\boldsymbol{x}\|_2^2 - 1 \right| \\ &= \sup_{\substack{\boldsymbol{x}\in S^{n-1} \\ |\operatorname{supp} x|\leq s}} \left| \left\| \frac{1}{\sqrt{m}} \boldsymbol{P}_\Omega \boldsymbol{x} * \boldsymbol{\epsilon} \right\|_2^2 - 1 \right| \\ &= \sup_{\substack{\boldsymbol{x}\in S^{n-1} \\ |\operatorname{supp}\boldsymbol{x}|\leq s}} \left| \|\boldsymbol{V}_{\boldsymbol{x}}\boldsymbol{\epsilon}\|_2^2 - \mathbb{E}\|\boldsymbol{V}_{\boldsymbol{x}}\boldsymbol{\epsilon}\|_2^2 \right|,\end{aligned}$$

where $\boldsymbol{V}_{\boldsymbol{x}}$ is defined through $\boldsymbol{V}_{\boldsymbol{x}}\boldsymbol{y} := \frac{1}{\sqrt{m}} \boldsymbol{P}_\Omega \boldsymbol{x} * \boldsymbol{y}$.

As it turns out, the expression $\|\boldsymbol{V}_{\boldsymbol{x}}\boldsymbol{\epsilon}\|_2^2$ is a Rademacher chaos process, that is, it is of the form $\langle \boldsymbol{\epsilon}, \boldsymbol{M}\boldsymbol{\epsilon}\rangle$. This observation was already exploited in [4] to obtain their suboptimal bounds. Our result, however, incorporates the additional observation that the matrix M in the above scenario is $\boldsymbol{V}\boldsymbol{x}^*\boldsymbol{V}\boldsymbol{x}$, hence positive semidefinite.

In the following, we will provide a bound for suprema of chaos processes under such structural assumptions. That is, we study expressions of the form

$$\sup_{\boldsymbol{A}\in\mathcal{A}} \left| \|\boldsymbol{A}\boldsymbol{\epsilon}\|_2^2 - \mathbb{E}\|\boldsymbol{A}\boldsymbol{\epsilon}\|_2^2 \right|.$$

Here $\mathcal{A}$ is an arbitrary set of matrices, which is assumed to be symmetric, i.e., $\mathcal{A} = -\mathcal{A}$.

Theorem IV.1 ([2]). *Let $\mathcal{A} \subset \mathbb{R}^{m\times n}$ be a symmetric set of matrices and let $\boldsymbol{\epsilon}$ be a Rademacher vector of length n. Then*

$$\begin{aligned}&\mathbb{E} \sup_{\boldsymbol{A}\in\mathcal{A}} \left| \|\boldsymbol{A}\boldsymbol{\epsilon}\|_2^2 - \mathbb{E}\|\boldsymbol{A}\boldsymbol{\epsilon}\|_2^2 \right| \\ &\quad \leq C_1 \left(d_F(\mathcal{A})\gamma_2(\mathcal{A}, \|\cdot\|_{2\to 2}) + \gamma_2(\mathcal{A}, \|\cdot\|_{2\to 2})^2 \right) \\ &\quad =: C_1 E.\end{aligned}$$

Furthermore, for $t > 0$,

$$\begin{aligned}&\mathbb{P}\left(\sup_{\boldsymbol{A}\in\mathcal{A}} \left| \|\boldsymbol{A}\boldsymbol{\epsilon}\|_2^2 - \mathbb{E}\|\boldsymbol{A}\boldsymbol{\epsilon}\|_2^2 \right| \geq C_2 E + t \right) \\ &\quad \leq 2\exp\left(-C_3 \min\left\{ \frac{t^2}{V^2}, \frac{t}{U} \right\} \right),\end{aligned}$$

where $U = d_{2\to 2}^2(\mathcal{A})$ and

$$V = d_{2\to 2}(\mathcal{A})(\gamma_2(\mathcal{A}, \|\cdot\|_{2\to 2}) + d_F(\mathcal{A})).$$

The constants $C_1, C_2, C_3 > 0$ are universal.

The proof of this theorem is based on decoupling and a chaining argument, see [2].

V. Proof of Theorem II.1

The proof will be mainly based on Theorem IV.1. Thus we need to control the parameters $d_{2\to 2}(\mathcal{A})$, $d_F(\mathcal{A})$, as well as $\gamma_2(\mathcal{A}, \|\cdot\|_{2\to 2})$ for the set

$$\mathcal{A} = \{\boldsymbol{V}_{\boldsymbol{x}} : \boldsymbol{x} \in D_{s,N}\},$$

where

$$D_{s,N} = \{\boldsymbol{x} \in \mathbb{R}^N : |\operatorname{supp}\boldsymbol{x}| \leq s\}.$$

Since the matrices $\boldsymbol{V}_{\boldsymbol{x}}$ consist of shifted copies of $\boldsymbol{x}$ in all of their m nonzero rows, the ℓ_2-norm of each nonzero row is $m^{-1/2}\|\boldsymbol{x}\|_2$; thus $\|\boldsymbol{V}_{\boldsymbol{x}}\|_F = \|\boldsymbol{x}\|_2 \leq 1$ for all $\boldsymbol{x} \in D_{s,N}$ and

$$d_F(\mathcal{A}) = 1.$$

To bound $d_{2\to 2}(\mathcal{A})$, we will use a Fourier domain description of $\boldsymbol{\Phi}$. Let $\boldsymbol{F}$ by the unnormalized Fourier transform with elements $F_{jk} = e^{2\pi ijk/n}$. As the Fourier transform diagonalizes the convolution operator, for every $1 \leq j \leq n$, $\boldsymbol{F}(\boldsymbol{x} * \boldsymbol{y})_j = (\boldsymbol{F}\boldsymbol{x})_j \cdot (\boldsymbol{F}\boldsymbol{y})_j$. Therefore,

$$\boldsymbol{V}_{\boldsymbol{x}}\boldsymbol{\xi} = \frac{1}{\sqrt{m}} \boldsymbol{P}_\Omega \boldsymbol{F}^{-1} \widehat{\boldsymbol{X}} \boldsymbol{F}\boldsymbol{\xi},$$

where $\widehat{\boldsymbol{X}} = \mathrm{diag}(\boldsymbol{F}\boldsymbol{x})$ is the diagonal matrix, whose diagonal is the Fourier transform $\boldsymbol{F}\boldsymbol{x}$. In short,

$$\boldsymbol{V}_{\boldsymbol{x}} = \frac{1}{\sqrt{m}} \widehat{\boldsymbol{P}}_\Omega \widehat{\boldsymbol{X}} \boldsymbol{F},$$

where $\widehat{\boldsymbol{P}}_\Omega = \boldsymbol{P}_\Omega \boldsymbol{F}^{-1}$. Now observe that for every $\boldsymbol{x} \in D_{s,N}$ with the associated diagonal matrix $\widehat{\boldsymbol{X}}$,

$$\begin{aligned} \|\boldsymbol{V}_{\boldsymbol{x}}\|_{2\to 2} &= \frac{1}{\sqrt{m}} \|\widehat{\boldsymbol{P}}_\Omega \widehat{\boldsymbol{X}} \boldsymbol{F}\|_{2\to 2} \\ &\le \sqrt{\frac{n}{m}} \|\boldsymbol{P}_\Omega \boldsymbol{F}^{-1}\|_{2\to 2} \|\widehat{\boldsymbol{X}}\|_{2\to 2} \\ &\le \frac{1}{\sqrt{m}} \|\widehat{\boldsymbol{X}}\|_{2\to 2} \\ &= \frac{1}{\sqrt{m}} \|\boldsymbol{F}\boldsymbol{x}\|_\infty. \end{aligned} \tag{3}$$

Setting $\|\boldsymbol{x}\|_{\widehat{\infty}} := \|\boldsymbol{F}\boldsymbol{x}\|_\infty$ we observe that

$$\|\boldsymbol{F}\boldsymbol{x}\|_\infty \le \|\boldsymbol{x}\|_1 \le \sqrt{s}\|\boldsymbol{x}\|_2 \le \sqrt{s}$$

for every $x \in D_{s,N}$, and hence

$$d_{2\to 2}(\mathcal{A}) \le \sqrt{s/m}.$$

Next, to estimate the γ_2 functional, recall from (2) that

$$\begin{aligned} &\gamma_2(\mathcal{A}, \|\cdot\|_{2\to 2}) \\ &\quad \le \int_0^{d_{2\to 2}(\mathcal{A})} \log^{1/2} N(\mathcal{A}, \|\cdot\|_{2\to 2}, u) du, \end{aligned}$$

where C is an absolute constant. By (3),

$$\|\boldsymbol{V}_{\boldsymbol{x}} - \boldsymbol{V}_{\boldsymbol{y}}\|_{2\to 2} = \|\boldsymbol{V}_{\boldsymbol{x}-\boldsymbol{y}}\|_{2\to 2} \le m^{-1/2} \|\boldsymbol{x} - \boldsymbol{y}\|_{\widehat{\infty}},$$

and hence for every $u > 0$,

$$N(\mathcal{A}, \|\cdot\|_{2\to 2}, u) \le N(D_{s,N}, m^{-1/2}\|\cdot\|_{\widehat{\infty}}, u).$$

Such covering numbers and the corresponding Dudley integral have been bounded before, e.g., in the context of proving the restricted isometry property for partial random Fourier matrices [7]. The resulting bound for the γ_2-functional is

$$\gamma_2(\mathcal{A}, \|\cdot\|_{2\to 2}) \le C\sqrt{\frac{s}{m}} (\log s)(\log n),$$

where C is an absolute constant. This implies that

$$\gamma_2(\mathcal{A}, \|\cdot\|_{2\to 2}) \le \frac{C}{c}\delta$$

for the given choice of m.

Now, by choosing the constant c in (1) appropriately, one obtains

$$E \le \frac{\delta}{2C_2},$$

where E and C_2 are chosen as in Theorem IV.1. Then Theorem IV.1 yields

$$\mathbb{P}(\delta_s \ge \delta) \le \mathbb{P}\left(\delta_s \ge C_2 E + \delta/2\right) \le \exp(-C_3(m/s)\delta^2),$$

which, after possibly increasing the value of c enough to compensate C_3, exactly amounts to the probability bound given in the theorem. □

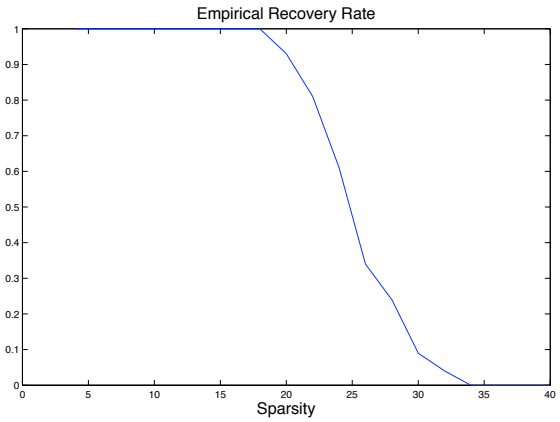


Fig. 1. Empirical recovery rate from partial random circulant measurements for $n = 500$, $m = 100$, and different sparsity levels

VI. Numerical Illustration

We illustrate our results by a numerical example, considering signals of length $n = 500$ and $m = 100$ measurements, letting the sparsity vary. We used a partial random circulant matrix based on a Bernoulli vector, where the rows are selected at random. The plot shows the empirical success rate, that is, in which fraction of the trials the correct signal was recovered (see Figure 1). One should note that our rather simple tests depict the non-uniform success rate: Given a signal, what is the probability that it can be recovered from randomly generated measurements? What we proved above are uniform recovery guarantees: With high probability, a single randomly chosen matrix allows for the recovery of all sparse vectors. This property is much harder to check, as one needs to find the worst vector. While we leave such tests in the context of partial random circulant matrices for future work, we note that strategies to check for this property have been investigated recently in [8].

VII. Conclusion

In this paper we derive bounds on the embedding dimension necessary for a partial random circulant matrix, which is linear in the sparsity. This improves on previous results, in which the sparsity appears with an exponent of $\frac{3}{2}$.

References

[1] E. J. Candès, J., T. Tao, and J. Romberg. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inform. Theory*, 52(2):489–509, 2006.

[2] F. Krahmer, S. Mendelson, and H. Rauhut. Suprema of chaos processes and the Restricted Isometry Property. *Comm. Pure Appl. Math.*, to appear.

[3] J. Haupt, W. Bajwa, G. Raz, and R. D. Nowak. Toeplitz compressed sensing matrices with applications to sparse channel estimation. *IEEE Trans. Inform. Theory*, 56(11):5862–5875, 2010.

[4] H. Rauhut, J. K. Romberg, and J. A. Tropp. Restricted isometries for partial random circulant matrices. *Appl. Comput. Harmon. Anal.*, 32(2):242–254, 2012.

[5] J. K. Romberg. Compressive sensing by random convolution. *SIAM J. Imaging Sci.*, 2(4):1098–1128, 2009.

[6] M. Talagrand. *The Generic Chaining*. Springer Monographs in Mathematics. Springer-Verlag, 2005.

[7] M. Rudelson and R. Vershynin. On sparse reconstruction from Fourier and Gaussian measurements. *Comm. Pure Appl. Math.*, 61:1025–1045, 2008.

[8] J. Blanchard, C. Cartis, and J. Tanner. Compressed Sensing: How sharp is the Restricted Isometry Property? *SIAM Review*, 53(1):105–125, 2011.

Multivariate sampling Kantorovich operators: approximation and applications to civil engineering

Federico Cluni
Dipartimento Ingegneria Civile ed Ambientale
Università degli Studi di Perugia
Perugia, 06125 Italy
Email: cluni@strutture.unipg.it

Danilo Costarelli
Dipartimento di Matematica
e Fisica, Università degli Studi
Roma Tre, Roma, 00146 Italy
Email: costarel@mat.uniroma3.it

Anna Maria Minotti and Gianluca Vinti
Dipartimento di Matematica e Informatica
Università degli Studi di Perugia
Perugia, 06123 Italy
Emails: annamaria.minotti@dmi.unipg.it
and mategian@unipg.it

***Abstract*—In this paper, we present the theory and some new applications of linear, multivariate, sampling Kantorovich operators. By means of the above operators, we are able to reconstruct pointwise, continuous and bounded signals (functions), and to approximate uniformly, uniformly continuous and bounded functions. Moreover, the reconstruction of signals belonging to Orlicz spaces are also considered. In the latter case, we show how our operators can be used to approximate not necessarily continuous signals/images, and an algorithm for image reconstruction is developed. Several applications of the theory in civil engineering are obtained. Thermographic images, such as masonries images, are processed to study the texture of the buildings, thus to separate the stones from the mortar and finally a real-world case-study is analyzed in terms of structural analysis.**

I. Introduction

In [1] the authors introduced the linear sampling Kantorovich operators and studied, in particular, their convergence in the general setting of Orlicz spaces, in one-dimesional case. Later these results have been extended in [8] to the multivariate setting, in [12], [9] to the nonlinear case and in a more general context in [13], [2].

In this paper, we obtain applications to civil engineering by using the linear multivariate sampling Kantorovich operators $(S_w)_{w>0}$, defined by

$$(S_w f)(\underline{x}) := \sum_{\underline{k}\in\mathbb{Z}^n} \chi(w\underline{x}-t_{\underline{k}}) \left[\frac{w^n}{A_{\underline{k}}}\int_{R^w_{\underline{k}}} f(\underline{u})\, d\underline{u}\right], \quad \textbf{(I)}$$

for every $\underline{x}\in\mathbb{R}^n$, where $f:\mathbb{R}^n\to\mathbb{R}$ is a locally integrable function such that the above series is convergent for every $\underline{x}\in\mathbb{R}^n$. Here $\chi:\mathbb{R}^n\to\mathbb{R}$ is a kernel function satisfying suitable properties, $t_{\underline{k}}=(t_{k_1},...,t_{k_n})$ is a vector where $(t_{k_i})_{k_i\in\mathbb{Z}}$, $i=1,...,n$ is a sequence of real numbers with some properties and where

$$R^w_{\underline{k}} := \left[\frac{t_{k_1}}{w},\frac{t_{k_1+1}}{w}\right]\times\left[\frac{t_{k_2}}{w},\frac{t_{k_2+1}}{w}\right]\times...\times\left[\frac{t_{k_n}}{w},\frac{t_{k_n+1}}{w}\right],$$

$w>0$ and $A_{\underline{k}}=\Delta_{k_1}\cdot\Delta_{k_2}\cdot...\cdot\Delta_{k_n}$ with $\Delta_{k_i}:=t_{k_i+1}-t_{k_i}$, $i=1,...,n$. For the study of the above family **(I)**, see [8].
The sampling series **(I)** represents a Kantorovich version of the generalized sampling operators introduced by P.L. Butzer and his school at Aachen (see e.g. [4]). Here, in place of the sample values $f(\underline{k}/w)$ one has an average of f in a small pluri-rectangle containing $\underline{k}/w$ (here instead of $\underline{k}$, we have a general sequence $t_{\underline{k}}$, obtaining a non uniform sampling). This situation very often occurs in Signal Processing, when one cannot match exactly the "node" $t_{\underline{k}}/w$: this represents the so called "time-jitter error". Therefore our theory reduces time-jitter errors calculating the information in a neighborhood of a point rather than exactly at that point.
For the sampling Kantorovich operators **(I)**, we study the pointwise convergence for continuous and bounded functions, the uniform convergence, for uniformly continuous and bounded functions, and the modular convergence, for functions belonging to Orlicz spaces (see e.g. [3]). The latter case, allows to treat the case of L^p-signals, i.e., not necessarily continuous signals; note that in multivariate setting, when one deals with images, discontinuities are concentrated in the contours or edges of the image itself, in terms of jumps of grey levels (see [8], [9]). To show the versatility of our theory, we study various applications to civil engineering images. In this subject the images, in particular thermographic images, are used to make non-invasive investigations of structures, to analyze the story of the buildings or of the building walls, to make diagnosis and monitoring buildings, and to make structural measurements. The thermography is a remote sensing technique, performed by the image acquisition in the infrared. Moreover, these images are also used in civil engineering for image texture, i.e., for the separation between the bricks and the mortar in masonries images. Unfortunately, the direct application of the image texture algorithm to the thermographic images, can produce errors, as an incorrect separation between the bricks and the mortar. Then, we use the sampling Kantorovich operators to process the thermographic images before to apply the texture. In this way, the result produced by the texture becomes more refined and therefore we can apply structural analysis to a real-world case-study after the calculation of the various parameters involved.

A. Approximation results

In this section, we treat the main approximation results for the multivariate sampling Kantorovich operators. In what follows, we denote by $t_{\underline{k}}=(t_{k_1},...,t_{k_n})$ a vector where each $(t_{k_i})_{k_i\in\mathbb{Z}}$, $i=1,...,n$ is a sequence of real numbers with $-\infty<t_{k_i}<t_{k_{i+1}}<+\infty$, $\lim_{k_i\to\pm\infty}t_{k_i}=\pm\infty$, for every

$i = 1, ..., n$, and such that there exist Δ, $\delta > 0$ for which $\delta \leq \Delta_{k_i} := t_{k_{i+1}} - t_{k_i} \leq \Delta$, for every $i = 1, ..., n$.

A function $\chi : \mathbb{R}^n \to \mathbb{R}$ will be called a kernel if it satisfies the following properties:

- $(\chi 1)$ $\chi \in L^1(\mathbb{R}^n)$ and is bounded in a neighborhood of $\underline{0} \in \mathbb{R}^n$;
- $(\chi 2)$ For every $\underline{u} \in \mathbb{R}^n$, $\sum_{\underline{k} \in \mathbb{Z}^n} \chi(\underline{u} - t_{\underline{k}}) = 1$;
- $(\chi 3)$ For some $\beta > 0$,

$$m_{\beta,\Pi^n}(\chi) = \sup_{\underline{u} \in \mathbb{R}^n} \sum_{\underline{k} \in \mathbb{Z}^n} \left|\chi(\underline{u} - t_{\underline{k}})\right| \cdot \left\|\underline{u} - t_{\underline{k}}\right\|_2^{\beta} < +\infty,$$

where $\|\cdot\|_2$ denotes the usual Euclidean norm.

We may now state the following theorem for the linear multivariate sampling Kantorovich operators **(I)** based upon the kernel function χ.

Theorem 1: Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuous and bounded function. Then, for every $\underline{x} \in \mathbb{R}^n$,

$$\lim_{w \to +\infty} (S_w f)(\underline{x}) = f(\underline{x}).$$

In particular, if $f : \mathbb{R}^n \to \mathbb{R}$ is uniformly continuous and bounded, then

$$\lim_{w \to +\infty} \|S_w f - f\|_{\infty} = 0,$$

where $\|\cdot\|_{\infty}$ denotes the usual sup-norm.

We now recall some basic fact concerning Orlicz spaces, see e.g. [11], [3].

Let $\varphi : \mathbb{R}_0^+ \to \mathbb{R}_0^+$ be a φ-function, i.e. φ satisfies the following assumptions:

1) $\varphi(0) = 0$, $\varphi(u) > 0$ for every $u > 0$;
2) φ is continuous and non decreasing on $\mathbb{R}_0^+$;
3) $\lim_{u \to \infty} \varphi(u) = +\infty$.

For a fixed φ-function φ, one can consider the functional $I^{\varphi} : M(\mathbb{R}^n) \to [0, +\infty]$, where $M(\mathbb{R}^n)$ denotes the set of all measurable functions $f : \mathbb{R}^n \to \mathbb{R}$. We define

$$I^{\varphi}[f] := \int_{\mathbb{R}^n} \varphi(|f(\underline{x})|)\, d\underline{x}, \qquad (f \in M(\mathbb{R}^n)).$$

The Orlicz space generated by φ is defined by

$$L^{\varphi}(\mathbb{R}^n) := \{f \in M(\mathbb{R}^n) : I^{\varphi}[\lambda f] < \infty, \text{ for some } \lambda > 0\}.$$

We can introduce in $L^{\varphi}(\mathbb{R}^n)$, a notion of convergence, called "modular convergence", which induces a topology (modular topology) on the space ([11], [3]). Namely, we will say that a net of functions $(f_w)_{w>0} \subset L^{\varphi}(\mathbb{R}^n)$ is modularly convergent to a function $f \in L^{\varphi}(\mathbb{R}^n)$ if

$$\lim_{w \to +\infty} I^{\varphi}[\lambda(f_w - f)] = 0$$

for some $\lambda > 0$.

Now, by means of a modular estimate for the operators **(I)** and using a density result, we may state the following modular convergence theorem for the sampling Kantorovich operators (based upon the kernel function χ) in Orlicz spaces.

Theorem 2: Let φ be a convex φ-function. For every $f \in L^{\varphi}(\mathbb{R}^n)$, there exists $\lambda > 0$ such that

$$\lim_{w \to +\infty} I^{\varphi}[\lambda(S_w f - f)] = 0.$$

Now, choosing $\varphi(u) = u^p$, $1 \leq p < \infty$, we have $L^{\varphi}(\mathbb{R}^n) = L^p(\mathbb{R}^n)$ and $I^{\varphi}[f] = \|f\|_p^p$, where $\|\cdot\|_p$ is the usual L^p-norm. Then, from Theorem 2 we obtain the following corollary.

Corollary 1: For every $f \in L^p(\mathbb{R}^n)$, $1 \leq p < +\infty$,

$$\lim_{w \to +\infty} \|S_w f - f\|_p = 0.$$

The corollary above, allows us to reconstruct L^p-signals (in L^p-sense), therefore not necessarily continuous. Other examples of Orlicz spaces for which the theory can be applied, are given by the Zygmund spaces (or interpolation spaces) and by the exponential spaces, see e.g. [11], [3], [1], [8].

B. The choice of the kernels

In the theory of sampling Kantorovich operators an important role is played by the kernels χ. A procedure to construct examples of multivariate kernel is to use product kernels by means of one-dimensional kernels. For a sake of simplicity, we consider our operators in case of uniform sampling ($t_{\underline{k}} = \underline{k}$), and denote by $\chi_1, ..., \chi_n$, the one-dimensional kernels $\chi_i : \mathbb{R} \to \mathbb{R}$ satisfying conditions $(\chi 1)$, $(\chi 2)$ and $(\chi 3)$ for $n = 1$. In [4], [8] is proved that the multivariate function

$$\chi(\underline{x}) := \prod_{i=1}^{n} \chi_i(x_i),$$

for every $\underline{x} = (x_1, ..., x_n) \in \mathbb{R}^n$, is a multivariate kernel for our operators $(S_w)_{w>0}$ satisfying the assumption of the theory. Then, it is now sufficient to give examples of one-dimensional kernels satisfying $(\chi 1)$, $(\chi 2)$ and $(\chi 3)$. Remarkable examples of kernels with compact support, are given by the well-known central B-spline of order $k \in \mathbb{N}$, defined by

$$M_k(x) := \frac{1}{(k-1)!} \sum_{i=0}^{k} (-1)^i \binom{k}{i} \left(\frac{k}{2} + x - i\right)_+^{k-1}.$$

where the function $(x)_+ := \max\{x, 0\}$ denotes the positive part of $x \in \mathbb{R}$ (see [1], [12], [8]). Other well-known examples of one-dimensional kernels are given by the Jackson-type kernels, defined by

$$J_k(x) = c_k \text{sinc}^{2k}\left(\frac{x}{2k\pi\alpha}\right), \qquad x \in \mathbb{R},$$

with $k \in \mathbb{N}$, $\alpha \geq 1$, for a suitable constant c_k and where the sinc-function is defined by

$$\text{sinc}(x) := \begin{cases} 1, & x = 0, \\ \dfrac{\sin(\pi x)}{\pi x}, & \text{otherwise}, \end{cases}$$

(see [5], [1]).

It is also possible to consider kernels which are not of product type. For instance, one can take into consideration *radial kernels*, i.e., functions for which the value depends on the Euclidean norm of the argument only. Example of such a

kernel can be given, for example, by the Bochner-Riesz kernel, defined as follows

$$b^{\alpha}(\underline{x}) := 2^{\alpha}\Gamma(\alpha+1)\|\underline{x}\|_2^{-(n/2)+\alpha}\mathcal{B}_{(n/2)+\alpha}(\|\underline{x}\|_2),$$

for $\underline{x} \in I\!R^n$, where $\alpha > (n-1)/2$, $\mathcal{B}_\lambda$ is the Bessel function of order λ and Γ is the Euler function. For more details about this matter, see e.g. [4].

C. Applications to Image Processing

In this section, we show how the multivariate sampling Kantorovich operators can be applied to process digital images, see [8], [9]. Every bi-dimensional grey scale image A (matrix) can be modeled as a step function I, with compact support, belonging to $L^p(I\!R^2)$, $1 \leq p < +\infty$. The most natural way to define I is:

$$I(x,y) \ := \ \sum_{i=1}^{m}\sum_{j=1}^{m} a_{ij}\cdot \mathbf{1}_{ij}(x,y) \quad ((x,y)\in I\!R^2),$$

where $\mathbf{1_{ij}}(\mathbf{x},\mathbf{y})$, $i,j = 1,2,...,m$, are the characteristics functions of the sets $(i-1,\ i]\times(j-1,\ j]$ (i.e. $\mathbf{1_{ij}}(\mathbf{x},\mathbf{y}) = \mathbf{1}$, for $(x,y) \in (i-1,\ i]\times(j-1,\ j]$ and $\mathbf{1_{ij}}(\mathbf{x},\mathbf{y}) = \mathbf{0}$ otherwise). Note that the above function $I(x,y)$ is defined in such a way that, to every pixel (i,j) it is associated the corresponding grey level a_{ij}. Then, we can now consider the family of bivariate sampling Kantorovich operators applied to the function I, $(S_wI)_{w>0}$ (for some kernel χ) that approximate I in L^p-sense. Now, in order to obtain a new image (matrix) that approximates the original one, it is sufficient to sample S_wI (for some $w>0$) with a fixed sampling rate. In particular, we can reconstruct the approximating images (matrices) taking into consideration different sampling rates and this is possible since we know the analytic expression of S_wI.

Obviously, if the sampling rate is chosen higher than the original sampling rate, one can get a new image that has a better resolution than the original one's. The above procedure has been implemented by using MATLAB, in order to obtain an algorithm based on the multivariate sampling Kantorovich theory.

D. Applications to civil engineering images

In this section, we propose some new applications of the algorithm, based on the multivariate sampling Kantorovich operators, to civil engineering images.

The most widely used images in this areas are the thermographic images, largely used to make diagnosis and monitoring buildings, and to make structural measurements. The thermography is a remote sensing technique, performed by the image acquisition, in the infrared. The thermographic images are obtained by the thermograph, that in practice consists in a thermal camera for detecting radiation in the infrared range of the electromagnetic spectrum, and perform measurements related with the emission of this radiation. This tool is able to detect the temperatures of the bodies analyzed by measuring the intensity of infrared radiation emitted by the body under examination. All the objects at a temperature above absolute zero emit radiation in the infrared range. The thermography allows to avoid the use of invasive techniques of investigation for buildings. Moreover, these images are also used in civil engineering for image texture, i.e., for the separation between the bricks and mortar in masonries images. The image texture algorithm performs as follows: first of all we apply a median filter to the image using a suitable mask, then the image is converted into a black and white image by means of a suitable thresholding, in order to obtain a consistent separation of the phases; the area consisting of white pixels denote the inclusions (stones or bricks) and the remaining areas of black pixels denotes the mortar joints. Finally, morphological operators are used to enhance the quality of the separation of the phase: closing of the area to eliminate salt-and-pepper noise, erosion and dilation to smooth the contours of the inclusions. The image obtained is characterized by a consistent separation of phases, where each stone is surrounded by mortar joints and unrealistic conjunction of inclusions is avoided as much as possible (see e.g. [6]).

The direct application of the image texture algorithm to the thermographic images, can produce errors (see e.g. Figure 2 (left) and (right)), as an incorrect separation between the bricks and the mortar. Then, we can use the sampling Kantorovich operators (see in Figure 1 (left) for the original termographic image of a masonry, and Figure 1 (right) for a reconstruction) to process the thermographic images before to apply the texture, in order to obtain images suitable for the application of the texture algorithm (see e.g. the comparison between Figure 2 (left) and (right)).

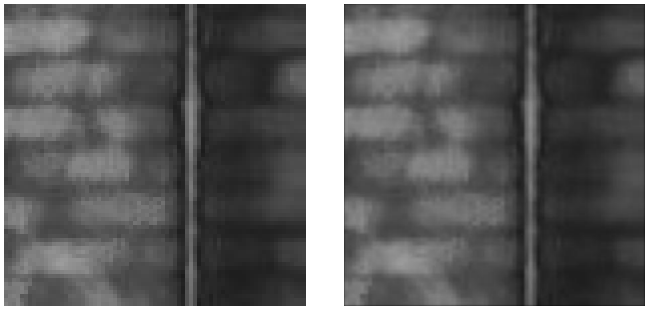


Fig. 1. Recontruction of the original image (left, 75 × 75 pixel) by the sampling Kantorovich operators with the bivariate Jackson kernel with $k = 4$ and $\alpha = 1$, for $w = 40$ (right, 450 × 450 pixel)

In order to perform structural analysis, the mechanical characteristics of an homogeneous material equivalent to the original heterogeneous material are sought (see e.g. [7]). The equivalence is in the sense that, when subjected to the same boundary conditions, the overall response in terms of mean values of stresses and deformations is the same, see e.g. [10]. In particular, the equivalent elastic properties taking into account the effective characteristics of the micro-structure can be estimated by a suitable choice of two kinds of boundary conditions: i) in terms of displacements (essential boundary conditions); ii) in terms of forces (natural boundary conditions). In order to solve the boundary condition problem, the Finite Element Method (F.E.M.) is used.

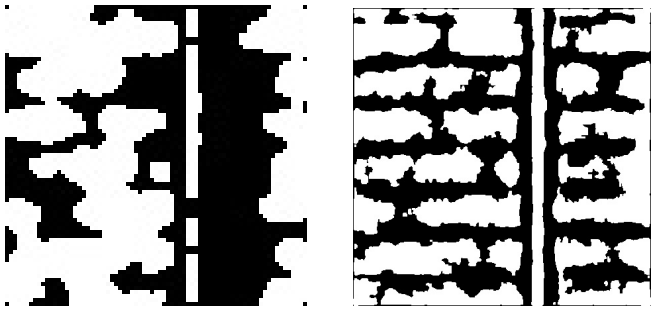


Fig. 2. On the left, we have the image texture of the original image of Figure 1 (left). On the right, we have the image texture of Figure 1 (right), reconstructed by the sampling Kantorovich operators.

The estimated mechanical properties can be used to analyze a real-world case-study. In particular the proposed approach allows to overcome some difficulties, that arise when dealing with the vulnerability analysis of existing structures, which are: i) the knowledge of the actual geometry of the walls (in particular the identification of hidden doors and windows); ii) the identification of the actual texture of the masonry and the distribution of inclusions and mortar joints, and from this iii) the estimation of the elastic characteristics of the masonry. It is noteworthy that, for item i) the engineer has limited knowledge, due to the lack of documentation, while for items ii) and iii) he usually use tables proposed in technical manuals and standards which however give large bounds in order to encompass the generality of the real masonries.

E. Future developments

A future development of the present paper is to study applications of the algorithm based on the sampling Kantorovich operators to biomedical images. In biomedicine, images cover a fundamental role for the clinical diagnosis, surgery (Endovascular aneurysm repair - EVAR), and for the patient follow up. For this purpose, it reveals of a certain importance that the contours of the biomedical images are clearly visible. Then becomes important having at disposal an algorithm for image reconstruction and enhancement. Our aim is to treat images in the field of Vascular Surgery, in collaboration with a group of radiologists and vascular surgeons of the sections of Vascular and Endovascular Surgery and Diagnostic Radiology and Interventional of the University of Perugia. In particular our aim is to apply our algorithm to images related to the aneurysmal aortic and steno-obstructive pathology of epiaortic and peripheral vessels in order to improve the medical diagnosis.

II. Conclusion

In this paper, we present the theory of the multivariate sampling Kantorovich operators. Approximation results are given in various settings. Applications of the theory to Image Processing are also shown. In particular, new applications of the algorithm based on the sampling Kantorovich operators to civil engineering images are obtained. The applications related to image texture algorithm is significant, and of practical utility in seismic engineering.

Acknowledgment

The authors would like to thank the Fondazione Cassa di Risparmio di Perugia, Project N. 2010.011.0403 and G.N.A.M.P.A. of I.N.D.A.M. for their support to this research.

References

[1] C. Bardaro, P.L. Butzer, R.L. Stens and G. Vinti, *Kantorovich-Type Generalized Sampling Series in the Setting of Orlicz Spaces*, Sampl. Theory Signal Image Process. 6 (1) 29-52, 2007.

[2] C. Bardaro and I. Mantellini, *On convergence properties for a class of Kantorovich discrete operators*, Numer. Funct. Anal. Optim. 33 (4) 374-396, 2012.

[3] C. Bardaro, J. Musielak and G. Vinti, *Nonlinear Integral Operators and Applications*, Gruyter Series in Nonlinear Analysis and Applications 9, New York - Berlin, 2003.

[4] P.L. Butzer, A. Fisher and R.L. Stens, *Generalized sampling approximation of multivariate signals: theory and applications*, Note di Matematica 10 (1), 173-191, 1990.

[5] P.L. Butzer and R.J. Nessel, *Fourier Analysis and Approximation*, I, Academic Press, New York-London, 1971.

[6] N. Cavalagli, F. Cluni and V. Gusella, *Evaluation of a Statistically Equivalent Periodic Unit Cell for a quasi-periodic masonry'*, International Journal of Solids and Structures, submitted, 2013.

[7] F. Cluni and V. Gusella, *Homogenization of non-periodic masonry structures*, International Journal of Solids and Structures 41, 1911-1923, 2004.

[8] D. Costarelli and G. Vinti, *Approximation by Multivariate Generalized Sampling Kantorovich Operators in the Setting of Orlicz Spaces*, Bollettino U.M.I. (9) IV 445-468, 2011.

[9] D. Costarelli and G. Vinti, *Approximation by Nonlinear Multivariate Sampling-Kantorovich Type Operators and Applications to Image Processing*, Numer. Funct. Anal. Optim. 34 (6) 1-26, 2013.

[10] R. Hill, *Elastic properties of reinforced solids: some theoretical principles*, Journal of the Mechanics and Physics of Solids 11, 357-372, 1963.

[11] J. Musielak, *Orlicz Spaces and Modular Spaces*, Springer-Verlag, Lecture Notes in Math. 1034, 1983.

[12] G. Vinti and L. Zampogni, *Approximation by means of nonlinear Kantorovich sampling type operators in Orlicz spaces*, J. Approx. Theory 161, 511-528, 2009.

[13] G. Vinti and L. Zampogni, *A Unifying Approach to Convergence of Linear Sampling Type Operators in Orlicz Spaces*, Adv. Differential Equations 16 (5-6) 573-600, 2011.

On the Number of Degrees of Freedom of Band-Limited Functions

Tatiana Levitina
Institut Computational Mathematics
Technische Universität Braunschweig
Braunschweig, D–38106, Germany
Email: t.levitina@tu-braunschweig.de

***Abstract*—The concept of the number of degrees of freedom of band-limited signals is discussed. Classes of band-limited signals obtained as a result of successive application of the truncated direct and truncated inverse Fourier transforms are shown to posses a finite number of degrees of freedom.**

I. Introduction

Let a signal f be Ω-band-limited, i.e. representable as $f(t) = \frac{1}{2\pi}\int_{-\Omega}^{\Omega} F(\omega)\, e^{it\omega}\, d\omega$. In accordance with the famous Whittaker–Kotel'nikov–Shannon (WKS) sampling theorem [1] f can be fully reconstructed from its uniformly distributed samples $f\left(\frac{\pi k}{\Omega}\right)$, $k = 0, \pm1, \pm2, \ldots,$

$$f(t) = \sum_{k=-\infty}^{\infty} f\left(\frac{\pi k}{\Omega}\right) \frac{\sin\Omega\left(t - \pi k/\Omega\right)}{\Omega\left(t - \pi k/\Omega\right)}. \qquad (1)$$

A common notion in the field is that, upon a certain duration T, this signal has no more than $2K+1$, $K = [\Omega T/(2\pi)]$[1] degrees of freedom, since it can be *completely* recovered from just $2K+1$ of its samples taken at the points $k\pi/\Omega \in (-T/2, T/2)$, as if $f(t) \equiv 0$ outside $(-T/2, T/2)$. This notion is refuted by the fact that a function cannot be simultaneously time- and band-limited [2].

The function f_K obtained from the sampling formula (1) by truncating the series to a finite number of terms, $|k| \le K$, is Ω-band-limited and coincides with f at each sampling point t_k, $k = 0, \pm1, \ldots, \pm K$. On the other hand, at any other time moment t the difference between $f(t)$ and $f_K(t)$ may be arbitrary large, depending on the values $f\left(\frac{\pi k}{\Omega}\right)$ from outside the interval $(-T/2, T/2)$.

Nonetheless, for band–limited functions essentially concentrated inside a finite time interval the concept of the number of degrees of freedom (NDF) makes a certain sense. For various definitions of signal concentration in the time domain we refer the reader to the monograph [3] and the literature cited therein. The $sinc$-function translates themselves are not highly concentrated inside this interval, therefore the classical sampling formula does not enable such a definition. Instead another formulation of the sampling theorem given by G. Walter and X. Shen in [4] will be helpful. The newly formulated sampling theorem employs the eigenfunctions of the finite, i.e. truncated to a finite interval, Fourier transform (TFT)[2]. The NDF of an essentially concentrated in the time domain signal can then be defined as the number of the TFT eigenfunctions which suffices to well-approximate this signal.

With a help of the sampling formula one can easily synthesize a signal of any desired NDF. This means that without an additional knowledge about the signal, the number of signal samples contributing significantly to the sampling series is not known *a priori*. However for particular classes of band–limited functions the upper bounds on this number and therewith on the NDF can be effectively computed. Thus in [4] it was shown that, if an Ω-band-limited signal is highly concentrated in $I_T = (-T/2, T/2)$ and its Fourier transform is sufficiently smooth, it has $[\Omega T/\pi] + 1$ degrees of freedom, the same number as the above erroneous explanation would give.

Yet the smoothness of the Fourier transform seems to be a too rigorous requirement. Even the TFT eigenfunctions, though proved to be the most concentrated in the time domain among other band-limited functions, have jumps in the frequency domain. The Fourier transforms of the convolutions of the Ω-band-limited TFT eigenfunctions are also discontinuous. Still they are highly concentrated in the interval $(-2\Omega, 2\Omega)$ and require no more than $2\Omega^2/\pi + 1$ of $\sqrt{2}\Omega$-band-limited TFT eigenfunctions for reconstruction via the Walter-Shen sampling formula [5].

In the present work we shall introduce a wide variety of classes of band-limited functions with a given NDF, one of them includes the convolutions of the TFT eigenfunctions as particular examples. The relevant upper bounds for the truncation error of the sampling series will be derived. We shall also touch upon a possible generalization to higher dimensions.

We begin with a brief survey of known results related to TFT eigenfunctions.

II. Band–limited functions

As is well-known [1], the Paley–Wiener space,

$$\mathcal{PW}_a := \left\{ f(x) \,\middle|\, f(x) = \frac{1}{2\pi}\int_{-a}^{a} e^{i\,xy} g(y)\, dy, g \in \mathcal{L}_2(-a, a) \right\}.$$

[1] here square brackets denote the integer part

[2] Since the acronym FFT stands commonly for the fast Fourier transform, we use the abbreviation TFT for the finite Fourier transform.

is a reproducing kernel Hilbert space with the reproducing kernel

$$G(x,y) := \frac{\sin a\,(x-y)}{\pi(x-y)}.$$

This follows from the Fourier inversion formula $\hat{F}\,[f]\,(y) = \hat{\chi}_a(y)\,g(y)$, which holds for all functions from $\mathcal{PW}_a$. Here $\hat{F}[\cdot]$ stands for the Fourier transform and $\hat{\chi}_a$ is the operator of multiplication by $\chi_a(\cdot)$, whereas $\chi_a(\cdot)$ is the characteristic function of the interval I_a

$$\chi_a(x) = \begin{cases} 1, & x \in I_a, \\ 0, & x \in \mathbb{R} \setminus I_a \end{cases} .$$

The classical WKS sampling formula

$$f(x) = \sum_{k=-\infty}^{\infty} f\left(\frac{\pi k}{a}\right) \frac{\sin a\left(x-\pi k/a\right)}{a\left(x-\pi k/a\right)}, \quad \text{for } \forall f \in \mathcal{PW}_a, \tag{2}$$

reflects the fact that the sequence of point evaluation functionals $G\left(\frac{\pi\,k}{a}, y\right)$, $k = 0, \pm 1, \ldots$, forms an orthonormal basis in $\mathcal{PW}_a$ [1].

Yet Eq. (2) is easy to obtain via the direct integration of the Fourier expansion of the associated function $g \in \mathcal{L}_2(I_a)$:

$$g(y) = \sqrt{\frac{1}{2a}} \sum_{k=-\infty}^{\infty} g_k\, e^{-i\pi k y/a} = \frac{1}{2a} \sum_{k=-\infty}^{\infty} f\left(\frac{\pi k}{a}\right) e^{-i\pi k y/a}, \tag{3}$$

since the Fourier coefficients g_k, $k = 0, 1, 2, \ldots$, are

$$g_k := \sqrt{\frac{1}{2a}} \int_{-a}^{a} e^{i\,\pi k\,y/a} g(y)\,dy = \sqrt{\frac{1}{2a}}\, f\left(\frac{\pi k}{a}\right).$$

The series in the right–hand side of Eq. (3) converges in $\mathcal{L}_2(-a,a)$. As a consequence, the sampling formula converges both in the $\mathcal{L}_2$–norm and uniformly on $\mathbb{R}$.

III. Prolates—eigenfunctions of the truncated Fourier transform

Another sampling formula invented and studied in [4] is written in terms of the TFT eigenfunctions. The TFT operator $\hat{F}_a$ is first introduced as acting on $\mathcal{L}_2(-a,a)$ by

$$\hat{F}_a[g](x) = \frac{1}{2\pi} \int_{-a}^{a} e^{i\,xy} g(y)\,dy, \quad x \in (-a,a). \tag{4}$$

Its eigenfunctions $\psi_l(a,x) = \psi_l(x)$ defined at $x \in (-a,a)$ via the equation

$$\mu_l\,\psi_l(x) \;=\; \frac{1}{2\pi} \int_{-a}^{a} e^{i\,xy} \psi_l(y)\,dy, \quad l = 0, 1, \ldots, \tag{5}$$

are ordered according to the absolute magnitude of the associated eigenvalues, $|\mu_0| > |\mu_1| > \ \ldots$. One can choose ψ_l to be real valued and normalized, so that

$$\|\psi_l\|_a^2 = \int_{-a}^{a} \left(\psi_l(x)\right)^2\,dx \;=\; 1.$$

Eqs. (4), (5) are then used to extend the functions in the left hand side to the entire real axis. Therewith an operator is defined that maps $\mathcal{L}_2(I_a)$ on the Paley–Winer space $\mathcal{PW}_a$. We shall keep for this operator the same notation $\hat{F}_a$, similarly the extensions of the TFT eigenfunctions are hereafter denoted by ψ_l, since it shall not cause any ambiguity. One can also define $\hat{F}_a$ as the composition $\hat{F}_a := \hat{F} \circ \hat{\chi}_a : \mathcal{L}_2(\mathbb{R}) \to \mathcal{PW}_a$. Evidently ψ_l extended to $\mathbb{R}$ are the eigenfunctions of $\hat{F} \circ \hat{\chi}_a$.

The double definition of the TFT operator provides a double set of properties of its eigenfunctions. Thus functions ψ_l are pairwise orthogonal both on the finite interval I_a and on $\mathbb{R}$,

$$\int_{-a}^{a} \psi_l(x)\,\psi_s(x)\,dx \;=\; \delta_{ls}\,, \qquad \int_{-\infty}^{\infty} \psi_l(x)\,\psi_s(x)\,dx \;=\; \frac{1}{\gamma_l}\,\delta_{ls},$$

where $\gamma_l := |\mu_l|^2/2\pi$. In addition to that ψ_l form a basis both in $\mathcal{L}_2(-a,a)$ and in $\mathcal{PW}_a$.

Besides, for all $l = 0, 1, \cdots,$

$$\frac{1}{2\pi}\int_{-a}^{a} e^{-ixy} \int_{-a}^{a} e^{i\xi y} \psi_l(\xi) d\xi = \int_{-a}^{a} \frac{\sin a\,(x-\xi)}{\pi(x-\xi)}\,\psi_l(\xi)\,d\xi = \gamma_l \psi_l.$$

This means that ψ_l are eigenfunctions of the operator $\hat{G}_a : \mathcal{L}_2(I_a) \to \mathcal{L}_2(I_a)$,

$$\hat{G}_a[g](x) := \int_{-a}^{a} \frac{\sin a\,(x-y)}{\pi(x-y)}\,g(y)\,dy, \quad g \in \mathcal{L}_2(I_a), \tag{6}$$

whereas γ_l are the corresponding eigenvalues, here γ_l are the same as defined above.

Like Eq. (5), the latter equation remains valid outside the interval I_a. Note that $\hat{G}_a = \hat{F}^{-1} \circ \hat{\chi}_a \circ \hat{F} \circ \hat{\chi}_a$. In other words, $\hat{G}_a$ is the truncated direct Fourier transforms followed by the truncated inverse Fourier transform. The operator $\hat{G}_a$ plays a key role in the further consideration.

Remarkable properties of the TFT eigenfunctions have been widely discussed, see e.g. [2], [4], [6], [7]. Below we shall cite only those properties which are important for the present analysis. Among others, of special interest is the concentration property of the TFT eigenfunctions [2], namely that in the Paley-Wiener space $\mathcal{PW}_a$ the function ψ_0 is the most concentrated inside the interval I_a, since $\gamma_0 = \dfrac{\|\psi_0\|_{I_a}}{\|\psi_0\|_{\mathbb{R}}}$ yields the largest possible value for the ratio $\dfrac{\|f\|_{I_a}}{\|f\|_{\mathbb{R}}}$. In general, denote by $\mathcal{PW}_a^l$ the orthogonal complement to $\text{Span}\{\psi_0, \psi_1, \ldots, \psi_{l-1}\}$ in $\mathcal{PW}_a$, then

$$\gamma_l \;=\; \frac{\|\psi_l\|_{I_a}}{\|\psi_l\|_{\mathbb{R}}} \;=\; \max_{f \in PW_a^l} \frac{\|f\|_{I_a}}{\|f\|_{\mathbb{R}}}\,.$$

An exceptional feature of eigenvalues γ_l is that γ_l are close either to zero or to one, and the number of γ_l close to one does not exceed $L = \left[\dfrac{2\,a^2}{\pi}\right]$ [6].

Thus, one can see a qualitative difference between the TFT eigenfunctions of indices $l < \dfrac{2\,a^2}{\pi}$ and those of $l > \dfrac{2\,a^2}{\pi}$:

although each ψ_l is the most concentrated among functions from $\mathcal{PW}_a^l$, only the first $\left[\frac{2\,a^2}{\pi}\right]$ are *really* concentrated on I_a. The integral equation (5) does not account for this difference. In order to understand this feature, we recall that after an appropriate scaling TFT eigenfunctions coincide with the prolate spheroidal wave functions of zero order [2], and they are therefore often referred to as prolates.

IV. Prolate spheroidal wave functions

A representative overview of the prolate spheroidal wave functions (PSWF) is given in [8] (see also the literature cited therein). At any point $\xi \neq \pm 1$, a PSWF of zero order $S(c,\xi) = S(\xi)$ obeys the prolate spheroidal wave equation

$$\frac{d}{d\xi}(1-\xi^2)\frac{d}{d\xi}S + \left[\lambda + c^2(1-\xi^2)\right] S = 0, \qquad (7)$$

remaining bounded at the singular points $\xi = \pm 1$,

$$|S(\xi)| < \infty,\ \xi \to \pm 1. \qquad (8)$$

Both singularities at the points $\xi = \pm 1$ are regular and limit-point. In the neighborhood of $\xi = 1$ Eq. (7) has two linearly independent solutions [8]

$$S^{(1)}(\xi) \sim \text{const}, \quad S^{(2)}(\xi) \sim \ln(1-\xi^2), \quad \xi \to 1,$$

of which only the first one is bounded.

Solutions of Eq. (7) which are bounded at both singular points simultaneously exist not for all λ. Eq. (7) and the boundedness conditions (8) define a self-adjoint singular Sturm–Liouville eigenvalue problem on the interval $(-1,1)$. The eigenfunctions $S_l(\xi)$ of this problem are called angular PSWF. They are ordered by the number of internal zeros and normalized by the condition $\|S_l\|_{(-1,1)} = \int\limits_{-1}^{1} S_l^2(\xi)\,d\xi = 1$. For the associated eigenvalues one can prove that

$$-c^2 < \lambda_0 < \lambda_1 < \ldots < \lambda_l < \ldots .$$

At infinity any solution of Eq. (7) vanishes as $(1/\xi)$. In particular, solutions bounded at $\xi = 1$ enjoy the asymptotic behaviour

$$S_l(\xi) = \frac{A_l}{c\xi}\cos\left(c\xi - \frac{l+1}{2}\pi\right) + \mathrm{O}\left(\frac{1}{\xi^2}\right), \quad \xi \to \infty. \quad (9)$$

In what follows A_l is chosen to match at $\xi = \pm 1$ the angular function. A simple relation links then the functions S_l and ψ_l:

$$a = \sqrt{c}, \quad \psi_l(a,x) = \frac{1}{\sqrt{a}}\,S_l\left(c, \frac{x}{a}\right), \quad l = 0, 1, \cdots . \quad (10)$$

This means that the prolates ψ_l and the $sinc$-function translates have the same rate of vanishing at infinity, which seems to contradict the concentration property of prolates. However Eq. (7) gives us a key to eliminating the apparent contradiction.

As was shown in [9], the properties of solutions of Eq. (7) depend dramatically on whether λ is positive or negative. Below we add to the detailed analysis provided in [9] a few more features explaining the concentration phenomenon. To this end, we study the behavior of a bounded solution of Eq. (7) inside the interval where the product of $(1-\xi^2)$ and the potential $Q(\xi) = \lambda + c^2(1-\xi^2)$ is negative. For $-c^2 < \lambda < 0$ this is the interval $(\xi_T, 1)$, while for $\lambda > 0$ it is $(1, \xi_T)$, where $\xi_T = \sqrt{1+\lambda/c^2}$ being the turning point of Eq. (7).

The following lemma is easy to prove.

Lemma 1. *Let $-c^2 < \lambda < 0$ and ξ_T be the turning point of Eq. (7), so that $Q(x) < 0$ on the interval $(\xi_T, 1)$. If $S(\xi)$ is a solution of Eq. (7) bounded at $\xi = 1$, then neither $S(\xi)$ nor $S'(\xi)$ vanish inside the interval $(\xi_T, 1)$.*

Proof: On integrating Eq. (7) multiplied by $S(\cdot)$ over an interval $(\xi, 1)$, one obtains:

$$(1-\xi^2)S'(\xi)S(\xi) = \int\limits_{\xi}^{1}\left\{Q(\eta)S^2(\eta) - (1-\eta^2)\left[S'(\eta)\right]^2\right\} d\eta.$$

The right hand side above is strictly negative on $(\xi_T, 1)$. ∎

As is readily seen, the logarithmic derivative of a bounded solution, $\beta(\xi) = S'(\xi)/S(\xi)$, satisfies the equation

$$(1-\xi^2)\,\beta' = -Q(\xi) + 2\,\xi\beta - \beta^2\,(1-\xi^2), \quad \xi \in (\xi_T, 1). \quad (11)$$

Besides, the expansion $\beta(\xi) = \lambda/2 + (c^2 + \lambda/2 + \lambda^2/4)(1-\xi)/2 + \ldots$ holds near the point $\xi = 1$ [10]. Straightforward but rather tiresome analysis of the direction field in (11) shows that for $\xi \in (\xi_T, 1)$

$$\beta(\xi) < \frac{Q(\xi)}{\xi + \sqrt{\xi^2 - (1-\xi^2)\,Q(\xi)}} < \frac{\lambda + c^2(1-\xi^2)}{1+\sqrt{1-\lambda}} < 0.$$

This means that $S(\xi)$ decays exponentially fast in $(\xi_T, 1)$. Thus, the smaller the index l of the eigenvalue λ_l, the higher the ratio

$$\frac{S_l(\xi_T)}{S_l(1)} = -\exp\left\{\int_{\xi_T}^{1}\beta_l(\xi)\,d\xi\right\},$$

hence the smaller the factor A_l in (9) and therewith the smaller the contribution from outside the interval$(-1,1)$ to the total norm $\|S_l\|_{\mathbb{R}}$.

Similar analysis done for $\lambda > 0$ shows that the factor A_l grows up with l in accordance with the exponential increase of S_l on $(1, \xi_T)$. As a result, the contribution from the interval $(-1,1)$ to $\|S_l\|_{\mathbb{R}}$ becomes negligibly small as $l \to \infty$.

Note that the number of negative eigenvalues λ_l of the problem (7)–(8) was proved in [9] not to exceed $2c/\pi = 2a^2/\pi$.

V. Walter-Shen Sampling formula and the range of the operator $\hat{G}_a$

In terms of prolates the sampling formula becomes [4]

$$\begin{aligned} f(x) &= \frac{\pi}{a}\sum_{k=-\infty}^{\infty} f\left(\frac{\pi k}{a}\right)\sum_{l=0}^{\infty}\gamma_l\,\psi_l\left(\frac{\pi k}{a}\right)\psi_l(x) \\ &= \frac{\pi}{a}\sum_{l=0}^{\infty}\gamma_l\left\{\sum_{k=-\infty}^{\infty} f\left(\frac{\pi k}{a}\right)\psi_l\left(\frac{\pi k}{a}\right)\right\}\psi_l(x). \quad (12) \end{aligned}$$

Here, the order of double summation is interchangeable, both double series converging in $\mathcal{L}_2$-norm and uniformly on $\mathbb{R}$.

Because of the double summation, the sampling formulae (12) look more cumbersome than (2), however in practical calculations series (12) may be more advantageous than the classical one. Moreover, the following estimate, similar to that of Lemma 4 in [4] (see also [3]), allows one to truncate the summation on k to a few terms:

$$\Theta_l := \gamma_l \sum_{|k|>a^2/\pi} \psi_l^2\left(\frac{\pi k}{a}\right) \le C\sqrt{\gamma_l(1-\gamma_l)}. \qquad (13)$$

One sees that the contribution from samples $\psi_l(\pi k/a)$, $|k| > a^2/\pi$, is small both for $l < 2a^2/\pi$ and for $l > 2a^2/\pi$.

Consider the range of the operator $\hat{G}_a$, $\mathrm{Rg}(\hat{G}_a)$. Clearly, prolates ψ_l are in $\mathrm{Rg}(\hat{G}_a)$.

Let $f(x) \in \mathrm{Rg}(\hat{G}_a)$, i.e. $f(x) = \int_{-a}^{a} \frac{\sin a\,(x-y)}{\pi(x-y)}\, g(y)\, dy$ for some $g \in \mathcal{L}_2(I_a)$. Denote the Fourier coefficients of the functions f and g in the basis of prolates by $\tilde{f}_l$ and $\tilde{g}_l$, respectively. Then the truncation error caused by neglect of the contribution from ψ_l at $l > L$ is

$$\varepsilon_L := \Big\| f - \sum_{l\le L} \tilde{f}_l \psi_l \Big\|_{\mathbb{R}} \le \sqrt{\gamma_{L+1}} \Big\| g - \sum_{l\le L} \tilde{g}_l \psi_l \Big\|_{I_a},$$

which is very small, provided that $L > 2a^2/\pi$.

In view of (13), the truncation of the inner sum in (12) at some $K > a^2/\pi$ causes the error

$$\begin{aligned}\varepsilon_{L,K}^2 &:= \Big\| f_L(x) - \frac{\pi}{a} \sum_{l\le L} \sum_{|k|\le K} \gamma_l\, f\left(\frac{\pi k}{a}\right)\, \psi_l\left(\frac{\pi k}{a}\right)\, \psi_l(x) \Big\|_{\mathbb{R}}^2 \\ &\le \sum_{|k|>K} \left[f\left(\frac{\pi k}{a}\right)\right]^2 \sum_{l\le L} \Theta_l.\end{aligned}$$

Summarizing, we conclude that the NDF of functions in $\mathrm{Rg}(\hat{G}_a)$ is $[2a^2/\pi] + 1$.

VI. Sampling in $\mathrm{Rg}\,(G_{\Omega,T})$

The range of the operator $\hat{G}_a$ is not the only class of band-limited functions for which the above truncation error estimates hold. Consider the operator $G_{\Omega,T}$:

$$\begin{aligned} G_{\Omega,T}\,[g]\,(x) &:= \frac{1}{2\,\pi} \hat{F}^{-1} \circ \hat{\chi}_\Omega \circ \hat{F} \circ \hat{\chi}_T\ [g]\,(x) \\ &= \int_{-T}^{T} \frac{\sin\,\Omega\,(x-y)}{\pi(x-y)}\, g(y)\, dy. \end{aligned}$$

On substituting into the above equation new variables $\eta = \sqrt{\Omega/T}\, y$ and $\xi = \sqrt{\Omega/T}\, x$, we obtain

$$\tilde{f}(\xi) = f\left(\sqrt{\frac{\Omega}{T}}\xi\right) = \int_{-a}^{a} \frac{\sin\,a\,(\xi-\eta)}{\pi(\xi-\eta)}\, g\left(\sqrt{\frac{T}{\Omega}}\eta\right)\, d\eta\,,$$

where $a^2 = \Omega\, T$. As a result, the function $\tilde{f}(\xi) \in \mathrm{Rg}\,(G_a)$ and has $[2\Omega\, T/\pi] + 1$ degrees of freedom.

The convolution of two TFT eigenfunctions $\Phi_{nm}(x) := \int_{-\infty}^{\infty} \psi_n(a, x-y)\, \psi_m(a, y)\, dy$ is a–band–limited and hence

$$\Phi_{nm}(x) \approx \frac{\pi}{a} \sum_{l\le L,\ |k|\le K} \gamma_l\,(a)\, \Phi_{nm}\left(\frac{\pi k}{a}\right)\, \psi_l\left(a, \frac{\pi k}{a}\right) \psi_l(a, x)\,.$$

On the other hand, one can prove that $\Phi_{nm}(x) \in \mathrm{Rg}\,(G_{a,2a})$. Therefore $\tilde{\Phi}_{nm}(x) = \Phi_{nm}(x/\sqrt{2}) \in \mathrm{Rg}\,\big(G_{\sqrt{2}a}\big)$ and

$$\begin{aligned}\Phi_{nm}(x) \;\approx\; & \frac{\pi}{\sqrt{2}a} \sum_{l,|k|\le 4a^2/\pi} \gamma_l\left(\sqrt{2}a\right) \Phi_{nm}\left(\frac{\pi k}{2a}\right) \\ & \times \quad \psi_l\Big(\sqrt{2}a, \frac{\pi k}{\sqrt{2}a}\Big)\, \psi_l\Big(\sqrt{2}a, \sqrt{2}x\Big)\,.\end{aligned}$$

The latter sampling formula shows better accuracy than the previous one, even if the number of samples and prolates involved in calculations is the same.

VII. Generalization to higher dimensions

In [12] the eigenfunctions of the 2D Fourier transform truncated to a circle of finite radius a were represented through the eigenfunctions of the truncated Hankel transforms (THT) of different angular numbers m. Recently in [11] the number of THT eigenvalues close to one was proved not to exceed $a^2/\pi - m/2$. The same number defined the NDF of a class of Hankel-band-limited functions analogous to $\mathrm{Rg}\big(\hat{G}_a\big)$. This results can easily be generalized to the case of higher dimensions in accordance with the discussion in [13].

References

[1] J. R. Higgins, *Five short stories about the cardinal series*, Bull. Amer. Math. Soc. (N.S.) **12**(1) (1985) 45–89.

[2] D. Slepian, H. O. Pollak, *Prolate spheroidal wave functions, Fourier analysis and uncertainty, I*, Bell Syst. Tech. J. **40**(1) (1961) 43–64.

[3] J. A. Hogan and J. D. Lakey, *Duration and Bandwidth Limiting: Prolate Functions, Sampling, and Applications*, Birkhäuser, Boston, 2012.

[4] G. G. Walter, X. Shen, *Sampling With Prolate Spheroidal Wave Functions*, Sampl. Theory Signal Image Proc. **2**(1) (2003) 25–52.

[5] T. Levitina and E. J. Brändas, *Filter diagonalization: Filtering and postprocessing with prolates*, Comp. Phys. Comm., **180**(9) (2009) 1448–1457.

[6] H. J. Landau, *Sampling, data transmission, and the Nyquist rate*, Proc. IEEE **55** (1967) 1701–1706.

[7] H. Xiao, V. Rokhlin, and N. Yarvin, *Prolate Spheroidal Wave Functions, Quadrature and Interpolation*, Inverse Problems **17** (2001) 805–838.

[8] I. V. Komarov, L. I. Ponomarev, and S. Yu. Slavyanov, *Spheroidal and Coulomb Spheroidal Functions,* [in Russian], Nauka, Moscow, 1976.

[9] A. Osipov and V. Rokhlin, *Detailed Analysis of Prolate Quadratures and Interpolation Formulas*, arXiv:1208.4816 [math.NA]

[10] T. V. Levitina and E. J. Brändas, *Computational techniques for Prolate Spheroidal Wave Functions in Signal Processing*, J. Comp. Meth. Sci. & Engrg. **1** (2001) 287–313.

[11] T. Levitina, *On the Eigenfunctions of the Finite Hankel Transform*, Sampl. Theory Signal Image Proc. **11**, (2012), 55–79.

[12] D. Slepian, *Prolate spheroidal wave functions. Fourier analysis and uncertainty, IV: Extensions to many dimensions; generalized prolate spheroidal functions*, Bell Sys. Tech. J. **43** (1964) 3009–3058.

[13] O. Brander and B. DeFacio, *A generalisation of Slepian's solution for the singular value decomposition of filtered Fourier transforms*, Inverse Problems **2** (4) (1986) 375–393.

Tracing Sound Objects in Audio Textures

Monika Dörfler
University of Vienna
Faculty of Mathematics
NuHAG
Email: monika.doerfler@univie.ac.at

Ewa Matusiak
University of Vienna
Faculty of Mathematics
NuHAG
Email: ewa.matusiak@univie.ac.at

***Abstract*—This contribution presents first results on two proposed methods to trace *sound objects* within *texture sounds*. We first discuss what we mean by these two notions and explain how the properties of a sound that is known to be textural are exploited in order to detect changes which suggest the presence of a distinct sound event. We introduce two approaches, one is based on Gabor multipliers mapping consecutive time-segments of the signal to each other, the other one on dictionary learning. We present the results of simulations based on real data.**

I. Introduction

Sound signals play a central role in human life and the manner sound is perceived is highly sophisticated, complex and context-dependent. In some applications, one may be interested in distinguishing between what may be called a "sound object" and more textural sound components constituting an acoustical background. The notion of sound object ("objet sonore") was introduced by Pierre Schaeffer [10] as a generalization of the concept of a musical note, in particular their definition implies a time-limitation of sound objects.

Human listeners tend to perceive sound in a structured manner, with the ability to focus and de-focus. Whether a particular event is experienced as a relevant sound structure as opposed to background, textural sound, seems to depend both on cultural and educational background, cp. [5], that may be shared by a group of listeners. From a certain point of view, the perception of sound components as background (textural) sound or object (compactly structured) sound, depends on the "zoom" the listener wishes to adopt or unconsciously assumes. In this contribution, we attempt to mimic these observations in a technical way, by "defining" a sound to be textural if it does not change certain characteristics which are first to be determined from a certain amount of data. In that sense, we need the a priori knowledge that a particular part of a signal represents textural sound segments. Any signal components representing a significant change are then considered to be new objects in the sense of not belonging to the previous texture sound or background. By definition, a characterizing feature of texture sounds, in particular as opposed to the signal components we would like to call sound objects, is some kind of stationarity over an extended period of time; while micro-changes are always present, the listener integrates them as part of the texture, at least after some time has passed. Therefore, any two sufficiently long slices of a pure texture sound can, and should, be assumed to be correlated. This observation leads us to the following approach: given a signal which is known to present a texture sound, we learn its inherent characteristics. Using the information gained from the learning step, we can then look for significantly different, hence salient, signal components, which we then define to represent a sound object.

For both the learning and the observation period, we divide the signal into overlapping time-slices. Then, during observation, we look for substantial changes from one part of the signal to another, which would indicate the presence of a sound object. We are going to quantify, what we mean by substantial changes, by means of two technical tools: sparsity in an appropriate dictionary and similarity of Gabor transforms. Based on these two tools, we introduce two methods to scan texture signals for the presence of what may be conceived as sound objects. While the proposed framework may also be useful for the task of detecting audio events, this application is not the primary motivation for our study. The latter, challenging task addressed in the framework of CASA [1] requires a much wider and more elaborate evaluation stage and is beyond the scope of the current contribution. Here, we are primarily motivated by a different challenge, which parallels the cognitive process sketched above: we mimic a situation in which a user/listener makes real-time decisions about the property of an event occurring in the signal to be or not to be a sound object which deserves attention. We divide the signal into overlapping slices. In the first approach we propose, we make use of Gabor transforms; more precisely, the variations of the Gabor coefficients between different slices of the signal are tracked by investigating corresponding Gabor multipliers.

The second proposed method is by means of the exploitation of sparsity constraints via dictionary learning. Given a part of a texture sound which is known to be free of sound objects, we learn a dictionary such that each slice admits a sparse approximate representation in that dictionary. We then scan the signal piece by piece by checking its reconstruction error with respect to the corresponding dictionary, in order to detect in which intervals of time an object may occur.

The two methods and the involved tools are presented in the next section. Then, some promising results of preliminary simulations are presented in Section III and we conclude with a short discussion and perspectives.

[1] Computational Auditory Scene Analysis, cp. [13].

II. Technical tools for sound object tracing

The proposed methods aim at deciding about the presence of distinct sound objects within a signal whose first section is known to be textural. Both methods give a decision about the presence or absence of an object within a given slice of the signal. This can be seen as a first step in exact object localization in terms of precise onset and offset times, and later extraction. We will see in Section III that the proposed methods are designed particularly for longer signals and should be applicable to online-applications.

Before describing the two methods in detail we recall some definitions from Gabor analysis and fix notation. We will be working with square integrable functions $L^2(\mathbb{R})$, with norm $\|\cdot\|_2$ induced by an inner product $\langle f, g\rangle = \int_{\mathbb{R}} f(t)\overline{g(t)}\,dt$, $f, g \in L^2(R)$. For $f \in L^2(\mathbb{R})$ and $\omega, \tau \in \mathbb{R}$, the operators $M_\omega f(t) = e^{2\pi i\omega t} f(t)$ and $T_\tau f(t) = f(t-\tau)$ are called frequency and time shift operators, respectively. A collection $\mathcal{G}(g, a, b) = \{g_{k,l} := M_{bl}T_{ak}g\}_{k,l\in\mathbb{Z}}$ is called a Gabor frame for $L^2(\mathbb{R})$ if the operator $S_{g,g}$

$$S_{g,g}f = \sum_{k,l\in\mathbb{Z}} \langle f, g_{k,l}\rangle g_{k,l} \qquad \text{for all} \quad f \in L^2(\mathbb{R}) \tag{1}$$

is bounded and invertible on $L^2(\mathbb{R})$.[2].
For every frame $\mathcal{G}(g, a, b)$ there exists a function γ, called dual window, such that $\mathcal{G}(\gamma, a, b)$ is again a frame, called dual Gabor frame, and $f = S_{g,\gamma}f = S_{\gamma,g}f$ for all $f \in L^2(\mathbb{R})$.

Let $f \in L^2(\mathbb{R})$ be a background, texture signal. We divide it into overlapping slices f_i, $i \in \mathbb{Z}$, in the following way: $f_i(t) = f(t)$ for $t \in [\alpha_i, \beta_i]$ with $\alpha_{i-1} < \alpha_i \leq \beta_{i-1}$ and $\alpha_{i+1} \leq \beta_i < \beta_{i+1}$.

A. Gabor Multipliers

We first describe the method based on Gabor multipliers. This method not only allows to detect a change but also gives more information on the time-frequency location of a potential object. Let $\mathcal{G}(g, a, b)$ be a Gabor frame and $\mathcal{G}(\gamma, a, b)$ its dual frame. Let $\mathbf{m} = \{m_{k,l}\}_{k,l\in\mathbb{Z}}$ be a bounded complex-valued sequence. Then the Gabor multiplier associated to (g, γ, a, b) with symbol, or mask, $\mathbf{m}$ is given by

$$G_{\mathbf{m}}f = \sum_{k,l\in\mathbb{Z}} m_{k,l}\langle f, g_{k,l}\rangle\gamma_{k,l}\,. \tag{2}$$

The operator $G_{\mathbf{m}}$ is well defined and bounded on $L^2(\mathbb{R})$ [4].

In [8] the authors addressed the problem of transforming one signal into another by means of linear operators. They focus on Gabor multipliers as the transforming operators. More precisely, for two signals f_1 and f_2, given dual frames $\mathcal{G}(g, a, b)$ and $\mathcal{G}(\gamma, a, b)$, the objective is to find a symbol $\mathbf{m}$ such that the Gabor multiplier $G_{\mathbf{m}}$ takes f_1 into f_2 subject to certain constraints on the mask $\mathbf{m}$. The constraints on the mask can be sparsity in time-frequency plane or total energy.

[2]Note that the coefficients $\langle f, g_{k,l}\rangle$ in $S_{g,g}$ are samples of a short-time Fourier transform of f at sampling points (ak, bl).

An optimal mask, subject to given constraints is a solution to the following minimization problem

$$\min_{\mathbf{m}} \|f_1 - G_{\mathbf{m}}f_2\|_2^2 \quad \text{subject to} \quad d(\mathbf{m}) < \epsilon\,, \tag{3}$$

where d we can chosen to be, for example $d(\mathbf{m}) = \lambda\||\mathbf{m}| - 1\|_1$, to promote sparsity, or $d(\mathbf{m}) = \lambda\|\mathbf{m} - 1\|_2^2$ to control total energy, where λ is a sparsity prior tuning the influence of the second term in (3).

For texture sounds, the slices f_i, as defined in the previous section, are similar, hence also their Gabor transforms. The grade of similarity is learned from the first part of the signal, which is known to be textural. Then, a symbol $\mathbf{m}_i$ of a Gabor multiplier transforming f_i to f_{i+1}, $f_{i+1} = G_{\mathbf{m}_i}f_i$, is close to one, or in other words $d(\mathbf{m}_i)$ is close to zero. During the learning phase, the parameter λ should be tuned to yield small deviations from the constant mask $\mathbf{m} = 1$.
Now, the problem of detecting a sound object versus a stationary background is based on studying masks $\mathbf{m}_i$. If $\mathbf{m}_i$ is significantly different from 1, or $d(\mathbf{m}_i) > \epsilon$ for some chosen $\epsilon > 0$, then the slices f_i and f_{i+1} differ significantly which leads us to assuming the presence of an object in slice f_{i+1}.
a

B. Dictionary Learning with Sparsity Prior

Given a dictionary $\mathbf{D} \in \mathbb{C}^{K\times L}$, $K < L$ and a signal $f \in \mathbb{C}^K$, we say that f admits an $S-$sparse approximation over $\mathbf{D}$ if one can find an approximation of f by S atoms from $\mathbf{D}$. In other words, we are looking for coefficients $x \in \mathbb{C}^L$, such that

$$f \approx \mathbf{D}x \quad \text{while} \quad \|x\|_0 \leq S. \tag{4}$$

Here, $\|\cdot\|_0$ is a pseudo-norm counting the non-zero entries in x. Finding the best solution to (4) is an NP-hard problem; however by relaxing the counting pseudo-norm to an ℓ_1 norm, it becomes a convex optimization problem that can be tackled with many existing efficient algorithms, such as basis pursuit (BP) [2], orthogonal matching pursuit (OMP) [12] or FOCUSS [9]. A dictionary yielding sparse approximate representation for a class of signals can be learned from a sufficient number of data samples. Let F be a set of N signals $f_i \in \mathbb{C}^K$, collected into a matrix of size $K \times N$, for which one would like to find a dictionary such that each signal in the group admits an $S-$sparse approximate representation. A dictionary with the desired properties can be built by finding a solution to the following minimization problem

$$\min_{\mathbf{X},\mathbf{D}} \sum_{i=0}^{N-1} \|f_i - \mathbf{D}x_i\|_2^2 \quad \text{subject to, for every } i \quad \|x_i\|_0 \leq S\,, \tag{5}$$

where $\mathbf{X} \in \mathbb{C}^{L\times N}$ is the matrix of coefficients $x_i \in \mathbb{C}^L$. Among many algorithms addressing the problem of dictionary learning are K-SVD [1], maximum likelihood methods [7] or the MOD method [3].

For a given texture sound f, we observe the first couple of seconds of the signal and learn a dictionary which gives a sparse approximate representation thereof. We build the

training data set F by considering slices of first L samples of f, each of length K with $M \geq 0$ samples of overlap, i.e. $f_i(k) = f(i(K-M)+k)$ where $k = 0, \ldots, K-1$ and $i = 0, \ldots, N-1$. Then, assuming ongoing textural characteristics of f, the slices f_i for $i \geq N$ also admit sparse approximate representation in the same dictionary while no significant changes occur. In detail, let $\epsilon > 0$ be given. If it is possible to find a vector x_i of coefficients such that $\|f_i - Dx_i\|_2 \leq \epsilon$ while $\|x_i\|_0 = S$ is satisfied, then we conclude no presence of a sound object. However, if the above relation is violated, we can assume additional components in f_i that are not correlated with elements of $\mathbf{D}$. We scan the signal f slice by slice and verify its representation in $\mathbf{D}$.

III. Simulations

We present numerical results based on two classes of texture sounds f: (heavy) rain and washing machine noise. In order to give a proof of concept, we apply the suggested methods to finding synthetic signals s which unambiguously qualify as sound objects within the background signals; we use damped sums of six different harmonics of 0.5 seconds length. The SNR[3] of the objects present in the texture sound is between $-5dB$ and $-7.5dB$. Note that the sound-files corresponding to the examples as well as supplementary examples, codes and extensions are available at the website homepage.univie.ac.at/monika.doerfler/SoundObj.html.

A. Gabor Multiplier

For the Gabor multiplier approach, we choose slices of approximately half a second (20480 samples) length with 75% overlap. We use a standard tight Gabor frame with a Hann window of length 1024 and 75% overlap. The spectrogram of the test signal is depicted in the upper plot of Fig. 1. The three harmonic and compactly supported synthetic signals are clearly visible. The lower plot shows the deviation $\||\mathbf{m}|-1\|_1$ for the mask corresponding to the transition between two time-slices. Based on the first, purely textural part of the signal, λ is tuned in order to allow only negligible deviation of the absolute value of $\mathbf{m}$ from 1. During our experiments, it turned out that the success depends heavily on an appropriate choice of λ, which was chosen to be 1.2 in the first example.

The second example, the distinct noise produced by a washing machine, is a more complex texture sound. Here, the situation is more difficult, since the "stationarity" of the texture is present on a larger scale, as visible in its spectrogram, shown in Figure 2, upper display. For this example, we had to allow for a much smaller $\lambda = 0.01$, i.e. for significant deviations from a constant mask, in oder to obtain meaningful results. Therefore, as opposed to the previous example, we obtain much higher values of the deviation $\||\mathbf{m}|-1\|_1$ also for the textural part. In Figure 3, we show two masks occurring in the investigation of this example; it is clearly visible, that this particular signal contains a lot of energy in low frequency

[3]We define the signal to noise ratio (SNR) by $SNR_{dB} = 10\log_{10}(\|s\|_2^2/\|f\|_2^2)$, given in dB, by where f is the background signal, which can be seen as "noise" in which s, the sound object is to be traced.

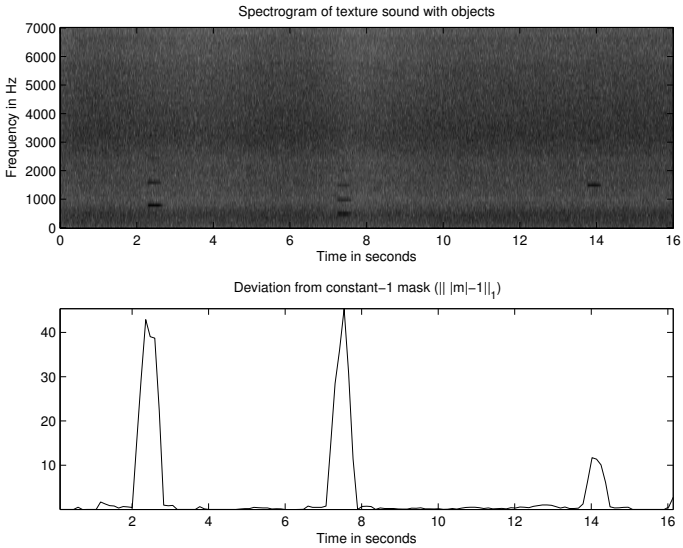


Fig. 1. Detection of sound objects in background noise (Rain) using Gabor multipliers approach.

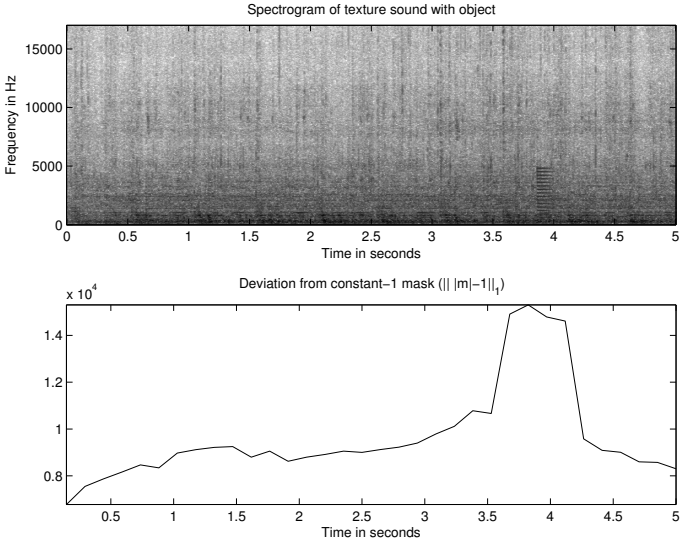


Fig. 2. Detection of a sound object in background noise (Washing machine) using Gabor multipliers approach.

bands, with a certain periodicity (also audible in the signal). It is quite obvious that, without taking these changes of energy into account, no meaningful transition can be expected. On the other hand, inspection of the part of the mask that is related to the sound object has a clear local persistence in time which the texture part lacks, but which is typical for harmonic signals. It is planned to exploit this kind of a priori knowledge - or assumption - about the objects one is interested in, in order to improve the method's success and reliability. In particular, the models introduced as structured or social sparsity, cp. [6], [11], show promising results in first experiments and will be further exploited.

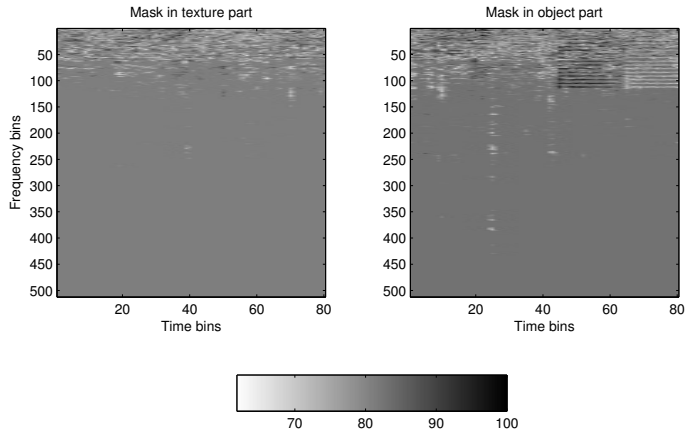


Fig. 3. Change of consecutive masks: between texture slices and between object slices; values are in dB.

B. Sparse Dictionary Representation

We applied the second, dictionary-based method to the signals presented in the previous section; we chose time-slices of 256 samples length and 50% overlap. It turned out during the experiments, that, in the evaluation step, smaller overlap is possible and does not deteriorate the results, since the time-resolution given by the slice-length of about 6ms is fine enough. The resulting evaluation criteria, namely approximation error for a maximum number of atoms and level of sparsity for a chosen error tolerance, are shown in Figure 4. Obviously, both criteria show significant deviation from the texture level during the duration of the sound objects. It should be noted that the amplitude of the time-signals don't visibly increase during the sound objects, also cf. homepage.univie.ac.at/monika.doerfler/SoundObj.html to listen to the audio files.

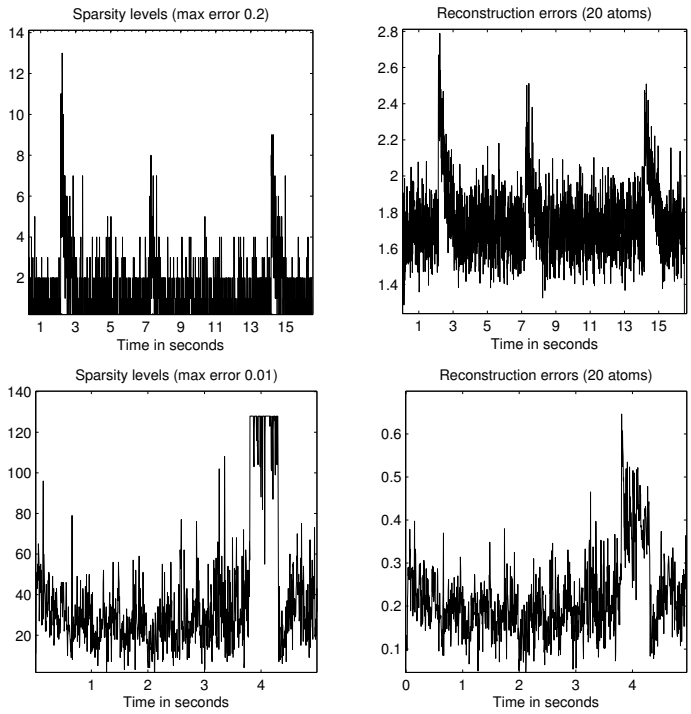


Fig. 4. Tracing sound objects by sparse dictionary representation. Top row: rain signal; bottom row: washing machine signal.

IV. Discussion and Perspectives

We presented two methods for sound object tracing in background, texture signals. Both methods exploit the assumed quasi-stationary character of texture signals and decide that a 'foreign' sound object should be present, if that stationarity is lost. The suggested methods and numerical experiments need to be extended to a much larger samples of both texture sounds and sound objects in order to draw reliable conclusions about the situations in which the proposed models give satisfactory results; furthermore, there are several open questions as to how long the slices of the signal should be, in both the sparsity method and Gabor multipliers, and what kind of Gabor frames to choose for the latter approach. These questions will be investigated in detail in ongoing work on the topic and results will be presented on the companion website.

Acknowledgments

This research was supported by the WWTF project Audio-Miner (MA09-024). The authors would like to thank Anaïk Olivero for sharing with us a source code for computing Gabor masks and Richard Kronland-Martinet and his team in LMA, CNRS Marseille, for giving us permission to use their software SPAD to synthesize texture sounds.

References

[1] M. Aharon, M. Elad, and A. Bruckstein. The K-SVD: An algorithm for designing of overcomplete dictionaries for sparse representation. *IEEE Trans. Signal Process.*, 54(11):4311–4322, 2006.

[2] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.*, 20(1):33–61, 1998.

[3] K. Engan, S. O. Aase, and J. H. Hakon Husoy. Method of optimal directions for frame design. *IEEE Int. Conf. Acoust.,Speech, Signal Process.*, 5:2443–2446, 1999.

[4] H. G. Feichtinger and K. Nowak. A first survey of Gabor multipliers. In H. G. Feichtinger and T. Strohmer, editors, *Advances in Gabor Analysis*, Appl. Numer. Harmon. Anal., pages 99–128. Birkhäuser, 2003.

[5] V. Klien, T. Grill, and A. Flexer. On Automated Annotation of Acousmatic Music. *Journal of New Music Research*, 41(2):153–173, 2012.

[6] M. Kowalski, K. Siedenburg, and M. Dörfler. Social Sparsity! Neighborhood Systems Enrich Structured Shrinkage Operators. *IEEE Trans. Signal Process.*, 61(10):2498 – 2511, 2013.

[7] M. S. Lewicki and T. J. Sejnowski. Learning overcomplete representations. *Neural Comp.*, 12:337365, 2000.

[8] A. Olivero. *Les multiplicateurs temps-fréquence. Applications à l'analyse et à la synthèse de signaux sonores et musicaux*. PhD thesis, 2012.

[9] B. Rao and K. Kreutz Delgado. An affine scaling methodology for best basis selection. *IEEE Trans. Signal Process.*, 47:187–200, 1999.

[10] P. Schaeffer. *On Automated Annotation of Acousmatic Music*. Editions du Seuil, Paris, France, 2002.

[11] K. Siedenburg and M. Dörfler. Persistent Time-Frequency Shrinkage for Audio Denoising. *J. Audio Eng. Soc.*, 61(1/2), 2013.

[12] J. A. Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Trans. Inform. Theory*, 50(10):2231–2242, 2004.

[13] D. Wang and G. Brown, editors. *Computational auditory scene analysis: Principles, Algorithms, and Applications*. IEEE Press/Wiley-Interscience, 2006.

An Uncertainty Principle for Discrete Signals

Sangnam Nam
Aix Marseille Université, CNRS, Centrale Marseille, LATP, UMR 7353
13453 Marseille, France
Email: nam.sangnam@cmi.univ-mrs.fr

***Abstract*—By use of window functions, time-frequency analysis tools like Short Time Fourier Transform overcome a shortcoming of the Fourier Transform and enable us to study the time-frequency characteristics of signals which exhibit transient oscillatory behavior. Since the resulting representations depend on the choice of the window functions, it is important to know how they influence the analyses. One crucial question on a window function is how accurate it permits us to analyze the signals in the time and frequency domains. In the continuous domain (for functions defined on the real line), the limit on the accuracy is well-established by the Heisenberg's uncertainty principle when the time-frequency spread is measured in terms of the variance measures. However, for the finite discrete signals (where we consider the Discrete Fourier Transform), the uncertainty relation is not as well understood. Our work fills in some of the gap in the understanding and states uncertainty relation for a subclass of finite discrete signals. Interestingly, the result is a close parallel to that of the continuous domain: the time-frequency spread measure is, in some sense, natural generalization of the variance measure in the continuous domain, the lower bound for the uncertainty is close to that of the continuous domain, and the lower bound is achieved approximately by the 'discrete Gaussians'.**

I. Introduction

Fourier Transform, due to the fact that it is a global transform, is not well-suited for the analysis of signals that exhibit transient behavior. This is a rather significant drawback since such signals exist in abundance. One way to remedy this shortcoming is the use of window functions: a window function enables us to localize the function to some specific interval of interest that we want to look at. This gives rise to time-frequency analysis and makes it possible for us to study the frequency structure of functions at varying points in time. Just like Fourier analysis, time-frequency analysis is a fundamental tool in science, especially in signal processing.

In this article, we define the Fourier Transform $\hat{f}$ of a complex-valued function f defined on the real line $\mathbb{R}$ via

$$\hat{f}(\xi) := \int_{\mathbb{R}} f(t) e^{-2\pi i \xi t}\, dt, \quad \xi \in \mathbb{R}. \tag{1}$$

The Windowed Fourier Transform of f with a given window function $g : \mathbb{R} \to \mathbb{C}$ would then be defined as

$$\mathcal{V}_g(\tau, \xi) := \int_{\mathbb{R}} f(t) \overline{g(t-\tau)} e^{-2\pi i \xi t}\, dt, \quad \tau, \xi \in \mathbb{R}.$$

If g and $\hat{g}$ are supported near the origin, one may interpret that $\mathcal{V}_g f(\tau, \xi)$ is the 'ξ-frequency content of f at time τ'.

Unfortunately, such an ideal interpretation cannot become a reality; the well-known uncertainty principles expresse the idea that there is a fundamental limit on how g and $\hat{g}$ can be simultaneously localized in the two domains. The most famous formulation of the uncertainty principle is given by the Heisenberg-Pauli-Weyl inequality (see, e.g., [1]):

Theorem 1: For $f \in L_2(\mathbb{R})$, define the variance of f by

$$v_f := \min_{a \in \mathbb{R}} \frac{1}{\|f\|_2^2} \int_{-\infty}^{\infty} (t-a)^2 |f(t)|^2\, dt. \tag{2}$$

Then,

$$v_f v_{\hat{f}} \geq \frac{1}{16\pi^2}.$$

Equality holds if and only if f is a multiple of $\varphi_{a,b}$, defined by

$$\varphi_{a,b}(t) := e^{2\pi i b(t-a)} e^{-\pi (t-a)^2/c}$$

for some $c > 0$.

We may define the mean of f by

$$\mu_f := \operatorname*{argmin}_{a \in \mathbb{R}} \frac{1}{\|f\|_2^2} \int_{-\infty}^{\infty} (t-a)^2 |f(t)|^2\, dt.$$

Clearly, the smaller v_f is, the more concentrated the function f is around μ_f. In other words, v_f is a measure of time-spreading of f. Similarly, $v_{\hat{f}}$ is a frequency-spreading measure of f. Thus, the Heisenberg-Pauli-Weyl inequality expresses the intrinsic limit on how well an $L_2(\mathbb{R})$ function can be localized on the time-frequency plane. Moreover, the theorem also tells us what the minimizing functions are.

While the Heisenberg Uncertainty Principle gives us a clear picture of what can be achieved for time-frequency localization for the continuous functions defined on $\mathbb{R}$, our discussion so far is somewhat detached from reality; we can only consider functions defined on finite intervals in real life. Furthermore, in this day and age of computers, processing can be done only when the signal can be stored in memory. Therefore, the signals are discrete and finite.

A pertinent question is: what can be said about the uncertainty for the time-frequency analysis when the Discrete Fourier Transform is used? Is there any relation between the uncertainties for the continuous and the discrete cases? To our knowledge, surprisingly little is known for this problem, and this is the area that we aim to contribute to with our work.

II. Discrete Uncertainty Relations: Some Related Works

In this section, we discuss some works in the literature which may serve as an introduction to the problem that we are interested in.

A. Uncertainty for Continuous Functions Defined on the Circle

The Fourier series for periodic functions may be viewed as something intermediate between the continuous Fourier Transform for functions on the real line and the discrete Fourier Transform for finite signals. It could be a good starting point of our discussion on the uncertainty for discrete signals.

For a 2π-periodic function f, the Fourier coefficients for f is defined by

$$\hat{f}(k) := \frac{1}{2\pi}\int_0^{2\pi} f(t)e^{-ikt}\,dt, \quad k \in \mathbb{Z}.$$

Remarks on notations: For lightness, we will sacrifice the precision and use the same notation $\hat{f}$ to mean various different Fourier Transforms whose meaning will become clear depending on what f is. Such a convention extends to $\|\cdot\|$ as well. We also point out that the definition of the continuous Fourier Transform $\hat{f}$ used in this subsection is defined without the 2π-factor in (1).

The question we are interested in is how concentrated, or conversely how spread, f and $\hat{f}$ are. We note that even though $\hat{f}$ is a discrete sequence, there is no problem in defining the variance of it; we need only to replace the integral in (2) with an analogous sum. The mean $\mu_{\hat{f}}$ can be similarly defined. The situation is different for f. The issue is that we cannot simply compute

$$\frac{1}{\|f\|_2^2}\int_0^{2\pi} t|f(t)|^2\,dt$$

for the mean of f. Such a quantity fails to take the periodicity into account.

A different way to characterize 'the mean value' had been proposed (see [2]):

$$\tau(f) := \frac{1}{\|f\|_2^2}\int_0^{2\pi} e^{it}|f(t)|^2\,dt.$$

The periodicity is clearly reflected in $\tau(f)$. With that, one defines 'the variance' of f as

$$\frac{1}{\|f\|_2^2}\int_0^{2\pi} |e^{it}-\tau(f)||f(t)|^2\,dt = 1-\tau(f)^2.$$

With these time-frequency spread measures, the uncertainty relation for the continuous functions on the circle was shown to be as follows:

$$\left(1-\tau(f)^2\right)v_{\hat{f}} \geq \frac{\tau(f)^2}{4}. \tag{3}$$

Note that unlike in the continuous setting, the quantity on the right-hand side depends on the function f. Therefore, if we were to use $(1-\tau(f)^2)v_{\hat{f}}$ as the measure of uncertainty of f, the equality in (3) does not immediately imply that f is a minimizer of the uncertainty. A simple way to bypass this issue is to define the time spread of f as

$$v_f := \frac{1-\tau(f)^2}{\tau(f)^2}.$$

A more precise description of the resulting uncertainty principle is given as follows [3], [4]:

Theorem 2: For a function $f \in AC_{2\pi}$ with $f' \in L_2([0,2\pi])$ where f is not of the form ce^{ikt} for any $c \in \mathbb{C}$, $k \in \mathbb{Z}$, it holds that

$$v_f v_{\hat{f}} > \frac{1}{4}.$$

The lower bound is not attained by any function, but is best possible. Here, $AC_{2\pi}$ is the class of 2π-periodic absolutely continuous functions.

One reservation towards this result is that the meaning of the so-called *angular spread* v_f is not very intuitive. In addition, the theorem does not give any guide on what functions may have the uncertainty product close to the lower bound.

A result in [5] sheds some light on the second problem. The authors used a process of periodization and dilation to show that a sequence of functions achieve the uncertainty for functions defined on the real line in the limit. They proved:

Theorem 3: For an admissible function f (defined on the real line),

$$\lim_{a\to\infty}\frac{1}{a^2}v_{f_a} = v_f, \quad \lim_{a\to\infty} a^2 v_{\hat{f}_a} = v_{\hat{f}},$$

where

$$f_a(t) := \sqrt{a}\sum_{k\in\mathbb{Z}} f(a(t+2\pi k)).$$

Therefore,

$$\lim_{a\to\infty} v_{f_a}v_{\hat{f}_a} = v_f v_{\hat{f}}.$$

We remind the reader that the definitions of v_{f_a} and v_f are quite different.

Since the minimum of $v_f v_{\hat{f}}$ is known to be $1/4$ and is achieved by (essentially) Gaussian functions, the theorem provides a way to build periodic functions that are asymptotically optimal in the given measure of time-frequency spreads. We will see that our result shares some similarity with this.

Another way to obtain periodic functions which nearly achieve the uncertainty bound is by computing directly with numerical optimization [6]. In this approach, Parhizkar et al. fixed the angular spread at a prescribed level and then searched for functions that minimize the frequency spread with the given angular spread. They formulated the problem as a quadratically constrained quadratic program and hence enabled efficient computations of desired window functions.

For more results for uncertainties for functions on the circle which include different spread measures, refer to [7]–[9].

B. Sparsity and Entropy

There are several works in the literature on the uncertainty relation for finite discrete signals where the Discrete Fourier Transform is considered; see, e.g., [10]–[16]. Most results in these can be generically stated as $\phi(\mathbf{x}) + \phi(\hat{\mathbf{x}}) \geq c_s$ or $\phi(\mathbf{x})\phi(\hat{\mathbf{x}}) \geq c_p$ for some constants c_s and c_p where $\phi(\mathbf{x})$ measures the spread of $\mathbf{x}$. In [10]–[13], $\phi(\mathbf{x})$ is chosen to be $\|\mathbf{x}\|_0$, i.e., the sparsity or the number of non-zero entries of $\mathbf{x}$. In [16], the entropy of $\mathbf{x}$, $\mathcal{S}(\mathbf{x})$, is used for $\phi(\mathbf{x})$. For more on

these and other topics regarding uncertainty principle, refer to [17].

While these results are deep and important with much impact, we note that $\|\mathbf{x}\|_0$ and $\mathcal{S}(\mathbf{x})$ (and other similar measures) do not reflect properly the underlying geometry. For example, if $\mathbf{x}$ consists of two pulses, $\|\mathbf{x}\|_0 = 2$ no matter where the pulses are. However, in many contexts, we clearly regard $\mathbf{x}$ is more localized/concentrated if the pulses are next to each other.

Another potential drawback is that the minimizers of these uncertainty measures tend to be the picket-fence signals (Dirac comb). From the perspective of window signals, those are intuitively regarded as poorly localized on the time-frequency plane. These are the reasons why we insist on the definitions in Section III-A.

Before closing the section, we mention the work [18]. In this work, they consider two operators (which may not even be self-adjoint) in a Hilbert space and derive related uncertainty relations. Since their result is general, one can apply it in the setting that we are interested in and obtain some uncertainty relation. For appropriate choice of operators, one may obtain a result that would be close to ours. While interesting, we think this is not a simple task. We also point out that our result links uncertainty relations in two different domains, which is not addressed by [18].

III. Connection between Discrete and Continuous Uncertainty Relations

In this section, we present the main result of this paper.

A. Discretized Time-Frequency Spreads Measures

Let us fix a positive integer N and consider the space $\mathbb{C}^N$ of N-dimensional signals. For our purposes, we will regard a vector $\mathbf{x} \in \mathbb{C}^N$ as defined on N uniformly spaced points

$$\mathcal{D}_N := \Big\{ -\frac{N}{2}+1, -\frac{N}{2}+1, \dots, \frac{N}{2} \Big\}/\sqrt{N}.$$

With this understanding, the Discrete Fourier Transform $\hat{\mathbf{x}} \in \mathbb{C}^N$ of $\mathbf{x} \in \mathbb{C}^N$ is defined by

$$\hat{\mathbf{x}}(k) := \frac{1}{\sqrt{N}} \sum_{j \in \mathcal{D}_N} \mathbf{x}(j) e^{-2\pi j k}, \quad k \in \mathcal{D}_N.$$

The inverse transform has the following form:

$$\mathbf{x}(j) = \frac{1}{\sqrt{N}} \sum_{k \in \mathcal{D}_N} \hat{\mathbf{x}}(k) e^{2\pi j k}, \quad j \in \mathcal{D}_N.$$

Next, we consider a measure of spread of a vector $\mathbf{x} \in \mathbb{C}^N$. For this, we go back to (2) and adapt it to our setting. Viewing $|t-a|$ as the distance between t and a, it is natural to define the variance $v_{\mathbf{x}}$ of $\mathbf{x} \in \mathbb{C}^N$ by

$$v_{\mathbf{x}} := \min_{a \in \mathcal{I}_N} \frac{1}{\|\mathbf{x}\|_2^2} \sum_{j \in \mathcal{D}_N} d(j,a)^2 |\mathbf{x}(j)|^2$$

where $\mathcal{I}_N$ denotes interval $(-\sqrt{N}, \sqrt{N}]/2$ and $d(j,a)$ is the distance between j and a. Now note that our definition of Discrete Fourier Transform assumes that the signals in $\mathbb{C}^N$ are $\sqrt{N}$-periodic. Taking this into account, we define the distance between two points j and a by

$$d(j,a) := \min_{l \in \sqrt{N}\mathbb{Z}} |j - a - l|.$$

Finally, we may define the mean $\mu_{\mathbf{x}}$ of $\mathbf{x}$ to be the minimizing value $a \in \mathcal{I}_N$ of the right-hand side expression above for $v_{\mathbf{x}}$. Note that $v_{\hat{\mathbf{x}}}$ is identically defined.

B. No Uncertainty?

With our definition of uncertainty $v_{\mathbf{x}} v_{\hat{\mathbf{x}}}$, there cannot be any uncertainty principle in the conventional sense. Clearly, for any $\mathbf{x} \in \mathbb{C}^N$, we have $v_{\mathbf{x}} \le N/4$ and $v_{\hat{\mathbf{x}}} \le N/4$. On the other hand, the vector $\mathbf{x}$ that is supported at the origin satisfies $v_{\mathbf{x}} = 0$. Hence, $v_{\mathbf{x}} v_{\hat{\mathbf{x}}} = 0$. It appears that there is no uncertainty at all and that we can do as well as we want!

Of course, such a claim is non-sense, and it runs counter to our intuition that we could not have a signal localized simultaneously in both domains as accurate as we wanted. A closer look at the case $v_{\mathbf{x}} v_{\hat{\mathbf{x}}} = 0$ reveals why we came to this conclusion. The signal $\hat{\mathbf{x}}$ is *globally* supported but $v_{\hat{\mathbf{x}}}$ fails to express the badness in frequency localization since it is always bounded above by $N/4$. In contrast, one would have had $v_{\hat{f}} = \infty$ in such cases. One way to resolve this issue would be to re-define $v_{\hat{\mathbf{x}}}$ (and $v_{\mathbf{x}}$) in a way so that $v_{\hat{\mathbf{x}}} = \infty$ in this kind of signals $\mathbf{x}$. However, we will not take this route since the argument in Section III-A shows that $v_{\mathbf{x}}$ is a sensible way to gauge the time spread of $\mathbf{x}$. How can we formulate a sensible uncertainty principle then?

C. Uncertainty for a Subclass of Discrete Signals

As seen in III-B, there are signals that we clearly want to exclude from our consideration. Thus, it makes sense to restrict our attention to a subclass of signals in $\mathbb{C}^N$ in order to exclude the cases where $\mathbf{x}$ or $\hat{\mathbf{x}}$ are 'globally supported'.

Based on the similarity between the discrete and the continuous Fourier Transforms, it is natural to suspect that discrete finite samples of Gaussian functions might be optimal windows for the Discrete Fourier Transform. While this appears reasonable, it looks difficult to show its validity rigorously. Moreover and perhaps obviously, taking discrete finite samples of Gaussian functions would be a bad idea *unless* they happen to be nearly zero outside the sampling interval. This leads us to introduce 'admissible functions' for our discussion.

We say that $f \in L_2(\mathbb{R})$ is (N,ϵ)-*localized* if

$$|f(t)| \le \frac{\epsilon}{|t|^2}, \quad |t| \ge \frac{\sqrt{N}}{2}, \tag{4}$$

and that a signal $\mathbf{x} \in \mathbb{C}^N$ is *admissible* with constant ϵ if

$$\mathbf{x}(j) = \mathbf{x}_f(j) := N^{-1/4} \sum_{l \in \sqrt{N}\mathbb{Z}} f(j+l), \quad j \in \mathcal{D}_N$$

for a function f with (N,ϵ)-localized functions f, f', $\hat{f}$, $\hat{f}'$. That is, admissible vectors in $\mathbb{C}^N$ are obtained by uniformly sampling $\sqrt{N}$-periodized localized functions in $L_2(\mathbb{R})$.

Our main result is the following:

Theorem 4: Suppose that $f \in L_2(\mathbb{R})$ is localized in time-frequency domain with constant ϵ. Then,

$$\sqrt{v_f v_{\hat{f}}}(1-\sqrt{\epsilon}) \leq \sqrt{v_{\mathbf{x}} v_{\hat{\mathbf{x}}}} \leq \sqrt{v_f v_{\hat{f}}}(1+\sqrt{\epsilon}),$$

where $\mathbf{x} := \mathbf{x}_f$. Thus, if $\mathbf{x}$ is an admissible signal, then

$$v_{\mathbf{x}} v_{\hat{\mathbf{x}}} \geq \frac{(1-\sqrt{\epsilon})^2}{16\pi^2}.$$

To give some idea of the proof, we ask first: Why do we associate $\mathbf{x}_f$ to f instead of sampling the function directly without periodizing it? Upon some reflection, the periodization seems to be natural given the well-known phenomenon of folding (aliasing) associated with sampling approach. It is the periodization that makes the two endpoints of $\mathcal{D}_N$ to be neighbors when the sampling is done. Another crucial reason for us to introduce $\mathbf{x}_f$ in that way is the observation that $\hat{\mathbf{x}}_f = \mathbf{x}_{\hat{f}}$, which is a standard consequence of Poisson Summation Formula. Thanks to this identity, we need only to show that $v_{\mathbf{x}}$ and $v_{\hat{\mathbf{x}}}$ are good approximations of v_f and $v_{\hat{f}}$, respectively. To show that $v_{\mathbf{x}}$ and v_f are close to each other, we show that relevant moments of $\mathbf{x}$ and f are very close. For this purpose, we apply the Poisson Summation Formula and the Parseval's identity. This is also where we use (N,ϵ)-localizedness of f, f', $\hat{f}$, and $\hat{f}'$. A detailed proof of Theorem 4 will be given in an up-coming work.

IV. Discussion and Conclusion

One implication of Theorem 4 is that, if we were to consider only the admissible signals in $\mathbb{C}^N$ as windows – which is not unreasonable in many applications since one would like to have 'smooth' and 'fast-decaying' windows for the time-frequency analysis – thanks to Theorem 1, we can easily construct nearly optimal windows for the Discrete Fourier Transforms by periodizing Gaussian functions and taking finite uniform samples as long as the Gaussian functions are supported essentially on the interval of sampling. This is a mild requirement due to the exponential decay of the Gaussian functions, especially when N is large.

We must keep in mind that 'discrete gaussians' above are, a priori, nearly optimal only among admissible signals in $\mathbb{C}^N$; however, we will demonstrate in the up-coming work that the near optimality of the discrete gaussians may be valid for 'all signals' in $\mathbb{C}^N$. More theoretical evidence related to the near optimality of the discrete Gaussians will be given there. We also show by numerical computation that the uncertainty bound is indeed very close to $1/(16\pi^2)$.

To conclude, we asserted that the uncertainty products of admissible signals with constant ϵ in $\mathbb{C}^N$ are bounded below by constant close to $1/(16\pi^2)$. Based on this claim, we derived that the discrete Gaussians are near optimal windows among the admissible signals.

Even though the near optimality of the discrete Gaussians among *all* signals is strongly suspected, a definitive proof is still missing and remains as future work. Also, as a side problem, it would be interesting to study the characteristics of the discrete Gaussians that arise from Gaussian functions with wide support. For example, are those signals near optimal in some other sense?

Finally, we mention the question of optimal windows for distinguishing, e.g., linear chirps. In our follow-up, we take the approach of this work and try to establish, at least formally, that modulated discrete Gaussians (so that they themselves are linear chirps) are nearly optimal as well.

Acknowledgment

This work is supported by the European project UNLocX (FET-OPEN, grant number 255931). The author would like to thank Bruno Torrésani and Benjamin Ricaud for helpful discussions on the subject. The author also thanks the reviewers for their constructive comments.

References

[1] K. Gröchenig, *Foundations of Time-Frequency Analysis*, ser. Appl. Numer. Harmon. Anal. Birkhäuser Boston, 2001.

[2] E. Breitenberger, "Uncertainty measures and uncertainty relations for angle observables," *Foundations of Physics*, vol. 15, pp. 353–364, 1983.

[3] F. J. Narcowich and J. D. Ward, "Wavelets associated with periodic basis functions," *Appl. Comput. Harmon. Anal.*, vol. 3, pp. 40–56, 1996.

[4] J. Prestin and E. Quak, "Optimal functions for a periodic uncertainty principle and multiresolution analysis," *Proceedings of the Edinburgh Mathematical Society*, vol. 42, pp. 225–242, 1999.

[5] J. Prestin, E. Quak, H. Rauhut, and K. Selig, "On the connection of uncertainty principles for functions on the circle and on the real line," *Journal of Fourier Analysis and Applications*, vol. 9, no. 4, pp. 387–409, 2003.

[6] R. Parhizkar, Y. Barbotin, and M. Vetterli, "Sequences with minimal time-frequency spreads," 2013.

[7] R. Ishii and K. Furukawa, "The uncertainty principle in discrete signals," *IEEE Transactions on Circuits and Systems*, vol. 33, no. 10, pp. 1032–1034, 1986.

[8] L. C. Calvez and P. Vilbe, "On the uncertainty principle in discrete signals," *IEEE Transactions on Circuits and Systems Part II: Analog and Digital Signal Processing*, vol. 6, no. 39, pp. 394–395, 1992.

[9] Y. V. Venkatesh, S. Kumar Raja, and G. Vidyasagar, "On the uncertainty inequality as applied to discrete signals," *International Journal of Mathematics and Mathematical Sciences*, vol. 2006, pp. Article ID 48 185, 22 pages, 2006.

[10] D. L. Donoho and P. B. Stark, "Uncertainty principles and signal recovery," *SIAM Journal on Applied Mathematics*, vol. 49, no. 3, pp. 906–931, 1989.

[11] D. L. Donoho and X. Huo, "Uncertainty principles and ideal atomic decomposition," *IEEE Transactions on Information Theory*, 2001.

[12] T. Tao, "An uncertainty principle for cyclic groups of prime order," *Math. Res. Lett.*, vol. 12, pp. 121–127, 2005.

[13] R. Meshulam, "An uncertainty inequality for finite abelian groups," *European Journal of Combinatorics*, vol. 27, pp. 63–67, 2006.

[14] S. Ghobber and P. Jaming, "On uncertainty principles in the finite dimensional setting," *Linear Algebra and Applications*, vol. 435, pp. 751–768, 2011.

[15] F. Krahmer, G. E. Pfander, and P. Rashkov, "Uncertainty in time-frequency representations on finite abelian groups and applications," *Appl. Comput. Harmon. Anal.*, vol. 25, no. 2, pp. 209–225, 2008. [Online]. Available: http://dx.doi.org/10.1016/j.acha.2007.09.008

[16] T. Przebinda, V. DeBrunner, and M. Ozaydin, "Using a new uncertainty measure to determine optimal bases for signal representations," 1999.

[17] B. Ricaud and B. Torrésani, "A survey of uncertainty principles and some signal processing applications," *Submitted*, 2012.

[18] S. Song Goh and C. A. Micchelli, "Uncertainty principles in hilbert spaces," *Journal of Fourier Analysis and Applications*, vol. 8, no. 4, pp. 335–374, 2002. [Online]. Available: http://dx.doi.org/10.1007/s00041-002-0017-2

[19] G. B. Folland and A. Sitaram, "The uncertainty principle: a mathematical survey," *The Journal of Fourier Analysis and Applications*, vol. 3, no. 3, pp. 207–238, 1997.

Efficient Simulation of Continuous Time Digital Signal Processing RF Systems

Alin Ratiu
CEA, LETI, Minatec Campus
Université de Lyon, Ampère, INSA de Lyon
Grenoble, France
Email: alin.ratiu@cea.fr

Dominique Morche,
Arnaud Arias
CEA, LETI, Minatec Campus
Grenoble, France
Email: dominique.morche@cea.fr,
arnaud.arias@cea.fr

Bruno Allard,
Xuefang Lin-Shi,
Jacques Verdier
Université de Lyon, Ampère, INSA de Lyon
Lyon, France
Email: bruno.allard@insa-lyon.fr,
xuefang.shi@insa-lyon.fr,
jacques.verdier@insa-lyon.fr

***Abstract*—A new simulation method for continuous time digital signal processing RF architectures is proposed. The approach is based on a discrete time representation of the input signal combined with a linear interpolation. Detailed theoretical calculations are presented, which prove the efficiency of the simulation when dealing with narrowband RF signals. We show that, when compared to a discrete time simulation, for the same simulation error, a decrease of almost two orders of magnitude is expected in the necessary number of input samples.**

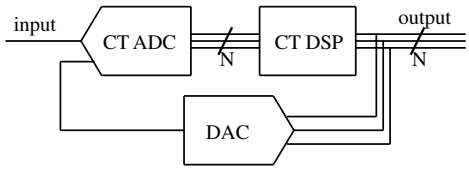


Fig. 1. Generic representation of a CT digital signal processing architecture. The system is composed of a CT ADC, a CT-DSP block and a DAC for feedback.

I. Introduction

Continuous time (CT) digital signal processing systems have been extensively studied over the past years. These systems rely on a continuous time level crossing (CT-ADC) analog to digital conversion stage followed by a continuous time digital signal processor (CT-DSP) block (Fig. 1). In some cases, the CT-DSP output is feedback to the input of the CT-ADC in order to enhance the performance of the conversion. One interesting property of these systems is that no quantization noise is observed because no quantization error is ever made. Furthermore, [1] shows that the data rate requirements are more relaxed as compared to those of synchronous systems. The CT design has been extensively deployed for low power, low frequency applications such as voice processing [2], fast control loops [3] and sensor interfacing [4].

With the recent advances in the deep submicron CMOS technologies, it has become possible to greatly increase the speed of the CT ADCs and of the CT-DSPs. The CT-ADCs have become good candidates for use in direct RF quantization architectures as they solve two important problems. First, the clockless design of these ADCs can greatly reduce the power consumption of the RF receiver. Second, their power consumption depends on input activity: when no input signal is present, the dynamic power consumption drops to 0. An RF implementation of a CT digital signal processing system has been presented in [5].

One of the remaining challenges in designing RF CT digital processing systems is efficient simulation. Fig. 1 presents a general representation of such a system. The simulation of the CT-DSP part can easily be done using an event driven approach, where level crossing events propagate from one system block to another with a certain behavioral and time delay model. However, the efficient generation of the level crossing events (referred to as timestamps) remains problematic since it is difficult to obtain a high precision on the time of the level crossings while using a fast simulator.

Analog simulation can be used for precise device level simulations, but it is too slow for architecture exploration needed to determine the specifications of the involved building blocks. On the other hand, discrete time simulation provides a fast and simple way of simulating CT digital signal processing architectures with a precision that depends on the ratio between the sampling frequency and the useful signal frequency (referred to as oversampling ratio - OSR). Since most of the current CT-DSP circuits are low frequency implementations, high OSR simulations are affordable. However, for RF implementations which operate at GHz frequency and require high frequency resolution, a brute increase of the OSR would greatly increase simulation time.

In this paper we present a hybrid simulation used for timestamp generation which attempts to minimize the simulation error while maintaining a low computational complexity. The rest of the article is organized as follows: Section 2 details the basic principles of the simulation; Section 3 provides a theoretical background for the simulation error computation for sinusoidal signals; Section 4 extends the validity of the results presented in Section 3 to general RF modulated signals. Section 5 applies these results to modulations used for IEEE 802.11a and IEEE 802.11b standards. The frequency representation of the error is discussed in Section 6. Lastly, Section 7 concludes the paper.

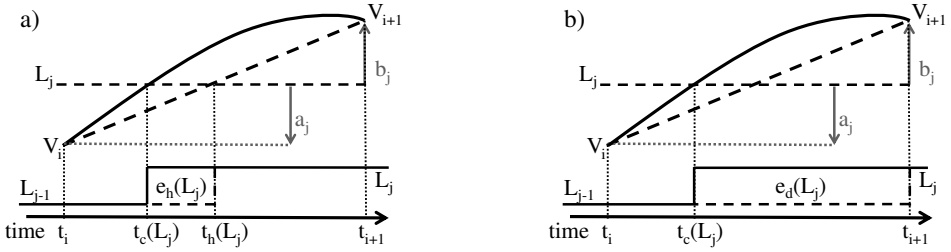


Fig. 2. Level crossing error committed by the proposed hybrid simulation (a) and a discrete time simulation (b).

II. Simulation Overview

The basic principle used by the proposed hybrid simulation is presented in Fig. 2(a). Like the discrete time simulation (Fig. 2(b)), the proposed method uses a discrete time representation of the input signal but with a lower OSR. For each level crossing, a first order interpolation is used to determine the timestamp. In this case, the input signal crosses the level L_j and is sampled at time instants t_i and t_{i+1}. The interpolation produces the timestamp $t_h(L_j)$ which, in the case of RF signals, is a much better approximation of the real level crossing time $t_c(L_j)$ than the result of a synchronous simulation: t_{i+1}.

In order to compare the different simulation methods, performance metrics need to be defined. The two possible performance criteria would be simulation time and simulation error. Since this paper aims to provide readers with insight into architecture level simulations of CT digital signal processing RF systems, the analog simulation is not a good candidate because it is too slow and complex. For the rest of this article the hybrid and the discrete time simulations will be compared, while analytical calculations will be used as benchmark for precision. Finally, as an objective simulation time evaluation, the required OSR for each simulation will be compared for a given upper bound on the committed error.

We define the simulation error energy as the energy of the difference between a perfect CT signal and the respective simulation output. Considering a quantization step of q, the error energy for a level crossing is expressed in (1) for the hybrid simulation and in (2) for the discrete time simulation.

$$e_h(L_j) = q^2 \left| t_c(L_j) - t_h(L_j) \right| \tag{1}$$

$$e_d(L_j) = q^2 \left| t_c(L_j) - t_{i+1} \right| \tag{2}$$

Using this we can now define the signal to noise ratio (SNR) of the simulation as the ratio between the signal energy and the error energy. This expression will be used for performance evaluation for the rest of this article. For an input signal $V(t)$ of duration T which crosses levels L_1 to L_n we have that:

$$SNR = \frac{\int_0^T V(t)^2 \mathrm{d}t}{\sum_{p=1}^{n} e(L_p)} \tag{3}$$

III. Sinusoidal Input

For a sinusoidal input in the form of $V(t) = A\sin(2\pi f t)$, the simulation error can be computed analytically for both simulation scenarios. As defined in Fig. 2, we have V_i and V_{i+1} the signal samples before and after the level crossing, $a_j = L_j - V_i$ and $b_j = V_{i+1} - L_j$. The error for a single level crossing as defined in (1), can be further developed as:

$$e_h(L_j) = q^2 \left| \frac{1}{2\pi f} \arcsin \frac{L_j}{A} - \left(t_i + a_j \frac{T_s}{V_{i+1} - V_i} \right) \right| \tag{4}$$

The difference between two consecutive samples (V_{i+1} and V_i) around the level L_j is approximated using a second order Taylor series:

$$V_{i+1} - V_i = \frac{2\pi}{OSR} \left(\sqrt{A^2 - (L_j - a_j)^2} - \pi \frac{L_j - a_j}{OSR} \right) \tag{5}$$

For simple input signals, precise expressions can be found for a_j, but since we require results which apply to more general cases, we opt for a statistical approach. In this case, a_j can be approximated as a uniform random variable bounded by the difference between two consecutive samples, as defined in (5). The resulting interval is expressed as follows:

$$\begin{aligned} -\frac{2\pi}{OSR} \left(\sqrt{A^2 - L_j^2} - \pi \frac{L_j}{OSR} \right) < a_j < \\ < \frac{2\pi}{OSR} \left(\sqrt{A^2 - L_j^2} - \pi \frac{L_j}{OSR} \right) \end{aligned} \tag{6}$$

Finally, the error corresponding to each level crossing can be computed using the definition of the expected value of (4) given the specific distribution of a_j:

$$E[e_h(L_j)] = \int_{a_j} e_h(L_j) P(a_j) \mathrm{d}a_j \tag{7}$$

This integral is nontrivial, but can easily be computed numerically. By replacing this expression in (3) the theoretical value of the SNR can be computed.

Similarly, the discrete time simulation SNR can be computed using:

$$e_d(L_j) = q^2 \left| \frac{1}{2\pi f} \left(\arcsin \frac{L_j}{A} - \arcsin \frac{L_j + b_j}{A} \right) \right| \tag{8}$$

By symmetry, b_j is a uniform random variable defined over the same interval as a_j (6).

These theoretical calculations are compared with simulation results. For a given number of bits, we plot the corresponding SNR obtained from the hybrid and the discrete time simulation as well as the values predicted by the theory. Using a sinusoid quantized over 5 bits, we plot the results in Fig. 3. Similarly,

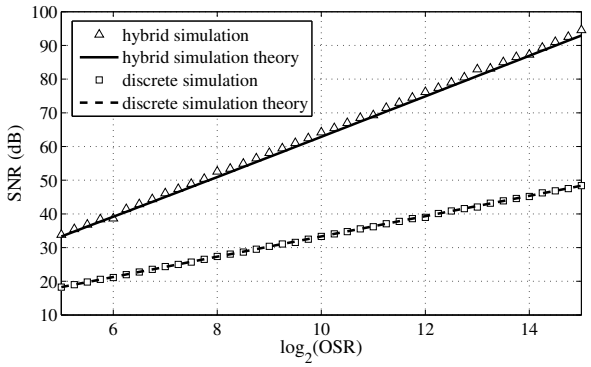


Fig. 3. SNR versus the OSR for the hybrid and discrete time simulation using a single sinusoidal input quantized over 5 bits.

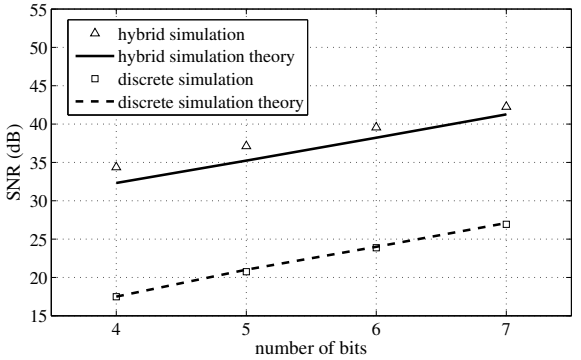


Fig. 4. SNR versus the number of bits for the hybrid and discrete time simulation using a single sinusoidal input oversampled by a factor of 64.

Fig. 4 shows the evolution of the simulation error versus the number of bits for an OSR of 64.

As we can see, the theoretical results are closely matched by the simulation. The maximum difference is only about 2dB and comes from the approximations related to the linearization of the sine function over small intervals. Moreover, there is a significant gain in using the hybrid simulation as opposed to the discrete time simulation. For a given SNR of, for example 50 dB, a reduction of a factor greater than 2^7 is expected in the representation of the input signal. The corresponding speedup gain of the hybrid simulation will be slightly less than 2^7 because extra interpolation points need to be computed.

IV. Simple Modulated Signals

In this section we will extend the previously derived results to a more general case. An RF signal can be represented as a sine wave carrier with phase and amplitude modulation:

$$V(t) = A(t)sin(2\pi ft + \phi(t)) \tag{9}$$

By supposing that $A(t)$ and $\phi(t)$ vary slowly with respect to the sinusoid and by following the same procedure as before, the error committed by the hybrid simulation for each level crossing L_j can be derived:

$$e_h(L_j) = q^2 \left| \frac{1}{2\pi f} \arcsin \frac{L_j}{A(t_i)} - \left(t_i + a_j \frac{T_s}{V_{i+1} - V_i} \right) \right| \tag{10}$$

As in the previous case, a_j can be defined as a uniform random variable. Its interval of variation can be derived from (6) by replacing the old constant amplitude value (A) with the new time varying amplitude defined in (9), equal to $A(t)$.

It is interesting to note that the phase component $\phi(t)$ completely cancels out in the error expression (10). This result can be easily interpreted by the slow variation of $\phi(t)$ with respect to the sinusoid. At any given time t, the phase is expected to remain constant for at least one period of the sinusoid, which is equivalent to a phase shifted version of the signal used in the previous section. Since the previously derived results do not depend on the initial phase of the signal, it follows that the simulation error for modulated signals is only determined by the amplitude modulation term $A(t)$.

In order to compare the theoretical results with a simulation, simple analytical expressions have been chosen for $A(t)$ and $\phi(t)$ so that the true level crossing times can be computed: $A(t) = k_1 t$ and $\phi(t) = k_2 t$. As in the previous case, we use a 5 bit quantizer. k_1 is chosen so that the amplitude modulation term varies from 0 to full scale over 32 periods of the carrier sinusoid. The phase modulation constant k_2 is chosen so that the phase varies from 0 to 2π over the same number of carrier periods. The results are shown in the Fig. 5.

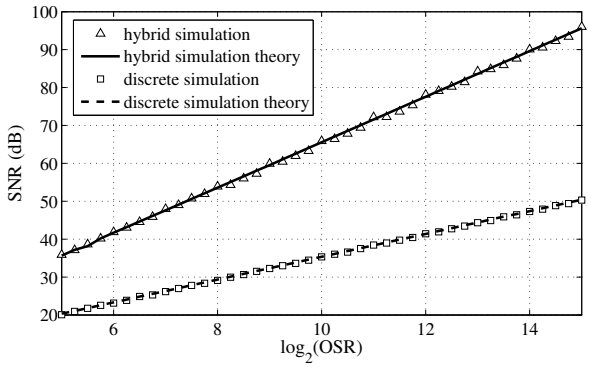


Fig. 5. SNR versus the OSR for the hybrid and discrete time simulation for an amplitude and phase modulated sine quantized over 5 bits.

Once again we see a very good agreement between the simulation and the theory. The proposed hybrid simulation greatly decreases the OSR required for a given SNR. More importantly, the results in this section prove that the previously derived formulae can be used to compute the expected value of the simulation error for any narrowband RF signal, as long as its discrete time baseband representation is known.

V. General Modulated Signals

In this section we will study the simulation performance for IEEE 802.11a and IEEE 802.11b signals. The signals are

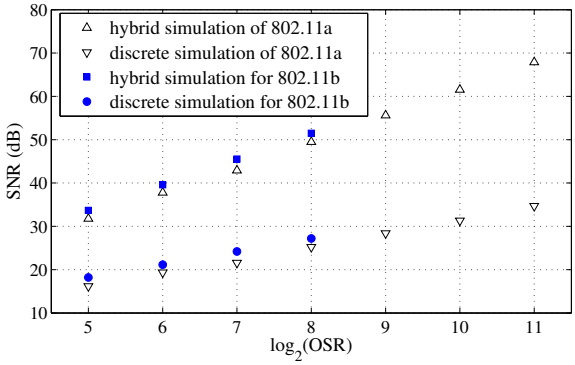


Fig. 6. SNR versus OSR for two wireless standards: 802.11a and 802.11b using the proposed hybrid simulation as well as a discrete time simulation

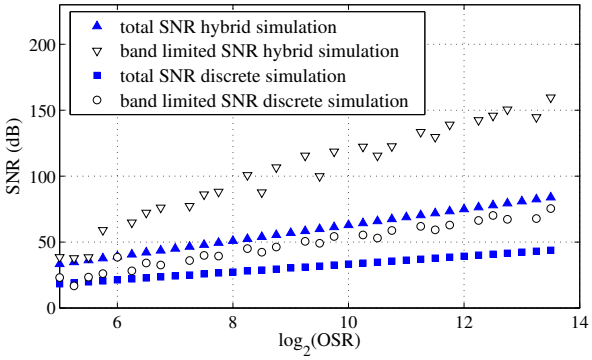


Fig. 7. Total SNR and band limited SNR versus the OSR for a sinusoidal input using the proposed hybrid simulation and the discrete time simulation

quantized over 5 bits and the results using the previously derived theory are plotted in Fig. 6. Since both standards have similar baseband amplitude distributions, the obtained SNRs are also very similar. Furthermore, the use of the hybrid simulation drastically reduces the necessary discrete time representation of the input signal for a given error limit. At 40 dB precision, a hybrid simulation only requires an input OSR of 64 instead of 4096 required by the discrete time simulation.

VI. Error in the Frequency Domain

Until now the total absolute error introduced by the simulation has been derived. However, when dealing with RF signals, it is interesting to study the repartition of this simulation error over the frequency spectrum. Since most RF signals use different frequency bands, in order to provide a fair comparison, the simulation error will be integrated between 0 Hz and the maximum useful frequency contained in the input signal.

For white noise, an SNR gain of the form $20 \log \frac{BW}{F_s}$ would be expected, but this is not be the case. The error, as expressed earlier, is the difference between the simulation output and a perfect level crossing version of it. However, as it has been shown in [6], a level crossing signal has frequency components which contain the fundamental frequency, as well as all of its harmonics with no noise in-between. By limiting the bandwidth of our signal, we will be cutting off high frequency noise as well as parts of our signal - more specifically - its harmonics.

In this section we will consider a sinusoidal input signal at 1 GHz frequency, quantized over 4 bits. The band limited simulation error is compared to the total error for both the hybrid and the discrete time simulation in the Fig. 7. Before analyzing the results, it is important to note that a non-uniform discrete time Fourier transform (NDFT) has been used, since the samples from the hybrid simulation are not periodic. A classic DFT would have required to synchronously resample the asynchronous output of the hybrid simulation which would have introduced another error component.

The band limited results show an even higher gain in precision for the hybrid simulation compared to the discrete time simulation. Using the proposed hybrid simulation, 50dB SNRs can be expected using input OSRs as low as 32.

VII. Conclusion

In this paper we have presented a detailed account of the error introduced by a new simulation method for CT digital signal processing RF systems. The proposed method combines the discrete time and the continuous time approaches by using an oversampled version of the input signal and a first degree interpolation in order to enhance the precision of the level crossing times. It has been shown that the proposed simulation greatly decreases the simulation time when compared to a purely discrete time approach. Moreover, the expressions derived in this paper enable the computation of the simulation error prior to the actual simulation for any RF signal which has a known baseband representation.

The proposed simulation method has been successfully used to study the tradeoffs for the building block parameters used in two CT digital signal processing RF architectures.

References

[1] M. Miskowicz, Send-On-Delta Concept: An Event-Based Data Reporting Strategy, Sensors, 6(1), 2006, pp. 49–63.

[2] B. Schell, Y. Tsividis, A Continuous-Time ADC/DSP/DAC System With No Clock and With Activity-Dependent Power Dissipation, IEEE Journal of Solid-State Circuits, 43(11), 2008, pp. 2472–2481.

[3] Z. Zhao, V. Smolyakov, A. Prodic, Continuous-Time Digital Signal Processing Based Controller for High-Frequency DC-DC Converters, Applied Power Electronics Conference, Anaheim, February 25 - March 1, 2007, pp. 882–886.

[4] V.N. Manyam, D. Chhetri, J.J. Wikner, Clockless Asynchronous Delta Modulator Based ADC for Smart Dust Applications, IEEE/IFIP 19th International Conference on VLSI and System-on-Chip, Hong Kong, October 1-3, 2011, pp. 331–336.

[5] M. Kurchuk, C. Weltin-Wu, D. Morche, Y. Tsividis, Event-Driven GHz-Range Continuous-Time Digital Signal Processor With Activity-Dependent Power Dissipation, IEEE Journal of Solid-State Circuits 47(9), 2012, pp. 2164–2173.

[6] B. Schell, Y. Tsividis, Analysis and simulation of continuous-time digital signal processors. Signal Processing, 89(10), 2009, pp. 2013–2026.

Shift-Variance and Cyclostationarity of Linear Periodically Shift-Variant Systems

Bashir Sadeghi and Runyi Yu
Department of Electrical and Electronic Engineering
Eastern Mediterranean University
Gazimagusa, via Mersin-10, Turkey 99628
bashir.sadeghi@cc.emu.edu.tr and yu@ieee.org

***Abstract*—We study shift-variance and cyclostationarity of linear periodically shift-variant (LPSV) systems. Both input and output spaces are assumed to be of continuous-time. We first determine how far an LPSV system is away from the space of linear shift-invariant systems. We then consider cyclostationarity of a random process based on its autocorrelation operator. The results allow us to investigate properties of output of an LPSV system when its input is a random process. Finally, we analyze shift-variance and cyclostationarity of generalized sampling-reconstruction processes.**

Keywords: cyclostationarity, generalized sampling processes, linear periodically shift-variant systems, shift-variance

I. Introduction

Shift-variance and cyclostationarity are two important issues in the study of linear shift-variant systems and random processes. They have found applications in many fields, including communication and signal processing. See [2], [4] and the reference therein. Recently, Aach and Führ studied shift-variance properties of multirate filterbanks with either deterministic or random inputs [2]. They analyzed shift-variance of the filterbank and calculated the cyclostationarity of its output. For generalized sampling processes, we also performed shift-variance analysis in the deterministic setting [11]. It is the purpose of this paper to report our extension of the results to linear periodically shift-variant LPSV systems whose inputs and outputs are both of continuous-time.

As in [2], we also consider the effect of LPSV systems on the deterministic and random signals. We apply a norm in a Hilbert space of linear systems. The distance between the LPSV system and the space of linear shift-invariant (LSI) systems is then used to measure the shift-variance of the LPSV system. To study cyclostationarity of random processes, we also follow the idea of [2] to link the cyclostationarity to the shift-variance of the associated autocorrelation operator (or function). This is because a random process is wide sense stationary (WSS) if and only if (iff) the operator is shift-invariant; and it is wide sense cyclostationary (WSCS) iff the operator is LPSV. We then obtain a kind of cyclostationarity based on the shift-variance level of the autocorrelation operator. This cyclostationarity also characterizes the distance from the autocorrelation of a random process to the autocorrelation of a nearest WSS process.

Finally we treat generalized sampling-reconstruction processes as a particular application. For minimum error reconstruction, we assume that the sampling and reconstruction kernels form Riesz dual basis [9]. The expected shift-variance and cyclostationarity of the output signal are then determined. Two illustrative examples are provided.

For brevity most derivations and proofs are omitted.

II. Shift-variance of LPSV Systems

We start this section with some basic definitions. The main aim is to determine the nearest shift-invariant system for any LPSV system.

Let L^2 be the Hilbert space of square integrable continuous-time functions. Let $H(L^2 \to L^2)\colon x(t) \mapsto y(t)$ be a bounded linear system. Denote by $\mathcal{B}$ the linear space of all bounded systems. For each $T > 0$, $\mathcal{B}_T$ denotes the subspace of bounded LPSV systems with period T (T-LPSV); and $\mathcal{B}_0$ the subspace of all bounded shift-invariant systems. Note that $\mathcal{B}_0 \subset \mathcal{B}_T$.

For every $H \in \mathcal{B}_T$, we can specify it with its response to shifted impulse function $\delta_s(\cdot) = \delta(\cdot - s)$. Let the response be $H\delta_s(t) = h(t, t-s)$. Then the output of H is given as

$$y(t) = Hx = \int_{-\infty}^{\infty} h(t,s)x(t-s)\,\mathrm{ds} \tag{1}$$

Throughout the paper we assume that $H \in \mathcal{B}_T$, or equivalently $h(t+T, s) = h(t, s)$.

Since $h(t, s)$ is periodic in t with period T, we can express the impulse response as Fourier series

$$h(t,s) = \sum_{k\in Z} h_k(s)\, e^{jk\omega_0 t} \tag{2}$$

where $\omega_0 = 2\pi/T$ and the coefficients are

$$h_k(s) = \frac{1}{T}\int_0^T h(t,s)\, e^{-jk\omega_0 t}\,\mathrm{dt} \tag{3}$$

Let $\hat{h}(t, \xi)$ be Fourier transform of $h(t, s)$ with respect to s. As a function of t, $\hat{h}(t, \xi)$ is also periodic with period T. Thus we can express it as Fourier series

$$\hat{h}(t,\xi) = \sum_{k\in\mathbb{Z}} \hat{h}_k(\xi)\, e^{jk\omega_0 t} \tag{4}$$

where

$$\hat{h}_k(\xi) = \frac{1}{T}\int_0^T \hat{h}(t,\xi)\, e^{-jk\omega_0 t}\,\mathrm{dt} \tag{5}$$

Note that $\hat{h}_k(\xi)$ is actually the Fourier transform of $h_k(s)$.

We define a norm of H by

$$\|H\|^2 = \frac{1}{T}\int_0^T \|H\delta_s(\cdot)\|_2^2 \,\mathrm{ds} \tag{6}$$

By change of variable, we get

$$\|H\|^2 = \frac{1}{T}\int_0^T \int_{-\infty}^{\infty} |h(t,s)|^2 \,\mathrm{ds}\,\mathrm{dt} \tag{7}$$

And using Parseval's relation, we can express the norm in the Fourier domain:

$$\begin{aligned}\|H\|^2 &= \sum_{k\in\mathbb{Z}} \int_{-\infty}^{\infty} |h_k(s)|^2 \,\mathrm{ds} \\ &= \frac{1}{2\pi}\sum_{k\in\mathbb{Z}} \int_{-\infty}^{\infty} |\hat{h}_k(\xi)|^2 \,\mathrm{d}\xi \end{aligned} \tag{8}$$

Let $G \in \mathcal{B}_0$ and g be its impulse response i.e., $g(t) = G\delta(t)$. The distance (squared) between H and G can be calculated as

$$\begin{aligned} d^2(H,G) &= \|H-G\|^2 \\ &= \frac{1}{T}\int_0^T \int_{-\infty}^{\infty} |h(t,s)-g(s)|^2 \,\mathrm{ds}\,\mathrm{dt} \\ &= \int_{-\infty}^{\infty} (|h_0(s)-g(s)|^2 + \sum_{k\neq 0} |h_k(s)|^2) \,\mathrm{ds} \end{aligned} \tag{9}$$

The above expression allows us to determine the nearest system $G_0 \in \mathcal{B}_0$. It is specified by the impulse response

$$g_0(s) = h_0(s) = \frac{1}{T}\int_0^T h(t,s)\mathrm{dt} \tag{10}$$

Note that G_0 is the orthogonal projection of H onto the subspace $\mathcal{B}_0$ and that the impulse response h_0 is the DC component of $h(t,s)$.

Then we have the distance between H and $\mathcal{B}_0$:

$$d^2(H,\mathcal{B}_0) = \frac{1}{T}\int_0^T \int_{-\infty}^{\infty} |h(t,s)-g_0(s)|^2 \,\mathrm{ds} \tag{11}$$

That is,

$$d^2(H,\mathcal{B}_0) = \sum_{k\neq 0} \int_{-\infty}^{\infty} |h_k(s)|^2 \mathrm{ds} \tag{12}$$

or

$$d^2(H,\mathcal{B}_0) = \frac{1}{2\pi}\sum_{k\neq 0} \int_{-\infty}^{\infty} |\hat{h}_k(\xi)|^2 \,\mathrm{d}\xi \tag{13}$$

Note that $h(t,s) - g_0(s)$ is in the orthogonal complement space of the shift-invariant subspace $\mathcal{B}_0$. Thus the LPSV system $H - G_0$ can be considered the shift-variant part of H. Following [2], we can also define $d(H,\mathcal{B}_0)$ as the shift-variance level (denoted by $\mathrm{SV}_2(H)$) of H.

III. Cyclostationarity of Random Processes

In this section we shall study cyclostationarity of a random process by linking it to the shift-variance of a linear system that is determined by autocorrelation function of the process.

Let $z : \mathbb{R} \to \mathbb{C}$ be a zero-mean continuous-time random process with $\mathcal{E}\{|z(t)|^2\} < \infty$, $t \in \mathbb{R}$, where $\mathcal{E}$ denotes the expectation operator. The autocorrelation function of z is defined as $r_z(t,s) = \mathcal{E}\{z(t+s)\, z^*(t)\}$. The random process z is called WSS if $r_z(t,s)$ is independent of time, t; and it is WSCS with period T (T-WSCS) if $r_z(t+T,\tau) = r_z(t,s)$. The notions for discrete-time random process are similarly defined.

We consider the autocorrelation operator R_z as a deterministic linear system whose impulse responses are specified as $R_z\,\delta_s = r_z(t,t-s)$. It is assumed that $R_z \in \mathcal{B}$. Note that z is WSS iff R_z is shift-invariant system; and z is T-WSCS iff R_z is T-LPSV system. This suggests that we can characterize cyclostationarity of random process z by shift-variance of linear system R_z. The amount of cyclostationarity of z can be assessed in terms of the shift-variance measure of R_z:

$$\mathrm{Cyc}(z) = \mathrm{SV}_2(R_z) \tag{14}$$

This measure quantifies the distance between the autocorrelation function $r_z(t,\tau)$ and the nearest autocorrelation function of a WSS random process.

We point out that the degree of cyclostationarity (DCS) defined in [9] is a normalized version of $\mathrm{Cyc}^2(z)$, specifically

$$\mathrm{DCS}(z) = \frac{\mathrm{Cyc}^2(z)}{\int_{-\infty}^{\infty} |r_{z_0}(s)|^2 \,\mathrm{ds}} \tag{15}$$

where $r_{z_0}(t)$ is the impulse response of the nearest system in $\mathcal{B}_0$.

IV. Expected Shift-variance of LPSV Systems With Random Input

Now assume that the input is random (for example, a WSS process), how can we quantify the shift-variance of an LPSV system? This problem was considered by Aach and Führ for multirate discrete-time systems. They introduced the notation of expected shift-variance, which is related not just to the system itself, but also to the random input.

Similar to [1], introduce the commutator

$$[H,\tau_s] = H\tau_s - \tau_s H \tag{16}$$

where $\tau_s : x(t) \mapsto x(t-s)$ is the shift operator. The expected shift-variance of H with input x can then be defined as

$$\mathrm{ESV}^2(H,x) = \frac{1}{T^2}\int_0^T \int_0^T \mathcal{E}(|[H,\tau_s]x(t)|^2) \,\mathrm{ds}\,\mathrm{dt} \tag{17}$$

After some tedious calculations, we obtain in the time-domain that

$$\mathrm{ESV}^2(H,x) = 2\sum_{k\neq 0} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h_k^*(t)\, h_k(t-s) r_x(s) \,\mathrm{ds}\,\mathrm{dt} \tag{18}$$

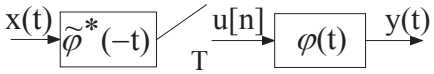


Fig. 1. A generalized sampling and reconstruction process

and in the Fourier domain that

$$\mathrm{ESV}^2(H,x) = \frac{1}{\pi}\sum_{k\neq 0}\int_{-\infty}^{\infty} |\hat{h}_k(\xi)|^2\, S_x(\xi)\,\mathrm{d}\xi \qquad (19)$$

where $S_x(\xi)$ is the power spectral density of x, i.e., the Fourier transform of r_x [7]. Note that the ESV tells how different the expected value of the output to a shifted input from that of shifted output.

Note that the ESV is zero iff the system is LSI. And the Fourier domain expression (19) provides some insight as when an LPSV system becomes LSI (see the examples at the end of Section V).

V. Generalized Sampling-reconstruction processes

Sampling-reconstruction process plays an important role in signal processing and communication. In particular, the generalized sampling-reconstruction theory of Unser and Aldroubi [9] offers a versatile framework in studying many problems of sampling beyond Shannon.

In this section, we investigate cyclostationarity and shift-variance of generalized sampling-reconstruction processes show in Fig. 1, where x is a zero-mean WSS random process; and for minimum error between input signal and the output signal (which is in the space of spanned by $\{\varphi(\cdot - nT)\}_n$), $\tilde{\varphi}(t)$ and $\varphi(t)$ are assumed to be dual Riesz basis [9]. It is well-known that sampling generally results in shift-variance whereas reconstruction introduces cyclostationarity.

Consider the sampling first. The output of sampling $u[n]$ is given by [1]

$$\begin{aligned} u[n] &= \langle x, \tilde{\varphi}(\cdot - nT)\rangle \\ &= \int_{-\infty}^{\infty} \tilde{\varphi}^*(t-nT)\, x(t)\,\mathrm{d}t \end{aligned} \qquad (20)$$

Note that u is of discrete-time and has autocorrelation function

$$r_u[n,k] = \mathcal{E}\{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \varphi^*(t_1-(n+k)T)x(t_1)\varphi(t_2-nT)x^*(t_2)\,\mathrm{d}t_1\mathrm{d}t_2\} \qquad (21)$$

By change of variable $t_1 - nT \to t_1$ and $t_2 - nT \to t_2$ we get

$$r_u[n,k] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \tilde{\varphi}^*(t_1+kT)\,\tilde{\varphi}(t_2)\, r_x(t_1-t_2)\mathrm{d}t_1\mathrm{d}t_2 \qquad (22)$$

Since r_u above is independent of n, thus it is a WSS discrete random process and the power spectral density of u is

$$S_u(e^{j\xi T}) = \frac{1}{T}\sum_{n\in Z} |\hat{\tilde{\varphi}}(\xi + 2n\pi/T)|^2\, S_x(\xi + 2n\pi/T) \qquad (23)$$

[1]Note that the integration for random signals is in the mean square sense [5].

In the reconstruction part, the output is

$$y(t) = \sum_{n\in Z} u[n]\,\varphi(t-nT) \qquad (24)$$

and its autocorrelation function becomes

$$r_y(t,s) = \sum_{n_1,n_2\in Z} \varphi(t+s-n_1T)\varphi^*(t-n_2T)r_u[n_1-n_2] \qquad (25)$$

Note that $r_y(t+T,s) = r_y(t,s)$, thus y is T-WSCS.

In order to analyze the shift-variance of system H in Fig. 1, we need to determine its input-output relation. By direct substitution and change of variable, we obtain that

$$y(t) = Hx = \int_{-\infty}^{\infty} h(t,s)\, x(t-s)\,\mathrm{d}s \qquad (26)$$

where

$$h(t,s) = \sum_{n\in Z} \tilde{\varphi}^*(t-s-nT)\,\varphi(t-nT) \qquad (27)$$

is the impulse response. It can be shown that $h(t+T,s) = h(t,s)$ and

$$\hat{h}_k(\xi) = \frac{1}{T}\hat{\tilde{\varphi}}^*(\xi)\hat{\varphi}(\xi + k\omega_0) \qquad (28)$$

Since $\hat{\tilde{\varphi}}(\xi)\sum_{k\in\mathbb{Z}}|\hat{\varphi}(\xi+k\omega_0)|^2 = T\hat{\varphi}(\xi)$ [6], hence

$$\|H\|^2 = \int_{-\infty}^{\infty} |\varphi(t)|^2\mathrm{d}t \qquad (29)$$

Apply the results in previous sections, we can obtain the following results:

$$\mathrm{Cyc}^2(y) = \frac{1}{2\pi\, T^2}\sum_{k\neq 0}\int_{-\infty}^{\infty} |\hat{\varphi}(\xi)\,\hat{\varphi}(\xi+2\pi k/T)\, S_u(e^{j\xi T})|^2\,\mathrm{d}\xi \qquad (30)$$

and

$$\mathrm{ESV}^2(H,x) = \frac{1}{\pi T^2}\sum_{k\neq 0}\int_{-\infty}^{\infty} |\hat{\tilde{\varphi}}(\xi)\hat{\varphi}(\xi+2\pi k/T)|^2 S_x(\xi)\mathrm{d}\xi \qquad (31)$$

Finally, let us consider two examples. The first one is about the traditional Shannon's sampling. In this case the kernels $\tilde{\varphi}(t) = \varphi(t) = \mathrm{sinc}(t/T)/\sqrt{T}$. From equation (25) and (27) it is not immediate that the output y is WSS for WSS input and that the sampling-reconstruction system is LSI. On the other hand if we examine (30) and (31), we can easily see that $\mathrm{Cyc}(y) = \mathrm{ESV}(H,x) = 0$ for each x, since the Fourier transform of φ is zero for $|\xi| > \pi/T$. Consequently the output is WSS and the sampling-reconstruction process is LSI.

In the other example, φ is taken to be B-spline of various order n [8] which is normalized such that $\int_{-\infty}^{\infty}|\varphi(t)|^2\mathrm{d}t = 1$. And for the input we take the unit variance white Gaussian noise, hence $S_x(\xi) = 1$. Now the expected shift-variance turns out to be equivalent to cyclostationarity: $\mathrm{ESV}(H,x) = \sqrt{2}\,\mathrm{Cyc}(y)$. Furthermore from (30) it follows that

$$\mathrm{Cyc}^2(y) = 1 - \frac{1}{2\pi\, T^2}\int_{-\infty}^{\infty} |\hat{\varphi}(\xi)\,\hat{\tilde{\varphi}}(\xi)|^2\,\mathrm{d}\xi \qquad (32)$$

Consequently $0 \le \mathrm{Cyc}(y) \le 1$.

For the zero order B-spline (a box), we obtain $\mathrm{Cyc}(y) = 0.5773 > 0.5$. This result indicates that output is quite non-stationary (in the wide sense). We also obtain numerical values of $\mathrm{Cyc}(y)$ for other orders: they are 0.3546 ($n = 1$), 0.2864 ($n = 2$), 0.2485 ($n = 3$), and 0.2227 ($n = 4$). Again the output is not WSS for all cases, but now the output y seems to be more stationary than non-stationary as the order n increases. We expect that $\mathrm{Cyc}(y)$ can become arbitrary small for n large enough.

VI. Conclusion

We reported our latest study on shift-variance and cyclostationarity analysis of LPSV systems. We extended recent similar results to systems with continuous-time input and output, rendering our treatment of generalized sampling-reconstruction processes. The extension enables us to define and compute the following:

- a distance of an LPSV system to the nearest linear shift-invariant system.
- a cyclostationarity of a WSCS random process
- the expect shift-variance of a generalized sampling process and cyclostationarity of its output when the input is WSS.

References

[1] T. Aach and H. Führ, "On bounds of shift-variance in two-channel mutirate filter banks," *IEEE Trans. Signal Process.*, vol. 57, pp. 4292–4303, Nov. 2009.

[2] T. Aach and H. Führ, "Shift-variance measures for multirate LPSV filter banks with random input signals," *IEEE Trans. Signal Process.*, vol. 60, pp. 5124–5134, Oct. 2012.

[3] T. Chen and L. Qiu, "Linear periodically time-varying discrete-time systems: aliasing and LTI approximation," *Syst. and Control Lett.*, vol. 30, pp. 225–235, 1997.

[4] W. A. Gardner, A. Napolitano, and L. Paura, "Cyclostationarity: half a century of research," *Signal Process.*, vol. 86, pp. 639–697, 2006.

[5] A. Leon-Garcia, *Probability, Statistics, and Random Processes for Electrical Engineering*. Pearson Prentice Hall, 2008.

[6] S. Mallat, *A Wavelet Tour of Signal Processing: The Sparse Way*. Amsterdam: Academic Press, 2009.

[7] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*. New York: McGraw-Hill, 1965.

[8] M. Unser, "Sampling — 50 years after Shannon," *Proc. IEEE*, vol. 88, pp. 569–587, Apr. 2000.

[9] M. Unser and A. Aldroubi, "A general sampling theory for nonideal acquisition devices," *IEEE Trans. Signal Process.*, vol. 42, pp. 2915–2925, Nov. 1994.

[10] R. Yu, "Shift-variance measure of multichannel multirate systems," *IEEE Trans. Signal Process.*, vol. 59, pp. 6245–6250, Dec. 2011.

[11] R. Yu, "Shift-variance analysis of generalized sampling process," *IEEE Trans. Signal Process.*, vol. 60, pp. 2840–2850, Jun. 2012.

[12] G. D. Zivanovic and W. A. Gardner, "Degrees of cyclostationarity and their application to signal detection and estimation," *Signal Process.*, vol. 22, pp. 287–297, 1991.

Constructive sampling for patch-based embedding

Moshe Salhov, Guy Wolf, Amit Bermanis and Amir Averbuch
School of Computer Science, Tel Aviv University
Tel Aviv 69978, Israel

***Abstract*—To process high-dimensional big data, we assume that sufficiently small patches (or neighborhoods) of the data are approximately linear. These patches represent the tangent spaces of an underlying manifold structure from which we assume the data is sampled. We use these tangent spaces to extend the *scalar* relations that are used by many kernel methods to *matrix* relations, which encompass multidimensional similarities between local neighborhoods in the data. The incorporation of these matrix relations improves the utilization of kernel-based data analysis methodologies. However, they also result in a larger kernel and a higher computational cost of its spectral decomposition. We propose a dictionary construction that approximates the oversized kernel in this case and its associated patch-to-tensor embedding. The performance of the proposed dictionary construction is demonstrated on a super-kernel example that utilizes the Diffusion Maps methodology together with linear-projection operators between tangent spaces in the manifold.**

I. Introduction

Recent methods for massive high dimensional data analysis utilize a manifold structure on which data points are assumed to lie. This manifold is immersed (or submersed) in an ambient space that is defined by observable parameters. Kernel methods such as Diffusion Maps (DM) [1] have provided good results in analyzing such massive high dimensional data. These methods are based on the spectral decomposition of a kernel designed to incorporate scalar similarities between data points. The resulting embedding of the data points into an Euclidean space preserves the qualities represented by the designed kernel.

Recently, DM was extended in several different ways to handle the orientation in local tangent spaces [2]–[4]. The relation between two patches is described by a matrix instead of a scalar value. The resulting kernel captures enriched similarities between local structures in the underlying manifold. These enriched similarities can be used to analyze local areas around data points instead of analyzing their specific locations.

The discussed enrichments increase considerably the kernel size, which is a limiting factor in the applicability of kernel methods to real problems. Considerable efforts have been invested for example in [5], [6] and others in approximating the spectral decomposition operator to become computationally feasible. In this paper, we combine the patch-based embedding from [3], [4] with the dictionary construction approach in [5] to approximate the spectral decomposition of a non-scalar kernel that utilizes the underlying patch structure inside the ambient space.

II. Problem setup

Let $\mathcal{M}$ be a d dimensional manifold that lies in the ambient space $\mathbb{R}^m$, where $d \ll m$, and let $M \subseteq \mathbb{R}^m$ be a set of n points sampled from it. Each point $x \in M$ has a d-dimensional tangent space $T_x(\mathcal{M})$, which is a subspace of $\mathbb{R}^m$. Let $O_x \in \mathbb{R}^{m\times d}$, $x \in M$, be a matrix whose columns $o_x^1, \ldots, o_x^d \in \mathbb{R}^m$ form an orthonormal basis of $T_x(\mathcal{M})$. If the manifold is densely sampled, $T_x(\mathcal{M})$ can be approximated by a small enough patch $N(x) \subseteq M$ around $x \in M$. We will assume that $o_x^1, \ldots, o_x^d$ are the principal directions of $N(x)$ and vectors in $T_x(\mathcal{M})$ are expressed according to this basis.

A. Diffusion Maps

The original diffusion maps method [1] is based on defining an isotropic kernel K as $k(x,y) \triangleq e^{-\frac{\|x-y\|}{\varepsilon}}$, for every $x, y \in \mathcal{M}$, where ε is a meta-parameter of the algorithm. This kernel represents the affinities between points on the manifold. The kernel is normalized with the degrees $q(x) \triangleq \int_{y\in\mathcal{M}} k(x,y)$, $x \in \mathcal{M}$ to produce a stochastic transition operator P, with $p(x,y) = \frac{k(x,y)}{q(x)}$, which defines a Markov process (i.e., a diffusion process) over the manifold $\mathcal{M}$. Its symmetric conjugate A, where $a(x,y) = \sqrt{q(x)}p(x,y)\frac{1}{\sqrt{q(y)}} = \frac{k(x,y)}{\sqrt{q(x)q(y)}}$, defines the diffusion affinities between data-points. Spectral analysis of the diffusion affinity kernel A yields the eigenvalues $1 = \sigma_0 \geq \sigma_1 \geq \ldots$ and their corresponding eigenvectors $\psi_0, \psi_1, \ldots$, which are used to construct the desired map that embeds each data point $x \in \mathcal{M}$ onto the point $\Psi(x) = (\sigma_i \psi_i(x))_{i=0}^{\delta}$ for a sufficiently small δ, which is the dimension of the embedded space and depends on the decay of the spectrum of A.

III. Super-kernel

For $x, y \in M$, let $O_{xy} = O_x^T O_y \in \mathbb{R}^{d\times d}$, where O_x and O_y represent bases of the tangent spaces $T_x(\mathcal{M})$ and $T_y(\mathcal{M})$, respectively. The matrix O_{xy} represents a linear-projection between these tangent spaces, and, in some sense, the similarity between them. We will refer to it as a tangent similarity matrix. We use the diffusion affinity kernel A and the tangent similarity matrices O_{xy} to define the following *super-kernel*:

Definition 1. *A super-kernel is a block matrix $G \in \mathbb{R}^{nd\times nd}$ with $n \times n$ blocks and each block in it is a $d \times d$ matrix. Each block $G_{xy} \in \mathbb{R}^{d\times d}$ of a* Linear-Projection Diffusion Super-kernel *is defined as $G_{xy} \triangleq a(x,y)O_{xy} = a(x,y)O_x^T O_y$, $x, y \in M$ and represents the affinity between the patches $N(x)$ and $N(y)$.*

We will use spectral decomposition for analyzing a super-kernel G, and utilize it to embed the patches $N(x)$ of the manifold (for $x \in M$) into a tensor space. Let $|\lambda_1| \geq |\lambda_2| \geq \ldots \geq |\lambda_\ell|$ be the ℓ most significant eigenvalues of G and let $\phi_1, \phi_2, \ldots, \phi_\ell$ be their corresponding eigenvectors. According to the spectral theorem, if ℓ is greater than the numerical rank of G, then $G \approx \sum_{i=1}^{\ell} \lambda_i \phi_i \phi_i^T$, where the eigenvectors are treated as column vectors.

Each eigenvector ϕ_i, $i = 1, \ldots, \ell$, is a vector of length nd. We denote each of its elements as $\phi_i(o_x^j)$ where $x \in M$ and $j = 1, \ldots, d$. An eigenvector ϕ_i can also be regarded as a vector of n sections, each of which is a vector of length d that corresponds to a point $x \in M$ on the manifold. To express this notion we use the notation $\varphi_i^j(x) = \phi_i(o_x^j)$ (for $x \in M, i = 1, \ldots, \ell, j = 1, \ldots, d$). Thus, the section in ϕ_i, which corresponds to $x \in M$, is the vector $(\varphi_i^1(x), \ldots, \varphi_i^d(x))^T$.

We use the eigenvalues and eigenvectors of G to construct a spectral map whose definition is similar to the standard (i.e., classic) diffusion map: $\Phi(o_x^j) = (\lambda_1 \phi_1(o_x^j), \ldots, \lambda_\ell \phi_\ell(o_x^j))^T$. By using this construction, we get nd vectors of length ℓ. Each $x \in M$ corresponds to d of these vectors, i.e., $\Phi(o_x^j)$, $j = 1, \ldots, d$. We use these vectors to construct the tensor $\mathcal{T}_x \in \mathbb{R}^\ell \otimes \mathbb{R}^d$ for each $x \in M$, whose coordinates are $[\mathcal{T}_x]_{ij} = \lambda_i \varphi_i^j(x)$, $x \in M, i = 1, \ldots, \ell, j = 1, \ldots, d$. Each tensor $\mathcal{T}_x$ represents an embedding of the patch $N(x)$, $x \in M$, into the tensor space $\mathbb{R}^\ell \otimes \mathbb{R}^d$.

A. Mathematical properties

1) Spectral properties: The linear-projection operators, which define the tangent similarity matrices by a LPD super-kernel, express some important properties of the manifold structure, e.g., curvatures between patches and differences in orientation. While there might be other ways to construct a super-kernel that expresses these properties, LPD super-kernels do have an important property, which is given by the following theorem:

Theorem 1. *A LPD super-kernel G is positive definite and its operator norm satisfies $\|G\| \leq 1$.*

Proof. Theorem 3.1 from [3] shows that linear-projection super-kernels have a non-negative spectrum that is bounded from above by the spectral norm of the used scalar affinities. Following the footsteps of that proof in our case, with the diffusion affinity kernel, which is positive definite and whose spectral norm is one, yields the result in the theorem. □

The patch-to-tensor embedding that is achieved by the LPD super-kernel is defined by the spectral analysis of this super-kernel. Therefore, the spectral properties of this super-kernel, which are shown in Theorem 1, are crucial for the patch-based data analysis that utilizes this embedding.

2) Embedded distances: The classical diffusion map provides an embedded space in which the Euclidean distance between data points is equal to a diffusion distance in the original ambient space. This diffusion distance measures the distance between two diffusion "bumps" $a(x, \cdot)$ and $a(y, \cdot)$, each of which is a row in the symmetric diffusion kernel that defines the diffusion map. From a technical point of view, this relation means that the Euclidean distance between two arbitrary points in the range of a diffusion map is equal to the Euclidean distances between the corresponding rows of its symmetric diffusion kernel. The following theorem (whose proof appears in [3]) shows a similar property of the LPD-based patch-to-tensor embedding:

Theorem 2. *Let $x, y \in M$ be two points on the manifold and let $\mathcal{T}_x$ and $\mathcal{T}_y$ be their embedded tensors, then $\|\mathcal{T}_x - \mathcal{T}_y\|_F^2 = \sum_{z \in M} \sum_{j=1}^{d} \|(a(x,z)O_x^T - a(y,z)O_y^T)o_z^j\|^2$, where the tensors are treated as matrices (i.e., their coordinate matrices) when computing the Frobenius distance between them.*

The vectors o_z^j in Theorem 2 are unit vectors that form an orthonormal basis of the tangent space $T_x(\mathcal{M})$ at the point $z \in M$. For each point $z \in M$, the matrix $[a(x,z)O_x^T - a(y,z)O_y^T]$ is applied to each of these unit vectors and the squared lengths of the resulting vectors are summed. These terms can be seen as extensions of the terms $(a(x,z) - a(y,z))$ of the original diffusion distance, which only consider the differences between scalar affinities. Further explanations about the meaning of the extended diffusion distance can be found in [3].

IV. Out-of-Sample Extension for Vector Fields

The presented patch-to-tensor embedding is based on the spectral analysis of a large super-kernel G. In order to approximate this spectral decomposition, we will use a dictionary (i.e., a set of representatives) and extend its results (using an out-of-sample extension) to the entire dataset. This extension method can also be utilized to extend this decomposition either from the dictionary or from the dataset to new data points. The super-kernel G can be regarded as an operator on tangent vector fields of the manifold $\mathcal{M}$ restricted to a dataset M. Therefore, the spectral decomposition of G consists of eigenvector fields that span the range of G. Hence, an out-of-sample extension of the eigenvector fields is equivalent to the out-of-sample extension of vector fields in the range of G.

Out-of-sample extension of vector fields assumes an a priori knowledge of a set of data points M and a corresponding vector field where each vector lies on the respective local tangent space. Consider a tangent vector field $\vec{v} : M \to \mathbb{R}^d$ such that $\vec{v}(x) \in T_x(\mathcal{M})$ for all $x \in M$. Then, the given data points are used to construct the super-kernel G. Since G is positive definite (see Theorem 1), it is also invertible and its range consists of all these vector fields.

The out-of-sample extension of a new data point under the PTE settings aims to find the new corresponding vector in the local tangent space of the new point. The extension coefficients $\vec{u}$ are designed to minimize $\|G\vec{u} - \vec{v}\|_2$ over the given set of training data points. These coefficients, which minimize the l_2 norm, are computed by using the inverse of G such that $\vec{u} = G^{-1}\vec{v}$.

The coefficient vector $\vec{u}$ can be interpreted as a vector field

$\vec{u} : M \to \mathbb{R}^d$ over the set of training points or, equivalently,

$$\vec{v}(x) = \sum_{y \in M} G_{(x,y)} \vec{u}(y),\ x \in M, \tag{1}$$

where $\vec{u}(y)$, $y \in M$, are considered as the coefficients of the vector field $\vec{v}$ according to the super-kernel G. Consider a new data point $x' \in \mathcal{M} \backslash M$ with the matrix $O_{x'}$ whose columns $o_{x'}^1, \ldots, o_{x'}^d$ form an orthonormal basis for the tangent space $T_{x'}(\mathcal{M})$. We can extend the vector field to a new data point x' by setting the value $\vec{v}(x')$ to be

$$\vec{v}(x') \triangleq \sum_{y \in M} \tilde{G}_{(x',y)} \vec{u}(y), \tag{2}$$

where $\tilde{G}_{(x',y)} = \bar{p}(x', y) O_{x'}^T O_y$, $y \in M$, are the non-scalar affinity blocks between the new data point and the data points in the dataset. The extension in Eq. 2 is consistent with the values $\vec{v}(x)$, $x \in M$, in Eq. 1.

While the new affinity blocks in Eq. 2 are not known in advance as part of the super-kernel, they are easily computed for any new data point. This approximation only considers values of the vector field $\vec{u}$ for the data points in M, which can be computed in advance by using the pseudo inverse of the super-kernel G. This computation is not complicated, but it is beyond the scope of this paper since it is not essential for the presented dictionary construction. Therefore, this provides a feasible out-of-sample extension of a vector field, which is similar to the methods shown in [7], [8] for the scalar case.

The extension in Eq. 2 can be interpreted geometrically by separately considering the projections and the scalar weights in the affinity blocks of the super-kernel. First, the extension projects the coefficient vector field $\vec{u}$ from the manifold M to the tangent space $T_{x'}(\mathcal{M})$ of the new data point x'. This projection expresses the coefficient vectors in local terms of the manifold around x'. Then, the value of the vector field $\vec{v}$ at x' is computed by using a weighted sum of the projected coefficient vectors on the tangent space $T_{x'}(\mathcal{M})$.

V. Constructive Patch Sampling

According to Lemma 3.3 in [3], the sum in Eq. 1 can be rephrased in terms of the embedded tensors $x \in M$ to be

$$\vec{v}(x) = \sum_{y \in M} \mathcal{T}_x^T \mathcal{T}_y \vec{u}(y). \tag{3}$$

However, due to linear dependencies between the embedded tensors, this sum may contain redundant elements. Indeed, if $\mathcal{T}_z = \sum_{z \neq y \in M} c_y^z \mathcal{T}_y$ for some scalar coefficients $c_y^z \in \mathbb{R}$, $z \neq y \in M$, then Eq. 3 becomes $\vec{v}(x) = \sum_{z \neq y \in M} \mathcal{T}_x^T \mathcal{T}_y (\vec{u}(y) + c_y^z \vec{u}(z))$. This enables us to eliminate the redundant tensors and by applying an iterative approach, we obtain a small subset linearly independent tensors that are sufficient for computing Eqs. 1 and 2.

Similarly, we can use matrix coefficients instead of scalar ones to incorporate reacher relations between tensors. Therefore, $\mathcal{T}_z$ is tensorialy dependent in $\{\mathcal{T}_y\}_{z \neq y \in M}$ if $\mathcal{T}_z = \sum_{z \neq y \in M} \mathcal{T}_y C_y^z$ for some matrix coefficients $C_y^z \in \mathbb{R}^{d \times d}$, $z \neq y \in M$. This dependency expresses more redundancies than the standard linear dependency. As a result, we obtain a sparser set of tensorialy independent tensors that enables us to efficiently compute Eqs. 1 and 2. This set of representative tensors constitutes a dictionary that compactly represents the embedded tensor space.

A. Dictionary Construction

We use an iterative approach to construct the described dictionary by a sequential scan of the data points in M. In the first iteration, we define the scanned set $X_1 = \{x_1\}$ and the dictionary $D_1 = \{x_1\}$. At each iteration $s = 2, \ldots, n$, we have a new data point x_s, the scanned set $X_{s-1} = \{x_1, \ldots, x_{s-1}\}$ from the previous iteration and the dictionary D_{s-1} that represents X_{s-1}. The dictionary D_{s-1} is in fact a subset of η_{s-1} data points from X_{s-1} that are sufficient to represent its embedded tensors. We define the scanned set $X_s = X_{s-1} \cup \{x_s\}$. Our goal is to define the dictionary D_s of X_s, based on the dictionary D_{s-1} with the new data point x_s. To do this, a dependency criterion has to be established. If this criterion is satisfied, then the dictionary remains the same such that $D_s = D_{s-1}$. Otherwise, it is updated to include the new data point $D_s = D_{s-1} \cup \{x_s\}$.

We use a dependency criterion that is similar to the approximated linear dependency (ALD) criterion from [5]. The ALD measures the distance between vector candidates and the span by the dictionary vectors. In our case, we want to approximate the tensorial dependency of $\mathcal{T}_{x_s}$ on the tensors in the dictionary D_{s-1}. Therefore, we define the distance of $\mathcal{T}_{x_s}$ from the dictionary D_{s-1} as $\delta_s \triangleq \min_{C_1, \ldots, C_{\eta_{s-1}}} \left\| \sum_{j=1}^{\eta_{s-1}} \mathcal{T}_{y_j} C_j - \mathcal{T}_{x_s} \right\|_F^2$, where $\|\cdot\|_F$ denotes the Frobenius norm, and $C_1, \ldots, C_{\eta_{s-1}} \in \mathbb{R}^{d \times d}$ are matrix coefficients. The approximated tensorial dependency (ATD) criterion is defined as $\delta_s \leq \mu$, for some accuracy threshold $\mu > 0$. If the ATD criterion is satisfied, then the tensor $\mathcal{T}_{x_s}$ can be approximated by the dictionary D_{s-1}, using the matrix coefficients $C_1^s, \ldots, C_{\eta_{s-1}}^s$ of δ_s. Otherwise, the dictionary has to be updated by adding x_s to it. Lemma 3 (whose proof appears in [9]) shows that δ_s and the dictionary-based approximation can be expressed in terms of the super-kernel and without requiring knowledge of the embedded tensors the embedded tensors.

Lemma 3. *Let $\hat{G}_{s-1} \in \mathbb{R}^{d\eta_{s-1} \times d\eta_{s-1}}$ be the super-kernel of the data points in D_{s-1}, and let $H_s \in \mathbb{R}^{d\eta_{s-1} \times d}$ be a $\eta_{s-1} \times 1$ block matrix whose j-th $d \times d$ block is $G_{(y_j, x_s)}$, $j = 1, \ldots, \eta_{s-1}$. Then, the optimal matrix coefficients in δ_s are the η_{s-1} blocks, of size $d \times d$, in $\hat{G}_{s-1}^{-1} H_s$. The achieved δ_s satisfies $\delta_s = \mathrm{tr}[G_{(x_s, x_s)} - H_s^T \hat{G}_{s-1}^{-1} H_s]$.*

Essentially, this lemma eliminates the need for prior knowledge of the embedded tensors during the dictionary construction. At each iteration s, the criterion $\delta_s < \mu$ is considered. Based on this condition, we decide whether to add x_s to the dictionary or just approximate its tensor. The threshold μ is given in advance as a meta-parameter and δ_s can be computed by using Lemma 3. Therefore, the dictionary construction process only requires knowledge of a relatively limited number

of super-kernel blocks, which is determined by the size of the dictionary and not by the size of the dataset.

VI. Example: MNIST Handwritten Digit Classification using Patch-based Analysis

The patch-based methodology provides a general framework that can be utilized to a wide spectrum of data analysis tasks such as clustering, classification, anomaly detection and related manifold learning tasks. In this section, we demonstrate its utilization of the task of MNIST Handwritten digit classification. This experiment was done utilizing an of-the-shelve computer with a $I7-2600$ quad core CPU and a $16GB$ of DDR3 memory.

The MNIST database of handwritten digits [10] (available from http://yann.lecun.com/exdb/mnist/) consists of a training set of $60,000$ examples and a test set of $10,000$ examples. Each digit example is given as a grey levels 28×28 image. The digit images were centered by computing the center of mass of the pixels, and a translation operation was preformed to position this point at the center of the 28×28 field. MNIST is a subset of a larger set available from NIST. Many machine learning methods have been tested on this data set, hence the recognition performance is highly competitive. Currently, convolutional networks show a state-of-the-art recognition accuracy with an error of 0.23% [11]. For our purpose, the MNIST dataset provides a dataset of $70,000$ data points of very high dimensional measurements of size 728 pixels per a measured digit. In our experiments, we used the images as is.

The dictionary approximated patch-based embedding was utilized to embed the MNIST dataset of $70,000$ examples by the following steps. First, in each data point we identified the 150 nearest neighbors and computed the corresponding local PCA. For each local tangent space, we kept the 3 significant eigenvectors. Secondly, the diffusion affinities were computed with $\varepsilon = 105$ (see Section II-A), which is the Euclidean distance mean of all pairwise data points. The proposed dictionary construction with ATD threshold $\mu = 0.0001$ identified 93 important patches and their corresponding local tangent spaces. Finally, the approximated tensors were constructed utilizing $\ell = 30$. The labeling of each test data-point was estimated using the label of the nearest training data-point, where the pairwise distance was computed as the Frobenius norm of the difference between the corresponding embedded tensors. The resulting labeling error of the patch-based recognition method is 5.8%. Table I compares the computational costs of the straightforward implementation of the PTE algorithm from [3] and the presented dictionary-based algorithm on the MNIST dataset.

Size	SVD Cost - Full G	Dict. Size	SVD Cost - Approx. G
$70,000$	$O\left(70,000^3 \times 3^3\right)$	93	$O\left(70,000 \times 77,841\right)$

TABLE I
Computational cost of the SVD step in the dictionary approximated PTE (*SVD Cost - Approx. G*) vs. the full SVD of the super-kernel (*SVD Cost - Approx. G*) of the NIST dataset.

Although we are not far away from the state-of-the-art in digit recognition, the proposed method has the following advantages: 1. It shows that patch processing can be practically utilized for recognition and data analysis tasks. 2. Big high-dimensional datasets can be processed on "cheap" hardware such as in our case where the algorithm ran on less than $1000\$$ worth of hardware.

VII. Conclusions

The proposed construction in the paper extends the dictionary construction in [5] by using the LPD super-kernel from [3], [4]. This is done by an efficient dictionary-based construction that assumes the data is sampled from an underlying manifold while utilizing the non-scalar relations between manifold patches instead of considering individual data-points. The constructed dictionary contains patches from the underlying manifold, which are represented by the embedded tensors from [3], instead of individual data points. Therefore, it encompasses multidimensional similarities between local areas of the data. The patch-based dictionary reduces the computational costs of the spectral analysis in comparison to the PTE [3], hence, it enables us to apply this patch processing approach for datasets that were impractical to process and embed before.

Acknowledgments

This research was supported by the Israel Science Foundation (Grant No. 1041/10) and the Eshkol Fellowship from the Israeli Ministry of Science & Technology.

References

[1] R. Coifman and S. Lafon, "Diffusion maps," *Applied and Computational Harmonic Analysis*, vol. 21, no. 1, pp. 5–30, 2006.

[2] A. Singer and H. Wu, "Vector diffusion maps and the connection laplacian," *Communications on Pure and Applied Mathematics*, vol. 65, no. 8, pp. 1067–1144, 2012.

[3] M. Salhov, G. Wolf, and A. Averbuch, "Patch-to-tensor embedding," *Applied and Computational Harmonic Analysis*, vol. 33, no. 2, pp. 182 – 203, 2012.

[4] G. Wolf and A. Averbuch, "Linear-projection diffusion on smooth Euclidean submanifolds," *Applied and Computational Harmonic Analysis*, vol. 34, pp. 1 – 14, 2013.

[5] Y. Engel, S. Mannor, and R. Meir, "The kernel recursive least-squares algorithm," *Signal Processing, IEEE Transactions on*, vol. 52, no. 8, pp. 2275 – 2285, aug. 2004.

[6] C. Baker, *The Numerical Treatment of Integral Equations*. Oxford: Clarendon Press, 1977.

[7] R. Coifman and S. Lafon, "Geometric harmonics: A novel tool for multiscale out-of-sample extension of empirical functions," *Applied and Computational Harmonic Analysis*, vol. 21, no. 1, pp. 31–52, 2006.

[8] A. Bermanis, A. Averbuch, and R. Coifman, "Multiscale data sampling and function extension," *Applied and Computational Harmonic Analysis*, vol. 34, pp. 182 – 203, 2013.

[9] M. Salhov, A. Bermanis, G. Wolf, and A. Averbuch, "Approximate patch-to-tensor embedding via dictionary construction," *Submitted*, 2012.

[10] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," in *Proceedings of the IEEE*, 1998, pp. 2278–2324.

[11] D. Ciresan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification," in *CVPR*, 2012, pp. 3642–3649.

The Constrained Earth Mover Distance Model, with Applications to Compressive Sensing

Ludwig Schmidt, Chinmay Hegde, Piotr Indyk
Computer Science and Artificial Intelligence Lab
Massachusetts Institute of Technology

***Abstract*—Sparse signal representations have emerged as powerful tools in signal processing theory and applications, and serve as the basis of the now-popular field of compressive sensing (CS). However, several practical signal ensembles exhibit additional, richer structure beyond mere sparsity. Our particular focus in this paper is on signals and images where, owing to physical constraints, the positions of the nonzero coefficients do not change significantly as a function of spatial (or temporal) location. Such signal and image classes are often encountered in seismic exploration, astronomical sensing, and biological imaging. Our contributions are threefold: (i) We propose a simple, deterministic model based on the *Earth Mover Distance* that effectively captures the structure of the sparse nonzeros of signals belonging to such classes. (ii) We formulate an approach for approximating any arbitrary signal by a signal belonging to our model. The key idea in our approach is a min-cost max-flow graph optimization problem that can be solved efficiently in polynomial time. (iii) We develop a CS algorithm for efficiently reconstructing signals belonging to our model, and numerically demonstrate its benefits over state-of-the-art CS approaches.**

I. Introduction

A signal (or image) is said to be k-sparse if only k of its coefficients in a given basis expansion are nonzero; in other words, the intrinsic information content in the signal is miniscule relative to its apparent size. This simple notion enables a wide variety of conceptual and algorithmic techniques to compress, reconstruct, denoise, and process practical high-dimensional signals and images. Notably, sparsity serves as the cornerstone of the field of compressive sensing (CS), an interesting alternative to the classical Shannon/Nyquist theory for signal sampling and reconstruction [1, 2]. A canonical result in CS states that for a k-sparse signal of length n, merely $O(k \log n/k)$ non-adaptive, *linear* measurements (samples) suffice to ensure robust, efficient reconstruction. When $k \ll n$, this can lead to significant practical benefits.

In several practical applications, the nonzero coefficients of signal ensembles exhibit additional, richer relationships that cannot be captured by mere sparsity. Consider, for example, a 2D "image" constructed by column-wise stacking of seismic time traces (or shot records) measured by geophones positioned on a uniform linear array. Assuming the presence of only a few subsurface reflectors, the physics of wave propagation dictates that such a 2D image would essentially consist of a number of curved lines, possibly contaminated with noise (see Figure 1). A convenient model for such an image is to simply assume that each column is sparse; indeed, such a sparsity assumption has been proven to be beneficial for efficient shot record sampling and reconstruction [3]. However, while this assumption may suffice for some situations, such a model cannot capture the the fact that the indices of the nonzeros change smoothly across adjacent columns. Such settings are commonplace; for example, similar "line" singularities are encountered in applications such as biological imaging and radio-astronomy.

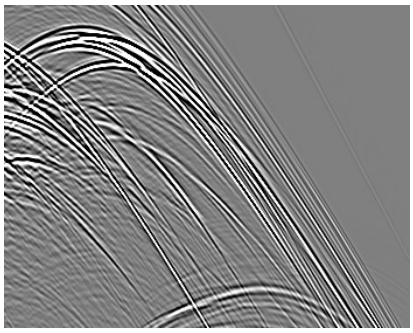


Fig. 1. *Example of a seismic shot record (Sigsbee2A data set). The horizontal axis corresponds with space (receiver) and the vertical axis with time. Note that the large coefficients of neighboring columns are at similar locations.*

In this paper, we propose a deterministic model for sparse signal ensembles where the locations of the nonzeros, or the support, of a signal transforms continuously as a function of spatial (or temporal) location. A key ingredient in our model is the classical Earth Mover Distance (EMD) [4], and we will call it the *Constrained EMD* model. Informally, our proposed model assumes that: (i) each signal in our ensemble is k-sparse, and (ii) the cumulative EMD between pairs of adjacent signal supports is constrained to be no greater than a nonnegative parameter B. The parameter B controls how dramatically the support can vary across different signals; a value of $B = 0$ indicates that the support remains invariant across all signals in our ensemble, while a large value of B admits potentially drastic changes across adjacent supports.

Next, given an arbitrary input signal (ensemble) x, we develop an efficient algorithm to find a near-optimal ℓ_2-approximation of x in the Constrained EMD model. We show that the support of the optimal approximation can be discovered by solving a small number of *min-cost max-flow* [5] problems over a specially defined graph. Each intermediate problem can be solved using existing, highly efficient network optimization methods, and therefore the overall signal approximation can be obtained in polynomial time.

Additionally, we demonstrate the advantages of the Con-

strained EMD model, and the associated approximation algorithm, in the context of compressive sensing. Geometrically, the model is equivalent to a particular *union of subspaces* of the ambient signal space. Therefore, we can leverage the framework of *model-based compressive sensing* [6] to build a new CS reconstruction algorithm that is specially tailored to signal ensembles well-described by the Constrained EMD model. We illustrate the numerical benefits of the new algorithm in comparison with existing state-of-the-art CS recovery approaches.

The rest of this paper is organized as follows. Section II provides a brief introduction to structured sparsity and compressive sensing. Section III introduces the constrained EMD model and describes our main algorithm. Section IV illustrates the advantages of our method with example reconstructions of images and quantitative results of algorithm performance. Section V concludes with a discussion of further directions.

II. Background

A. Preliminaries

A signal $x \in \mathbb{R}^n$ is said to be k-sparse in the orthobasis Ψ if at most $k < n$ coefficients of the basis expansion $\alpha = \Psi^T x$ are nonzero. In this paper, we assume that the basis Ψ is the identity matrix, while noting that all our results are conceptually valid for general Ψ. The *support* of x is defined as the set of indices corresponding to nonzero entries of x; this can be represented by a binary vector $s(x) \in \{0,1\}^n$ with at most k ones. Denote the set of *all* k-sparse signals by Σ_k. Geometrically, this set is equivalent to the union of the $\binom{n}{k}$ canonical k-dimensional subspaces of $\mathbb{R}^n$.

B. Structured sparsity

Often, we possess some additional information about the support of a sparse signal x. For example, suppose we are interested in k-sparse signals with only a few permitted configurations of $s(x)$. This defines a *union of subspaces model* $\mathcal{A}$ [7], comprising only m_k canonical k-dimensional subspaces of $\mathbb{R}^n$, with $m_k < \binom{n}{k}$. Let $x|_\Omega$ represent the entries of x corresponding to the set of indices $\Omega \subseteq \{1,\dots,n\}$, and let Ω^C denote the complement of the set Ω. Then, define:

$$\mathcal{A} = \bigcup_{m=1}^{m_k} \mathcal{X}_m,\ \mathcal{X}_m := \{x : x|_{\Omega_m} \in \mathbb{R}^k, x|_{\Omega_m^C} = 0\}, \quad (1)$$

where each subspace $\mathcal{X}_m$ contains all signals x with $\mathrm{supp}(x) \in \Omega_m$. In light of this definition, we view any such union of subspaces as a *structured sparsity model*. As in the general k-sparse case, given a signal x, we seek a signal x^* such that $x^* \in \mathcal{A}$, and $\|x - x^*\|_2$ is minimized. We define a *model-projection* algorithm as a procedure $\mathbb{M}(x,k)$ which returns the best k-term approximation of a given signal under the model $\mathcal{A}$, i.e., $x^* = \mathbb{M}(x,k)$.

C. Compressive Sensing

Suppose instead of collecting all the coefficients of a vector $x \in \mathbb{R}^n$, we merely record $m = O(k \log n/k)$ inner products (measurements) of x with $m < n$ pre-selected vectors, i.e., we observe an m-dimensional $y = \Phi x$, where $\Phi \in \mathbb{R}^{m \times n}$. The central tenet of compressive sensing (CS) is that x can be *exactly* recovered from y, even though Φ is rank-deficient (and therefore has a nontrivial nullspace). Numerous algorithms for signal recovery have been developed; particularly, iterative support selection algorithms (such as CoSaMP [8] and IHT [9]) have emerged that are both numerically stable and computationally efficient. Also, an added advantage is that such iterative algorithms can be easily *tailored to any arbitrary structured sparsity model*; this forms the central premise of *model-based* compressive sensing framework, initially proposed in [6]. In Section III below, we describe this further.

D. Related Work

There has been prior research on reconstructing time sequences of spatially sparse signals (e.g., [10]). Such approaches assume that the support of the signal (or even the signal itself) does not change much between two consecutive time steps. However, the variation between two columns a and b was defined according to the ℓ_0 distance between the supports $\|s(a) - s(b)\|_0$. In contrast, in this paper we measure this difference according to the classical Earth Mover Distance (EMD) (also variously known as the *Mallows* or the *Wasserstein* distance) between the supports. As a result, our model easily handles signals such as those in Figure 3, where the supports of any two consecutive columns can potentially be *even disjoint*, yet differ very little according to the EMD.

Another related work is that of [11], who proposed the use of the EMD in a compressive sensing context in order to measure the *approximation error* of the recovered signal. In contrast, in this paper we are using the EMD to constrain the *support set* of the signals.

III. The Constrained EMD Model

Below, we interpret the signal $x \in \mathbb{R}^n$ as a matrix $X \in \mathbb{R}^{h \times w}$ with $n = h\,w$. Furthermore, we denote the individual columns of X with $x_i \in \mathbb{R}^h$ for $i \in [w]$.

A. Definitions

Definition 1: The EMD of two index sets A and B with $|A| = |B|$ is defined as:

$$\mathrm{EMD}(A,B) = \min_{\pi : A \to B} \sum_{a \in A} |a - \pi(a)|, \quad (2)$$

where π ranges over all one-to-one mappings from A to B.

Definition 2: The support-EMD of two k-sparse vectors $a, b \in \mathbb{R}^h$ is defined as:

$$\mathrm{sEMD}(a,b) = \mathrm{EMD}(\mathrm{supp}(a), \mathrm{supp}(b)). \quad (3)$$

Definition 3: The *Constrained EMD model* is the set:

$$\mathcal{A}_{k,B} = \{X \in \mathbb{R}^{h \times w} : |\mathrm{supp}(x_i)| = k \text{ for } i \in [w], \sum_{i=1}^{w-1} \mathrm{sEMD}(x_i, x_{i+1}) \le B\}. \quad (4)$$

$$X = \begin{bmatrix} 1 & 3 \\ 0 & -1 \\ 2 & 1 \end{bmatrix} \qquad G_{X,k,\lambda} =$$

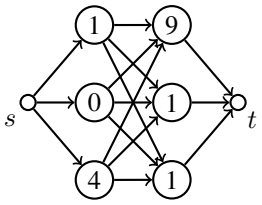


Fig. 2. *A signal X with the corresponding flow network $G_{X,k,\lambda}$. The node costs are the squared amplitudes of the corresponding signal components (negation omitted here). The capacities and edge costs are omitted for clarity. All capacities in the flow network are 1. The edge costs are the vertical distances between the start and end nodes.*

The set $\mathcal{A}_{k,B}$ in (4) is a subset of the set of all $k-$sparse signals Σ_k, and therefore the Constrained EMD model constitutes a specific instance of a structured sparsity model (1). For given dimensions of X, the Constrained EMD model has two parameters: (i) k, the sparsity of each column x_i and (ii) B, the cumulative support-EMD of adjacent columns x_i and x_{i+1}. Importantly, we note that we only constrain the EMD between adjacent signal *supports* and not the actual signal coefficients.

B. Graph-Based Model-Projection

In order to use our Constrained EMD signal model within a model-based compressive sensing framework, we need an algorithm that approximates arbitrary signals with signals in our model. Formally, we need a model-projection algorithm $\mathbb{M}(x, k, B)$ that returns a $\widehat{x} \in \mathcal{A}_{k,B}$ minimizing $\|x - x'\|_2$ for all $x' \in \mathcal{A}_{k,B}$.

To achieve this, we use the following *graph-based* approach. Observe that the support-EMD (3) of a pair of signals is the minimal cost of a maximum bipartite matching of the two support sets, where the edge costs are given by the absolute difference between the indices. We extend this intuition to ensembles of signals, via the notion of a *flow network*.

Definition 4: For a given signal X, sparsity k and parameter λ, the *flow network* $G_{X,k,\lambda}$ consists of the following elements:

- The *nodes* comprise a source s, a sink t and a node $v_{i,j}$ for $i \in [h]$, $j \in [w]$, i.e. one node per signal coefficient.
- G has an *edge* from every $v_{i,j}$ to every $v_{k,j+1}$ for $i, k \in [h]$, $j \in [w-1]$. Moreover, there is an edge from s to every $v_{i,1}$ and from every $v_{i,w}$ to t for $i \in [h]$.
- The *capacity* of every edge and node is 1.
- The *cost of a node* $v_{i,j}$ is $-x_{i,j}^2$. The *cost of an edge* from $v_{i,j}$ to $v_{k,j+1}$ is $\lambda|i-k|$. The cost of the source, the sink and all edges incident to the source or sink is 0.
- The *supply* at the source, and the *demand* at the sink, both equal k.

Figure 2 illustrates the construction of an example $G_{X,k,\lambda}$. Observe that for any $G_{X,k,\lambda}$, a standard min-cost max-flow optimization [5] through this network reveals a subset of nodes S that corresponds to exactly k indices per column. Moreover, this optimal flow minimizes the cost $-\|X|_S\|^2 + \lambda \sum_{i=1}^{w-1} \text{EMD}(s_i, s_{i+1})$ over all choices of S. This cost includes both the fidelity of the signal projection as well as the cumulative support-EMD across columns. The trade-off between these two quantities is determined by the parameter λ; for small values of λ, the resulting flow has a large support-EMD and vice versa. Setting $\lambda = 0$ removes the EMD-constraint while $\lambda = +\infty$ is equivalent to selecting the k rows with the largest amplitude sums. By systematically varying the parameter λ, we can find a support S that belongs to the Constrained EMD model $\mathcal{A}_{k,B}$ for a target B and simultaneously maximizes the quality of the projection under this constraint.

Algorithm 1 Model projection $\mathbb{M}(x, k, B)$

$\lambda_l \leftarrow 0, \lambda_h \leftarrow 1$
do
  $\lambda_h \leftarrow 2\lambda_h$
  Run min-cost max-flow on G_{X,k,λ_h}
while resulting support has total support-EMD $> B$.
do
  $\lambda_m \leftarrow (\lambda_h + \lambda_l)/2$
  Run min-cost max-flow on G_{X,k,λ_m}
  if resulting support has total support-EMD $> B$
    $\lambda_l \leftarrow \lambda_m$
  else
    $\lambda_h \leftarrow \lambda_m$
while $\lambda_h - \lambda_l > \epsilon_\lambda$
return $\widehat{x}$ corresponding to min-cost max flow on G_{X,k,λ_h}

Algorithm 1 describes the entire model projection algorithm. In order to solve the min-cost max-flow instances, it is possible to exploit the special structure of the graph. Since all edges and nodes have unit capacity, it is sufficient to find k cheapest augmenting paths in the flow network. Using Dijkstra's algorithm and assuming a square X, i.e. $h = w = \sqrt{n}$, each min-cost max-flow can be found in $O(kn^{3/2})$ time.

C. Compressive Sensing

The model projection method (Alg. 1) is useful in a number of contexts. Here, we use Alg. 1 in order to develop a new compressive sensing (CS) reconstruction algorithm specially tailored to signals and images with line singularities. Since the constrained EMD model essentially is a special structured sparsity model $\mathcal{A}_k$, as in (1), Alg. 1 provides an projection algorithm for this model. Given such a projection algorithm, the framework of model-based compressive sensing [6] suggests that iterative support selection algorithms, such as CoSaMP and IHT, can easily be modified in order to be tailored for signals belonging to the constrained EMD model. Further, the modified algorithms are provably stable, as well as provably achieve successful recovery using fewer measurements than the conventional (unmodified) algorithms.

We summarize our proposed CS recovery method as Alg. 2; we call it EMD-CoSaMP. The modification is simple: simply replace the signal thresholding steps (3 and 6) by an appropriate model projection step. A similar modification of IHT can also be developed (the description of which we omit); we will call it EMD-IHT. Below, we empirically illustrate the benefits of our proposed model-based CS recovery algorithms.

Algorithm 2 EMD-CoSaMP(Φ, y)

$\widehat{x}_0 \leftarrow 0, r \leftarrow y, i \leftarrow 0$

while not converged **do**

1. $i \leftarrow i + 1$
2. $e \leftarrow \Phi^T r$
3. $\Omega \leftarrow \text{supp}(\mathbb{M}(e, 2k, 2B))$
4. $T \leftarrow \Omega \cup \text{supp}(\widehat{x}_{i-1})$
5. $z|_T \leftarrow \Phi_T^\dagger y,\ z|_{T^C} = 0$
6. $\widehat{x}_i \leftarrow \mathbb{M}(z, k, B)$
7. $r \leftarrow y - \Phi \widehat{x}_i$

return $\widehat{x} \leftarrow \widehat{x}_i$

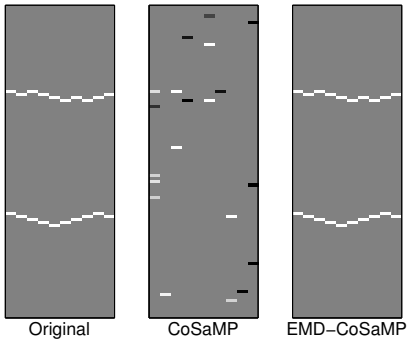


Fig. 3. *Benefits of CS reconstruction using EMD-CoSaMP. (left) Original image with parameters $h = 100$, $w = 10$, $k = 2$, $B = 20$, $m = 80$. (center) CS reconstruction using CoSaMP [8]. (right) CS reconstruction using EMD-CoSaMP. CoSaMP fails, while our proposed algorithm is able to perfectly recover the image.*

IV. Numerical Experiments

In all our experiments, we use the LEMON library [12] in order to solve the min-cost max-flow subroutine in Alg. 1. Figure 3 displays a test grayscale image of size 100×10 with edge discontinuities such that the total sparsity is $2 \times 10 = 20$ and the cumulative EMD across pairs of adjacent columns is equal to $B = 20$. We measure linear samples of this image using merely $m = 80$ random Gaussian measurements, and reconstruct using CoSaMP as well our proposed approach (EMD-CoSaMP). Each iteration of EMD-CoSaMP takes less than three seconds to execute. As visually evident from Fig. 3, CoSaMP fails to reconstruct the image, while our proposed algorithm provides an accurate reconstruction.

Figure 4 displays the results of a Monte Carlo experiment to quantify the effect of the number of random measurements M required by different CS reconstruction algorithms to enable accurate reconstruction. Each data point in Fig. 4 was generated using 100 sample trials over randomly generated measurement matrices. Successful recovery is declared when the converged solution is within an ℓ_2 distance of 5% relative to the Euclidean norm of the original image. We observe that our proposed EMD-CoSaMP and EMD-IHT algorithms achieve successful recovery with far fewer measurements than their conventional (unmodified) counterparts.

V. Conclusions

We have proposed a deterministic structured sparsity model, and associated model projection algorithm, based on the Earth Mover Distance (EMD) for signals and images with line

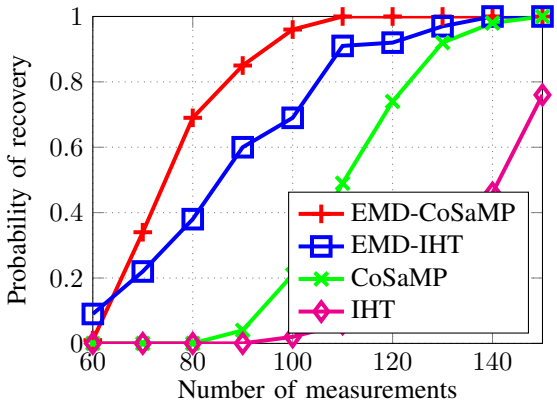


Fig. 4. *Comparison of several reconstruction algorithms. The signal is the same as in Figure 3. The probability of recovery is with respect to the measurment matrix and generated using 100 trial runs. The recovery algorithms using our constrained EMD model have a higher probability of recovery than standard algorithms.*

singularities. We leverage this algorithm to develop a new compressive sensing (CS) recovery algorithm with significant numerical benefits. We defer a full theoretical characterization of our proposed CS recovery algorithm, as well as a thorough study of practical applications such as seismic shot record acquisition, to a future expanded version of this work.

Acknowledgements

The authors would like to thank Lei Hamilton, Chris Yu, and Detlef Hohl for helpful discussions. This work was supported in part by a grant from the MITEI-Shell program, by the MADALGO center, and by the Packard Foundation.

References

[1] D. Donoho, "Compressed sensing," *IEEE Trans. Inform. Theory*, vol. 52, no. 4, pp. 1289–1306, September 2006.

[2] E. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inform. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[3] F. Herrmann, "Randomized sampling and sparsity: Getting more information from fewer samples," *Geophysics*, vol. 75, no. 6, pp. WB173–WB187, 2010.

[4] Y. Rubner, C. Tomasi, and L. J. Guibas, "The Earth Mover's Distance as a metric for image retrieval," *Intl. J. Comp. Vision*, vol. 40, no. 2, pp. 99–121, 2000.

[5] R. Ahuja, T. Magnanti, and J. Orlin, *Network Flows: Theory, Algorithms, and Applications*. Prentice Hall, 1993.

[6] R. Baraniuk, V. Cevher, M. Duarte, and C. Hegde, "Model-based compressive sensing," *IEEE Trans. Inform. Theory*, vol. 56, no. 4, pp. 1982–2001, 2010.

[7] Y. Eldar and M. Mishali, "Robust recovery of signals from a structured union of subspaces," *IEEE Trans. Inform. Theory*, vol. 55, no. 11, pp. 5302–5316, 2009.

[8] D. Needell and J. Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples," *Appl. Comput. Harmon. Anal.*, vol. 26, no. 3, pp. 301–321, 2009.

[9] T. Blumensath and M. Davies, "Iterative hard thresholding for compressive sensing," *Appl. Comput. Harmon. Anal.*, vol. 27, no. 3, pp. 265–274, 2009.

[10] N. Vaswani and W. Lu, "Modified-CS: Modifying compressive sensing for problems with partially known support," in *Proc. IEEE Int. Symp. Inform. Theory (ISIT)*, Seoul, Korea, Jun. 2009.

[11] P. Indyk and E. Price, "K-median clustering, model-based compressive sensing, and sparse recovery for Earth Mover Distance," in *Proc. ACM Symp. Theory of Comput.*, San Jose, CA, June 2011.

[12] B. Dezs, A. Jüttner, and P. Kovács, "LEMON: an open source C++ graph template library," *Electron. Notes Theor. Comput. Sci.*, vol. 264, no. 5, pp. 23–45, 2011.

Orlicz Modulation Spaces

Catherine Schnackers
RWTH Aachen University
Lehrstuhl A für Mathematik
Aachen, Germany
Email: catherine.schnackers@matha.rwth-aachen.de

Hartmut Führ
RWTH Aachen University
Lehrstuhl A für Mathematik
Aachen, Germany
Email: fuehr@matha.rwth-aachen.de

***Abstract*—In this work we extend the definition of modulation spaces associated to Lebesgue spaces to Orlicz spaces and mixed-norm Orlicz spaces. We give the definition of the Orlicz spaces L^Φ, a generalisation of the L^p spaces of Lebesgue. Therefore we characterise the Young function Φ and give some basic properties of this spaces. We collect some facts about this spaces that we need for the time frequency analysis, then we introduce the Orlicz modulation spaces. Finally we present a discretisation of the Orlicz space and mixed-norm Orlicz space and a characterisation of the modulation space by discretisation.**

I. Introduction

The modulation spaces were introduced in 1983 by H. Feichtinger. The idea is to impose a norm on the short-time Fourier transform and to define Banach spaces of signals with a given time-frequency behavior. Especially, the modulation space $M^{p,q}$ consists of all tempered distributions such that the short-time Fourier transform is a function in the mixed-norm Lebesgue space $L^{p,q}$. We will extend this concept and examine modulation spaces associated to Orlicz spaces and mixed-norm Orlicz spaces. The Orlicz spaces L^Φ are a generalisation of the L^p spaces of Lebesgue. For the Young function $\Phi(x) = |x|^p$ with $p \geq 1$, $L^\Phi(\mu) = L^p(\mu)$. In general, the function Φ is a convex function, precisely a Young function. The mixed-norm Orlicz spaces $L^{\Phi_1\Phi_2}$ are vector-valued L^{Φ_2} spaces where Φ_1, Φ_2 are Young functions. Since the function $x \mapsto f(\cdot, x)$ takes values in the Banach space L^{Φ_2}, the mixed-norm Orlicz spaces $L^{\Phi_1\Phi_2}$ arise by taking a L^{Φ_2} norm with respect to the time variable x and an L^{Φ_1} norm with respect to the frequency variable w. This can be considered as a generalisation of the mixed-norm Lebesgue spaces $L^{p,q}$. As general setting let (Ω, Σ, μ) be a measure space, where Ω is a set, Σ is a σ-algebra of Ω and μ a σ-additive measure on Σ and $f : \Omega \to \overline{\mathbb{C}}$ is a measurable function. We also assume that the measure μ has the finite subset property, i.e., for $E \in \Sigma$ with $\mu(E) > 0$ there exists a subset $F \in \Sigma$ with $F \subset E$ and $0 < \mu(F) < \infty$.

II. Orlicz spaces and mixed-norm Orlicz spaces

A. Definition and properties

Firstly we give the definition of a Young function Φ and the Δ_2-condition, which is a growth condition.
After that we introduce the Orlicz spaces and characterise norms so that these spaces are Banach spaces. Then we determine their corresponding dual spaces.

This section is based on the book [8] *Theory of Orlicz spaces* of Rao and Ren.

Definition 1: (Young function) A convex function $\Phi : \mathbb{R} \to \overline{\mathbb{R}^+}$ which satisfies the conditions:

1) $\Phi(-x) = \Phi(x)$,$\Phi(0) = 0$,
2) $\lim_{x\to\infty}\Phi(x) = +\infty$,

is called *Young function*.
In the theory of Lebesgue spaces, the conjugate exponent q to p is related to the dual space. By analogy, one can define the so called complementary function, this function is the counterpart to the conjugate exponent.

Definition 2: (Complementary function) If $\Psi : \mathbb{R} \to \overline{\mathbb{R}^+}$ is defined by $\Psi(y) = \sup\{x|y| - \Phi(x); x \geq 0\}$. Then Ψ is called the *complementary function* to the Young function Φ.

In the structure theory of Orlicz spaces a classification of the Young function based on properties of their growth plays a central role. Of particular importance for us will be the Δ_2-condition.

Definition 3: (Δ_2-condition) A Young function $\Phi : \mathbb{R} \to \mathbb{R}^+$ is said to satisfy the Δ_2-*condition*, if there exists a constant $K > 0$ and $x_0 \in \mathbb{R}_0^+$, such that

$$\Phi(2x) \leq K\Phi(x) \qquad \text{for all } x \geq x_0 \geq 0.$$

Hereafter we say that a Δ_2-condition for Φ is *regular* if it holds locally (for a $x_0 > 0$) when the measure in $L^\Phi(\mu)$ is finite and globally (for $x_0 = 0$) when the measure is infinite.

Definition 4: (Orlicz space) The function space

$$L^\Phi(\mu) = \left\{ f\colon \Omega \to \overline{\mathbb{C}} \text{ (equivalence classes of) } \Sigma\text{-measurable: } \int_\Omega \Phi\left(\alpha|f|\right)\, \mathrm{d}\mu < \infty \text{ for at least one } \alpha > 0 \right\}$$

with $\Phi : \mathbb{R} \to \mathbb{R}^+$ a Young function, is called *Orlicz space*.

We next define norms on $L^\Phi(\mu)$.

Definition 5: (Gauge norm and Orlicz norm) The norm

$$N_\Phi(f) = \inf\left\{ k > 0 : \int_\Omega \Phi\left(\frac{|f|}{k}\right)\, \mathrm{d}\mu \leq 1 \right\}$$

is called *gauge norm* of the Orlicz space $L^\Phi(\mu)$ for a Young function $\Phi : \mathbb{R} \to \mathbb{R}^+$.
By using the complementary Young function we can define another norm on $L^\Phi(\mu)$.

Let (Φ, Ψ) be a complementary pair of Young functions, then we define the *Orlicz norm* as:

$$\|.\|_{L^\Phi}\colon f \mapsto \|f\|_{L^\Phi} = \sup\left\{\int_\Omega |fg| \,\mathrm{d}\mu : \int_\Omega \Psi(|g|) \,\mathrm{d}\mu \leq 1\right\}.$$

The two norms defined on the Orlicz spaces are equivalent, furthermore the Orlicz spaces with the corresponding norms are Banach spaces.

Theorem 1: [8] [Proposition 4 3.3.III, Corollary 12 III.3.3] Let (Ω, Σ, μ) be a measure space, (Φ, Ψ) be a complementary Young pair, then $N_\Phi(f) \leq \|f\|_{L^\Phi} \leq 2N_\Phi(f)$ for $f \in L^\Phi(\mu)$. $\left(L^\Phi(\mu), N_\Phi(\cdot)\right)$ and $\left(L^\Phi(\mu), \|\cdot\|_{L^\Phi}\right)$ are Banach spaces.

Since it is often useful to work with duality arguments in proofs, we give a characterisation of the dual space to the Orlicz space in the next theorem.

Theorem 2: [8] [Theorem 7, Corollary 9 IV.4.1] Let (Φ, Ψ) be a complementary Young pair and Φ be Δ_2-regular and (Ω, Σ, μ) be σ-finite. Then $\left(L^\Phi(\mu)\right)^*$ is isometrically isomorphic to $L^\Psi(\mu)$.

We next extend the Orlicz space theory of $\mathbb{C}$-valued functions $f : \Omega \subset \mathbb{R}^d \to \mathbb{C}$ to functions $f : \Omega \subset \mathbb{R}^d \to X$ whose values lie in a Banach space X. Candidates for X are Orlicz spaces L^{Φ_2} associated to a Young function Φ_2.

Definition 6: (Mixed-norm Orlicz space) Let $(\Omega_i, \Sigma_i, \mu_i)$ be measure spaces, (Φ_i, Ψ_i) be complementary Young pairs for $i = 1, 2$. Then the *mixed-norm Orlicz space* is

$$\begin{aligned} L^{\Phi_1\Phi_2} &= L^{\Phi_1}(\mu_1, L^{\Phi_2}(\mu_2)) \\ &= \Bigg\{ f\colon \Omega_1 \to L^{\Phi_2}(\mu_2) \text{ strongly measurable on } (\Omega_1, \Sigma_1, \mu_1)\colon \\ &\qquad \int_{\Omega_1} \Phi_1(\alpha N_{\Phi_2}(f)) \,\mathrm{d}\mu_1 < \infty \text{ for some } \alpha > 0 \Bigg\}. \end{aligned}$$

The corresponding *gauge norm* $N_{\Phi_1\Phi_2}(\cdot) = N_{\Phi_1}(N_{\Phi_2}(\cdot))$ is given by:

$$N_{\Phi_1\Phi_2}(f) = \inf\Bigg\{k > 0 : \int_{\Omega_1} \Phi_1\left(\frac{|N_{\Phi_2}(f(\cdot, w_1))|}{k}\right) \mathrm{d}\mu_1(w_1) \leq 1\Bigg\}.$$

The *Orlicz norm* is similarly defined by

$$\|f\|_{\Phi_1\Phi_2} = \sup\Bigg\{\int_{\Omega_1}\Big|\|f(\cdot, w_1)\|_{L^{\Phi_2}} \cdot g(w_1)\Big| \,\mathrm{d}\mu_1(w_1) : \int_{\Omega_1} \Psi_1(|g(w_1)|) \,\mathrm{d}\mu_1(w_1) \leq 1\Bigg\},$$

As in the case of the Orlicz spaces the mixed-norm Orlicz spaces are also Banach spaces and it can be shown that the norms are equivalent.

Theorem 3: Let $(\Omega_i, \Sigma_i, \mu_i)$ be measure spaces, (Φ_i, Ψ_i) be complementary Young pairs for $i = 1, 2$, then $(L^{\Phi_1}(\mu_1, L^{\Phi_2}(\mu_2)), N_{\Phi_1\Phi_2}(\cdot))$ and $(L^{\Phi_1}(\mu_1, L^{\Phi_2}(\mu_2)), \|\cdot\|_{L^{\Phi_1\Phi_2}})$ are Banach spaces and the norms are equivalent. Furthermore it follows

$$N_{\Phi_1\Phi_2}(f) \leq \|f\|_{L^{\Phi_1\Phi_2}} \leq 4N_{\Phi_1\Phi_2}(f) \text{ for } f \in L^{\Phi_1\Phi_2}.$$

If we assume that the Young functions are also strictly convex the dual space to L^{Φ_1,Φ_2} is isometrically isomorphic to the space L^{Ψ_1,Ψ_2} to the complementary functions.

Theorem 4: [8] [Theorem 4 VII.7.5] Let $(\Omega_i, \Sigma_i, \mu_i)$ be measure spaces, (Φ_i, Ψ_i) be complementary Young pairs which are Δ_2-regular and strictly convex for $i = 1, 2$. Then $\left(L^{\Phi_1\Phi_2}\right)^*$ is isometrically isomorphic to $L^{\Psi_1\Psi_2}$.

B. Useful properties for time frequency analysis

In this section we list properties of the Orlicz spaces which are useful for time-frequency analysis. At first we mention that the Orlicz norm and the mixed Orlicz norm are invariant under translations, if the measure spaces are the Lebesgue space $(\Omega_i, \Sigma_i, \mu_i) = (\mathbb{R}^d, \mathcal{B}^d, \boldsymbol{\lambda}^d)$ for $i = 1, 2$.

Lemma 1: Let Φ_i be Young functions for $i = 1, 2$, then $L^{\Phi_1}(\boldsymbol{\lambda}^d)$ and $L^{\Phi_1\Phi_2}(\boldsymbol{\lambda}^{2d}) = L^{\Phi_1}(\boldsymbol{\lambda}^d, L^{\Phi_2}(\boldsymbol{\lambda}^d))$ are invariant under $T_z F := F(\cdot - z)$ and we have

$$N_{\Phi_1}(T_z F) = N_{\Phi_1}(F) \text{ for } F \in L^{\Phi_1}(\boldsymbol{\lambda}^d), z \in \mathbb{R}^d \quad \text{and}$$
$$N_{\Phi_1\Phi_2}(T_z F) = N_{\Phi_1\Phi_2}(F) \text{ for } F \in L^{\Phi_1\Phi_2}(\boldsymbol{\lambda}^{2d}), z \in \mathbb{R}^{2d}.$$

Futher one can also prove a Hölder inequality for Orlicz spaces.

Lemma 2: (Hölder inequality)[8] [Proposition 1 III.3.3] Let $(\Omega_i, \Sigma_i, \mu_i) = (\mathbb{R}^d, \mathcal{B}^d, \boldsymbol{\lambda}^d)$ and (Φ_i, Ψ_i) be complementary Young pairs for $i = 1, 2$. If $F \in L^{\Phi_1}(\boldsymbol{\lambda}^d)$ and $G \in L^{\Psi_1}(\boldsymbol{\lambda}^d)$, then one has $\displaystyle\int_{\mathbb{R}^d} |F \cdot G| \,\mathrm{d}\boldsymbol{\lambda}^d \leq 2 \cdot N_{\Phi_1}(F) N_{\Psi_1}(G)$.

If we assume in addition that Φ_2 is Δ_2-regular, then one has for $F \in L^{\Phi_1\Phi_2}(\boldsymbol{\lambda}^{2d})$ and $G \in L^{\Psi_1\Psi_2}(\boldsymbol{\lambda}^{2d})$ the estimate

$$\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} |F \cdot G| \,\mathrm{d}\boldsymbol{\lambda}^d \,\mathrm{d}\boldsymbol{\lambda}^d \leq 4 \cdot N_{\Phi_1\Phi_2}(F) N_{\Psi_1\Psi_2}(G).$$

Now, we have a look at inclusion properties. If Φ is continuous the Schwartz class $\mathcal{S}(\mathbb{R}^d)$ is embedded into the Orlicz space $L^\Phi(\boldsymbol{\lambda}^d)$ and if also the complementary function Ψ is continuous then the functions in the Orlicz space define tempered distributions.

Lemma 3: Let (Φ_i, Ψ_i) be pairs of complementary Young functions and Φ_i be continuous for $i = 1, 2$, then

$$\begin{aligned} &\mathcal{S}(\mathbb{R}^d) \subset L^{\Phi_1}(\boldsymbol{\lambda}^d), \\ \text{and} \quad &L^{\Phi_1}(\boldsymbol{\lambda}^d) \subset \mathcal{S}'(\mathbb{R}^d), \quad \text{if } \Psi_1 \text{ is continuous.} \\ \text{And} \quad &\mathcal{S}(\mathbb{R}^{2d}) \subset L^{\Phi_1\Phi_2}(\boldsymbol{\lambda}^{2d}), \\ \text{and} \quad &L^{\Phi_1\Phi_2}(\boldsymbol{\lambda}^{2d}) \subset \mathcal{S}'(\mathbb{R}^{2d}), \quad \text{if } \Psi_1, \Psi_2 \text{ are continuous.} \end{aligned}$$

With the fact that $(L^\Phi)^* \cong L^\Psi$, we can extend a well known convolution relation $L^1(\mathbb{R}^d) * L^p(\mathbb{R}^d) \subset L^p(\mathbb{R}^d)$ of the Lebesgue spaces to the Orlicz spaces. Further one can prove the following Young inequality.

Theorem 5: If $F \in L^1(\mathbb{R}^{2d})$, $G \in L^\Phi(\boldsymbol{\lambda}^{2d})$ and Φ is a Δ_2-regular Young function, then

$$\|F * G\|_{L^\Phi} \leq 2\|F\|_{L^1}\|G\|_{L^\Phi}.$$

If $F \in L^{1,1}(\mathbb{R}^{2d})$, $G \in L^{\Phi_1\Phi_2}$ and Φ_i are Δ_2-regular and strictly convex Young functions for $i = 1, 2$, then

$$\|F * G\|_{L^{\Phi_1\Phi_2}} \leq 4\|F\|_{L^{1,1}}\|G\|_{L^{\Phi_1\Phi_2}}.$$

III. Orlicz Modulation spaces and mixed-norm Orlicz modulation spaces

We now have all the tools in place that we need to define and analyse the modulation space associated to the Orlicz space.

Definition 7: (Orlicz modulation space)
Fix a non-zero window $g \in \mathcal{S}(\mathbb{R}^d)$ and a Young function Φ. Then the *Orlicz modulation space* $M^\Phi(\mathbb{R}^d)$ is defined by

$$M^\Phi(\mathbb{R}^d) = \{f \in \mathcal{S}'(\mathbb{R}^d) : V_g f \in L^\Phi(\mathbb{R}^{2d})\}.$$

The norm on M^Φ is $\|f\|_{M^\Phi} = \|V_g f\|_{L^\Phi}$.

In the same way we define the mixed-norm Orlicz modulation space.
Therefore we replace only the Orlicz space L^Φ by the mixed-norm Orlicz space $L^{\Phi_1\Phi_2}$.

Definition 8: (Mixed-norm Orlicz modulation space)
Fix a non-zero window $g \in \mathcal{S}(\mathbb{R}^d)$ and Young functions Φ_i for $i = 1, 2$. Then the *Orlicz modulation space* $M^{\Phi_1\Phi_2}(\mathbb{R}^d)$ is defined by

$$M^{\Phi_1\Phi_2}(\mathbb{R}^d) = \{f \in \mathcal{S}'(\mathbb{R}^d) : V_g f \in L^{\Phi_1\Phi_2}(\mathbb{R}^{2d})\}.$$

The norm on $M^{\Phi_1\Phi_2}$ is $\|f\|_{M^{\Phi_1\Phi_2}} = \|V_g f\|_{L^{\Phi_1\Phi_2}}$.

Remark 1: Modulation spaces are a special case of the coorbit spaces defined by H. Feichtinger and K.H. Gröchenig [1], and Orlicz spaces are mentioned, without proof, as classes of Banach function spaces Y suitable to define coorbit spaces CoY. In this paper we make this remark more explicit by providing additional details such as associated discrete coefficient spaces, relationship to tempered distributions, dual spaces, etc. We would also like to point out that to our knowledge, mixed-norm Orlicz spaces have not been considered previously.

Now we analyse a few properties of the Orlicz modulation spaces. We start with the observation that the definitions of these spaces are independent of the choice of a window g. In addition, if the Young function is Δ_2-regular, these spaces are also Banach spaces.

Theorem 6: Assume that Φ is a Δ_2-regular Young function and its complementary function Ψ is continuous. Then the definition of $M^\Phi(\mathbb{R}^d)$ is independent of the window $g \in \mathcal{S}(\mathbb{R}^d)$ and $M^\Phi(\mathbb{R}^d)$ is a Banach space.

If we assume that the Young functions are also strictly convex, we can show an analogous statement for the mixed-norm Orlicz spaces.

Theorem 7: Let (Φ_i, Ψ_i) be complementary Young pairs which are Δ_2-regular, strictly convex and continuous for $i = 1, 2$. Then the definition of $M^{\Phi_1\Phi_2}(\mathbb{R}^d)$ is independent of the window $g \in \mathcal{S}(\mathbb{R}^d)$ and $M^{\Phi_1\Phi_2}(\mathbb{R}^d)$ is a Banach space.

Furthermore, the duality between the Orlicz spaces L^Φ and L^Ψ suggests a similar statement for their modulation spaces. This can be proved in the following theorem by using the Δ_2-condition for the Young function.

Theorem 8: If (Φ, Ψ) is a complementary Young pair and if Φ is Δ_2-regular and continuous, then $\left(M^\Phi(\mathbb{R}^d)\right)^* \cong M^\Psi(\mathbb{R}^d)$ under the duality

$$\langle f, h\rangle = \iint_{\mathbb{R}^{2d}} V_{g_0} f(z)\overline{V_{g_0} h(z)}\,\mathrm{d}z$$

for $f \in M^\Phi(\mathbb{R}^d)$ and $h \in M^\Psi(\mathbb{R}^d)$, $g_0 \in \mathcal{S}(\mathbb{R}^d)$.
Let (Φ_i, Ψ_i) be complementary Young pairs which are Δ_2-regular, strictly convex and continuous for $i = 1, 2$. Then $\left(M^{\Phi_1\Phi_2}(\mathbb{R}^d)\right)^* \cong M^{\Psi_1\Psi_2}(\mathbb{R}^d)$ under the duality

$$\langle f, h\rangle = \iint_{\mathbb{R}^{2d}} V_{g_0} f(z)\overline{V_{g_0} h(z)}\,\mathrm{d}z$$

for $f \in M^{\Phi_1\Phi_2}(\mathbb{R}^d)$ and $h \in M^{\Psi_1\Psi_2}(\mathbb{R}^d)$, $g_0 \in \mathcal{S}(\mathbb{R}^d)$.

IV. Discrete Orlicz space and discrete mixed-norm Orlicz space

This space consists of all sequences for which the discrete norm defined by the next definition is finite.

Definition 9: (Discrete Orlicz space) Let Φ be a Young function, then the *discrete Orlicz space* is defined by

$$l^\Phi(\mathbb{Z}^d) = \{a = (a_n)_{n\in\mathbb{Z}^d} : n_\Phi(a) < \infty\},$$

where $$n_\Phi(a) = \inf\left\{\lambda > 0 : \sum_{n\in\mathbb{Z}^d} \Phi\left(\frac{|a_n|}{\lambda}\right) \le 1\right\}.$$

Definition 10: (Discrete mixed-norm Orlicz space) Let Φ_1, Φ_2 be Young functions, then the *discrete mixed-norm Orlicz space* is defined by

$$l^{\Phi_1\Phi_2}(\mathbb{Z}^{2d}) = \{a = (a_{kn})_{k,n\in\mathbb{Z}^d} : n_{\Phi_1\Phi_2}(a) < \infty\},$$

where
$$n_{\Phi_1\Phi_2}(a) = \inf\left\{\lambda > 0 : \sum_{k,n\in\mathbb{Z}^d} \Phi_1\left(\frac{n_{\Phi_2}(|a_{kn}|)}{\lambda}\right) \le 1\right\}.$$

With these definitions we can apply the theory of Atomic Decomposition of H. G. Feichtinger and K. H. Gröchenig presented in the paper [1]. In the context of Orlicz modulation spaces we get the following result.

Theorem 9: (The Atomic Decomposition in M^Φ) [1] Let Φ be a Δ_2-regular Young function. For any $g \in \mathcal{S}(\mathbb{R}^d)$ there exist positive constants C_0 and C_1 (depending only on g) and a neighbourhood U of the identity such that for an arbitrary U-dense and relatively separated family $X = (x_i)_{i\in I} \subset \mathbb{R}^{2d}$ the following is true:

1) Analysis: There exists a bounded linear operator $A : M^\Phi \to l^\Phi(X)$, i.e., writing $\Lambda := (\lambda_i)_{i\in I} := A(f)$ one has $n_\Phi(\Lambda) \le C_0\|f\|_{M^\Phi}$, such that every $f \in M^\Phi$ can be represented as $f = \sum_{i\in I} \lambda_i \rho(x_i) g$, where ρ is the Schrödinger representation.
2) Synthesis: Conversely, assuming that $X = (x_i)_{i\in I}$ is relatively separated, every $\Lambda \in l^\Phi$ defines an element $f = \sum_{i\in I} \lambda_i \rho(x_i) g$ in M^Φ with $\|f\|_{M^\Phi} \le C_1 n_\Phi(\Lambda)$.

In both cases convergence takes place in the norm of M^Φ.

Moreover by using the results in [3] of H.G. Feichtinger, K. H. Gröchenig and D. Walnut the orthonormal Wilson bases are unconditional bases for some Orlicz modulation spaces. Consequently in these cases M^Φ and l^Φ are isomorphic Banach spaces. Simple Wilson bases of exponential type are given by the following construction.

Definition 11: [3] A real-valued function ψ constructed, such that $|\psi(x)| \leq Ce^{-a|x|}$ and $|\hat{\psi}(t)| \leq Ce^{-b|t|}$ and such that the $\psi_{ln}, l \in \mathbb{N}, n \in \mathbb{Z}$, defined by
$\psi_{0n}(x) = \psi(x-n)$
$\psi_{ln}(x) = \sqrt{2}\psi\left(x - \frac{n}{2}\right)\cos(2\pi lx) \quad l \neq 0, l+n \in 2\mathbb{Z}$
$\psi_{ln}(x) = \sqrt{2}\psi\left(x - \frac{n}{2}\right)\sin(2\pi lx) \quad l \neq 0, l+n \in 2\mathbb{Z}+1$
constitute an orthonormal basis for $L^2(\mathbb{R})$.

In their work [3], H.G. Feichtinger, K. H. Gröchenig and D. Walnut use the density of the functions with compact support in the Banach function space. The following lemma gives a characterisation of this density for Orlicz spaces.

Lemma 4: Let Φ be a Young function, $\Phi(x) = 0$ if and only if $x = 0$, and $L^{\Phi}(\mathbb{R}^2)$ be the associated Orlicz space on $\mathbb{R}^2$. Then the bounded functions with compact support are dense in $L^{\Phi}(\mathbb{R}^2)$ if the Young function satisfies the Δ_2 condition.

If we have the Δ_2-regularity of the Young function it follows from [3].

Theorem 10: Assume that the Young function Φ satisfies the Δ_2 condition. Then the orthonormal Wilson bases are unconditional bases for $M^{\Phi}(\mathbb{R})$.

Moreover we can characterise inclusion properties of Orlicz modulation spaces by using properties of the corresponding Orlicz sequence spaces as in [2]. Additionally we can translate this to a comparison of Young functions.

Theorem 11: Let $\Phi_1, \Phi_2, \Phi_1', \Phi_2'$ unbounded Young functions. Then $M^{\Phi_1\Phi_1'} \subset M^{\Phi_2\Phi_2'}$ if and only if $l^{\Phi_1\Phi_1'} \subset l^{\Phi_2\Phi_2'}$ if and only if there are constants $C_1, C_2 > 0$ and $t_1, t_2 \geq 0$ such that $\Phi_2(t) \leq C_1\Phi_1(t)$ for all $0 \leq t \leq t_1$ and $\Phi_2'(t) \leq C_2\Phi_1'(t)$ for all $0 \leq t \leq t_2$.

Next, we wanted to give, without proof, an example of an embedding relation between Fourier-Lebesgue spaces and a concrete Orlicz modulation spaces. This result is an extension of the embedding theorems that Y.V. Galperin and K.H. Gröchenig gave in her work [5].

Theorem 12: Suppose that $g \in \mathcal{S}(\mathbb{R}^d), f \in \mathcal{S}'(\mathbb{R}^d)$, $C > 0, N \geq 0$ and $|V_g f(x,w)| \leq C(1+|x|+|w|)^N$ for alle $x, w \in \mathbb{R}^d$ and $0 < p \leq 2, p \leq r, s \leq 2, \frac{1}{s} + \frac{1}{s'} = 1, \frac{1}{r} + \frac{1}{r'} = 1$. If

$$\left(\frac{ap-N}{pd} + \frac{1}{r} - \frac{1}{p}\right)\left(\frac{bp-N}{pd} + \frac{1}{s} - \frac{1}{p}\right) > \left(\frac{N}{pd} + \frac{1}{p} - \frac{1}{s'}\right)\left(\frac{N}{pd} + \frac{1}{p} - \frac{1}{r'}\right)$$

with all factors positive, then $L^r_a \cap \mathcal{F}L^s_b \hookrightarrow M^{p,p}_{\frac{N}{p},\frac{N}{p}} \subset M_{l^p \ln l}$.

V. Conclusion and outlook

In this work we have presented and analysed modulation spaces associated to Orlicz spaces and mixed-norm Orlicz spaces. It is possible to extend the theory of modulation spaces associated to Lebesgue space (understood as spaces of tempered distributions) to more general Orlicz spaces and mixed-norm Orlicz spaces. For some results the adaptation was straightforward, but in other cases further conditions on the Young function, in particular the Δ_2 condition, are necessary to obtain analogs to the results known for classical modulation spaces.

The most general approach to modulation spaces follows [1] and [2]. Here, the Δ_2 condition is needed to characterise duals of Orlicz modulation spaces (in particular in the mixed-norm setting), but also to establish density of bounded functions with compact support in the Orlicz space (needed, e.g., in *Lemma 4*, and subsequently in *Theorem 10*).
A more accessible, but less general approach is developed in [6]. The adaptation of the arguments in [6] is often feasible (and instructive), however, since duality plays a stronger role here, the Δ_2 condition is needed more often than in the general case.

Furthermore we derive embedding results between Orlicz modulation spaces by using the discretisation of the Orlicz spaces, especially by using comparison of Young functions.
For a special Orlicz modulation space we can also give an embedding relation of Fourier-Lebesgue spaces into this Orlicz modulation space. But at this time it isn't clear if this result has also an interpretation as uncertainty principles as in [5].
Another topic of interest for further work are relations applications to entropy estimates.

Acknowledgment

The authors would like to thank H. G. Feichtinger for his helpful comments.

References

[1] H. G. Feichtinger and K. H. Gröchenig, *Banach spaces related to integrable group representations and their atomic decompositions. I*, J. Funct. Anal., 86(1989),307-340.
[2] H. G. Feichtinger and K. H. Gröchenig, *Banach spaces related to integrable group representations and their atomic decompositions. II*, Monatsth. Math., 108(1989),129-148.
[3] H. G. Feichtinger, K. H. Gröchenig and D. Walnut, *Wilson bases and modulation spaces*, Math. Nachrichten, 155(1992), 7-17.
[4] G. B. Folland and A. Sitaram, *The Uncertainty Principle: A Mathematical Survey*, J. Fourier Anal. and Appl. 3 (1997), 207-238.
[5] Y. V. Galperin and K. H. Gröchenig, *Uncertainty principles as embeddings of modulation spaces*, J. Math. Anal. Appl., 274 (2002),181-202.
[6] K. H. Gröchenig, *Foundations of time-frequency analysis*, Applied and Numerical Harmonic Analysis, Birkhäuser Boston Inc., Boston, MA, 2001.
[7] M. A. Krasnosel'skiĭ and YA. B. Rutickiĭ, *Convex functions and Orlicz spaces*, P. Noordhoff Ltd., Groningen, 1961.
[8] M. M. Rao and Z. D. Ren, *Theory of Orlicz spaces*, New York: Marcel Dekker Inc., 1991.
[9] M. Růžička, *Nichtlineare Funktionalanalysis: Eine Einführung*, Springer-Lehrbuch Masterclass, Springer, 2004.

Binary Reduced Row Echelon Form Approach for Subspace Segmentation

Akram Aldroubi
Department of Mathematics
Vanderbilt University, Nashville, TN, 37212 USA
Email: akram.aldroubi@vanderbilt.edu

Ali Sekmen
Department of Computer Science
Tennessee State University, Nashville, TN 37209
Email:asekmen@tnstate.edu

***Abstract*—This paper introduces a subspace segmentation and data clustering method for a set of data drawn from a union of subspaces. The proposed method works perfectly in absence of noise, i.e., it can find the number of subspaces, their dimensions, and an orthonormal basis for each subspace. The effect of noise on this approach depends on the noise level and relative positions of subspaces. We provide a performance analysis in presence of noise and outliers.**

I. Introduction

The goal of subspace clustering is to identify all of the subspaces that a set of data $\mathbf{W} = \{w_1, ..., w_N\} \in \mathbb{R}^D$ is drawn from and assign each data point w_i to the subspace it belongs to. The number of subspaces, their dimensions, and a basis for each subspace are to be determined even in presence of noise, missing data, and outliers. In some subspace clustering problems, the number M of subspaces or the dimensions of the subspaces $\{d_i\}_{i=1}^M$ are known. A number of approaches have been devised to solve the problem above or some of its special cases. They are based on sparsity methods [1], [2], [3], [4], algebraic methods [5], [6], iterative and statistical methods [7], [8], [9], [10], [11], [12], and spectral clustering methods [2], [3], [13], [14], [15], [16], [17], [18], [19].

In this work, we develop an algebraic method for solving the general subspace segmentation problem for noiseless data. For the case where all the subspaces are four dimensional, Gear observed, without proof, that the reduced echelon form can be used to segment motions in videos [20]. In this paper, we develop this idea and show that the reduced row echelon form can completely solve the subspace segmentation problem in its most general version for noiseless data. For noisy data, the reduced echelon form method does not work, and a thresholding must be applied. The effect of the noise on the reduced echelon form method depends on the noise level and the relative positions of the subspaces.

A. Non-Linear Approximation Formulation

When M is known, the subspace segmentation problem, for both the finite and infinite dimensional space cases, can be formulated as follows:

Let $\mathcal{B}$ be a Banach space, $\mathbf{W} = \{w_1, \dots, w_N\}$ a finite set of vectors in $\mathcal{B}$. For $i = 1, \dots, M$, let $\mathcal{C} = C_1 \times C_2 \times \cdots \times C_M$ be the cartesian product of M family C_i of closed subspaces of $\mathcal{B}$ each containing the trivial subspace $\{0\}$. Thus, an element $\mathbf{S} \in \mathcal{C}$ is a sequence $\{S_1, \dots, S_M\}$ of M subspaces of $\mathcal{B}$ with $S_i \in C_i$. An example for finite dimensions is when $\mathcal{B} = \mathbb{R}^D$ and $\mathcal{C}$ is the family of all subspaces of $\mathbb{R}^D$ of dimensions less than or equal to D. An example for infinite dimensions is when $\mathcal{B} = L^2(\mathbb{R}^D)$ and $\mathcal{C}$ is a family of closed, shift-invariant subspaces of $L^2(\mathbb{R}^D)$ that are generated by finite generators.

Problem 1.

1) *Given a finite set $\mathbf{W} \subset \mathcal{B}$, a fixed p with $0 < p \leq \infty$, and a fixed integer $M \geq 1$, find the infimum of the expression*

$$e(\mathbf{W}, \mathbf{S}) := \sum_{w \in \mathbf{W}} \min_{1 \leq j \leq M} d^p(w, S_j),$$

over $\mathbf{S} = \{S_1, \dots, S_M\} \in \mathcal{C}$, and $d(x, y) := \|x - y\|_{\mathcal{B}}$.

2) *Find a sequence of M-subspaces $\mathbf{S}^o = \{S_1^o, \dots, S_M^o\} \in \mathcal{C}$ (if it exists) such that*

$$e(\mathbf{W}, \mathbf{S}^o) = \inf\{e(\mathbf{W}, \mathbf{S}) : \mathbf{S} \in \mathcal{C}\}. \qquad \text{(I.1)}$$

In the presence of outliers, it is shown that $p = 1$ is a good choice [21] and a good choice for light-tailed noise is $p = 2$. The necessary and sufficient conditions for the existence of a solution when $p = 2$ and $\mathcal{B}$ is a Hilbert space can be found in [22].

Definition 1. *For $0 < p \leq \infty$, a set of closed subspaces C of a Banach space $\mathcal{B}$ has the Minimum Subspace Approximation Property p-(MSAP) if for every finite subset $\mathbf{W} \subset \mathcal{B}$ there exists an element $S \in C$ that minimizes the expression $e(\mathbf{W}, S) = \sum_{w \in \mathbf{W}} d^p(w, S)$ over all $S \in C$.*

Under the assumption that each family of subspaces C_i satisfies p-(MSAP), problem 1 has a minimizer [23]:

Theorem 1. *If for each $i = 1, \dots M$, C_i satisfies p-(MSAP), then Problem 1 has a minimizing set of subspaces for all finite sets of data.*

Theorem 1 suggests an iterative search algorithm for the optimal solution $\mathbf{S}^o$. Obviously, this solution can be obtained by Algorithm 1. This algorithm will work well if a good initial partition is chosen. Otherwise, the algorithm may terminate in a local minimum instead of the global minimum.

Algorithm 1 Optimal Solution $\mathbf{S}^o$

Pick any partition $P \in \mathcal{P}(\mathbf{W})$
For each subset $\mathbf{W}_i$ in the partition P find the subspace $S_i^o(P) \in C_i$ that minimizes the expression $e(\mathbf{W}_i, S) = \sum_{w \in \mathbf{W}_i} d^p(w, S)$
while $\sum_{i=1}^{M} e(\mathbf{W}_i, S_i^o(P)) > e(\mathbf{W}, \mathbf{S}^o(P))$ **do**
  for all i from 1 to M **do**
    Update $\mathbf{W}_i = \{w \in \mathbf{W} : d(w, S_i^o(P)) \leq d(w, S_k^o(P)),\ k = 1, \ldots, M\}$
    Update $S_i^o(P) = \underset{S \in C_i}{\operatorname{argmin}}\, e(\mathbf{W}_i, S)$
  end for
  Update $P = \{\mathbf{W}_1, \ldots, \mathbf{W}_M\}$
end while
$\mathbf{S}^o = \{S_1^o(P), \ldots, S_M^o(P)\}$

II. Subspace Segmentation - Noiseless Case

In this section we consider the problem in which a set of vectors $\mathbf{W} = \{w_1, \ldots, w_N\}$ are drawn from a union $\mathcal{U} = \bigcup_{i \in I} S_i$ of M subspaces $S_i \in \mathbb{R}^D$ of dimension d_i. In order to find the M subspaces from the data set $\mathbf{W}$ it is clear that we need enough vectors $\mathbf{W} = \{w_1, \ldots, w_N\}$. In particular for the problem of subspace segmentation, it is necessary that the set $\mathbf{W}$ can be partitioned into M sets $\mathbf{W} = \{\mathbf{W}_1, \ldots, \mathbf{W}_M\}$ such that span $\mathbf{W}_i = S_i$, $i = 1, \ldots, M$. Thus, we need to assume that we have enough data for solving the problem. In particular, we assume that any $k \leq d$ vectors drawn from a subspace S of dimension d are linearly independent, and we make the following definition.

Definition 1. *Let S be a linear subspace of $\mathbb{R}^D$ with dimension d. A set of data $\mathbf{W}$ drawn from $S \subset \mathbb{R}^D$ with dimension d is said to be* generic *if (i) $|\mathbf{W}| > d$, and (ii) every d vectors from $\mathbf{W}$ form a basis for S.*

Another assumption that we will make is that the union of subspaces $\mathcal{U} = \bigcup_{i \in I} S_i$ from which the data is drawn consists of independent subspaces:

Definition 2. *(Independent Subspaces) Subspaces $\{S_i \subset \mathbb{R}^D\}_{i=1}^n$ are called independent if $dim(S_1 + \cdots + S_n) = dim(S_1) + \cdots + dim(S_n)$.*

Definition 3. *Matrix R is said to be the* binary reduced row echelon form *of matrix A if all non-pivot column vectors are converted to binary vectors, i.e., non-zero entries are set to one.*

The following theorem suggests a very simple yet effective approach to cluster the data points. The proofs of the following Theorems can be found in [23].

Theorem 2. *Let $\{S_i\}_{i=1}^M$ be a set of non-trivial linearly independent subspaces of $\mathbb{R}^D$ with corresponding dimensions $\{d_i\}_{i=1}^M$. Let $\mathbf{W} = [w_1 \cdots w_N] \in \mathbb{R}^{D \times N}$ be a matrix whose columns are drawn from $\bigcup_{i=1}^M S_i$. Assume the data is drawn from each subspace and that it is generic. Let Brref($\mathbf{W}$) be the binary reduced row echelon form of $\mathbf{W}$. Then*

1) *The inner product $\langle e_i, b_j \rangle$ of a pivot column e_i and a non-pivot column b_j in Brref($\mathbf{W}$) is one, if and only if the corresponding column vectors $\{w_i, w_j\}$ in $\mathbf{W}$ belong to the same subspace S_l for some $l = 1, \ldots, M$.*
2) *Moreover, $\dim(S_l) = \|b_j\|_1$, where $\|b_j\|_1$ is the l_1-norm of b_j.*
3) *Finally, $w_p \in S_l$ if and only if $b_p = b_j$ or $\langle b_p, b_j \rangle = 1$.*

The data $\mathbf{W}$ can be partitioned into M clusters $\{\mathbf{W}_1, \ldots, \mathbf{W}_M\}$, such that span$\mathbf{W}_l = S_l$. The clusters can be formed as follows: Pick a non-pivot element b_j in Brref($\mathbf{W}$), and group together all columns b_p in Brref($\mathbf{W}$) such that $\langle b_j, b_p \rangle > 0$. Repeat the process with a different non-pivot column until all columns are exhausted.

III. Subspace Segmentation - Noisy Case

In practice the data $\mathbf{W}$ is corrupted by noise. In this case, the Reduced Row Echelon Form (RREF)-based algorithm cannot work, even under the assumption of Theorem 2, since the noise will have two effects: 1) The rank of the data corrupted by noise $\mathbf{W} + \eta \subset \mathbb{R}^D$ becomes full; i.e., $\operatorname{rank}(\mathbf{W} + \eta) = D$; and 2) Even under the assumption that $r = D$, none of the entries of the non-pivot columns of rref$(\mathbf{W} + \eta)$ will be zero. One way of circumventing this problem, is to use the RREF-based algorithm in combination with thresholding to set to zero those entries that are small. The choice of the threshold depends on the noise characteristics and the position of the subspaces relative to each other.

In general, $dim(\sum_{i=1}^M S_i) = \operatorname{rank}(\mathbf{W}) \leq D$, where D is the dimension of the ambient space $\mathbb{R}^D$. After projection of $\mathbf{W}$, the new ambient space is isomorphic to $\mathbb{R}^r$, where $r = \operatorname{rank}(\mathbf{W})$, and we may assume that $\operatorname{rank}(\mathbf{W}) = D$. Without loss of generality, let us assume that $\mathbf{W} = \begin{bmatrix} A & B \end{bmatrix}$ where the columns of A form basis for $\mathbb{R}^D$, i.e., the columns of A consist of d_i linearly independent vectors from each subspace S_i, $i = 1, \ldots, M$. Let $\widetilde{\mathbf{W}} = \mathbf{W} + \mathbf{N}$ be the data with additive noise. Then the reduced echelon form applied to $\widetilde{\mathbf{W}}$ is given by rref$(\widetilde{\mathbf{W}}) = \begin{bmatrix} I & \widetilde{A}^{-1}\widetilde{B} \end{bmatrix}$. Let b_i and $\tilde{b}_i$ denote the columns of B and $\widetilde{B}$ respectively, $e_i = \widetilde{A}^{-1}\tilde{b}_i - A^{-1}b_i$, $\Delta = \widetilde{A} - A$, and $\nu_i = \tilde{b}_i - b_i$. Let σ_{min} denote the smallest singular value of A, then if $||\Delta|| \leq \sigma_{min}(A)$, we get

$$\|e_i\|_2 \leq \frac{\|\nu_i\|_2}{\sigma_{min}(A)} + \frac{\|\Delta\|}{\sigma_{min}^2(A)} \left(\frac{1}{1 - \frac{\|\Delta\|}{\sigma_{min}(A)}} \right) (\|b_i\|_2 + \|\nu_i\|_2), \tag{III.1}$$

where $\|\cdot\|$ denotes the operator norm $\|\cdot\|_{\ell^2 \to \ell^2}$. Unless specified otherwise, the noise $\mathbf{N}$ will be assumed to consist of entries that are i.i.d. $\mathcal{N}(0, \sigma^2)$ Gaussian noise with zero mean and variance σ^2. For this case, the expected value of $\|\Delta\|$ can be estimated by $\mathbb{E}\|\Delta\| \leq C\sqrt{D}\sigma$ [24]. Note that to estimate the error in (III.1) we still need to estimate $\sigma_{min}(A)$. This singular value depends on the position of the subspaces $\{S_i\}_{i=1}^M$ relative to each other which can be

measured by the principle angles between them. The principle angles between two subspaces $\mathcal{F}, \mathcal{G}$, can be obtained using any pair of orthogonal bases for $\mathcal{F}, \mathcal{G}$ as described in the following Lemma [25]:

Lemma 1. *Let $\mathcal{F}$ and $\mathcal{G}$ be two subspaces of $\mathbb{R}^D$ with $p = dim(\mathcal{F}) \leq dim(\mathcal{G}) = q$. Assume that $Q_{\mathcal{F}} \in \mathbb{R}^{D\times p}$ and $Q_{\mathcal{G}} \in \mathbb{R}^{D\times q}$ are matrices whose columns form orthonormal bases for the subspaces $\mathcal{F}$ and $\mathcal{G}$. If $1 \geq \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0$ are the singular values of $Q_{\mathcal{F}}^t Q_{\mathcal{G}}$, then the principle angles are given by*

$$\theta_k = \arccos(\sigma_k) \quad k = 1, \ldots, p. \tag{III.2}$$

The dependence of the minimum singular value $\sigma_{\min}(A)$ on the principle angles between the subspaces $\{S_i\}_{i=1}^M$ is given in the theorem below, which is one of the two main theorems of this section. The proofs are provided in [23].

Theorem 3. *Assume that $\{S_i\}_{i=1}^M$ are independent subspaces of $\mathbb{R}^D$ with corresponding dimensions $\{d_i\}_{i=1}^M$ such that $\sum_{i=1}^M d_i = D$. Let $\{\theta_j(S_i)\}_{j=1}^{\min(d_i, D-d_i)}$ be the principle angles between S_i and $\sum_{\ell\neq i} S_\ell$. If $A = \begin{bmatrix} a_1 & \ldots & a_D \end{bmatrix}$ is a matrix whose columns $\{a_1, \ldots, a_D\} \subset \cup_{i=1}^M S_i$ form a basis for $\mathbb{R}^D$, with $\|a_i\|_2 = 1$, $i = 1, \ldots, D$, then*

$$\sigma_{\min}^2(A) \leq \min_i \left(\prod_{j=1}^{\min(d_i, D-d_i)} \left(1 - \cos^2(\theta_j(S_i))\right) \right)^{1/D}, \tag{III.3}$$

where $\sigma_{\min}(A)$ is the smallest singular value of A.

Corollary 1. *Under the same conditions of Theorem 3, a simpler but possibly larger upper bound is given by:*

$$\sigma_{\min}^2(A) \leq \min_i (1 - \cos(\theta_1(S_i))^{1/D} 4^{1/D}, \tag{III.4}$$

where $\theta_1(S_i)$ is the minimum angle between S_i and $\sum_{\ell\neq i} S_\ell$.

Theorem 4. *Assume that $\{S_i\}_{i=1}^M$ are independent subspaces of $\mathbb{R}^D$ with corresponding dimensions $\{d_i\}_{i=1}^M$ such that $\sum_{i=1}^M d_i = D$. Let $\{\theta_j(S_i)\}_{j=1}^{\min(d_i, D-d_i)}$ be the principle angles between S_i and $\sum_{\ell\neq i} S_\ell$. Assume that $\mathbf{W} = [w_1 \cdots w_N] \in \mathbb{R}^{D\times N}$ is a matrix whose columns are drawn from $\bigcup_{i=1}^M S_i$ and the data is generic for each subspace S_i. If P is a permutation matrix such that $\mathbf{W}P = \begin{bmatrix} A_P & B_P \end{bmatrix}$, and A_P is invertible, then*

$$\sup_P \{\sigma_{\min}^2(A_P)\} \leq \min_i \left(\prod_{j=1}^{\min(d_i, D-d_i)} \left(1 - \cos^2(\theta_j(S_i))\right) \right)^{1/D} \tag{III.5}$$

In particular,

$$\sup_P \{\sigma_{\min}^2(A_P)\} \leq \min_i (1 - \cos(\theta_1(S_i))^{1/D} 4^{1/D}, \tag{III.6}$$

where $\theta_1(S_i)$ is the minimum angle between S_i and $\sum_{\ell\neq i} S_\ell$.

Remark 1. *The value $\sigma_{\min}(A_P)$ can be arbitrarily close to zero, thus, one of the goals is to find D columns of $\mathbf{W}$ that form a basis such that $\sigma_{\min}(A_P)$ is as close to the upper bound as possible without an exhaustive search.*

IV. Subspace Segmentation Algorithm for Noisy Data

Algorithm 1 works perfectly in noiseless data. For noisy data, the success of the algorithm depends on finding a good initial partition. Otherwise, the algorithm may terminate at a local minimum. Theorem 2 works perfectly for *noiseless* data (it determines a basis for each subspace and it correctly clusters all of the data points). An algorithm for implementing Theorem 2 is given in [23]. However, it does not perform very well when sufficiently large noise is present because any threshold value will keep some of the values that need to be zeroed out and will zero out some of the values that need to be kept. However, the thresholded reduced echelon form can be used to determine a set of clusters that can in turn be used to determine a good initial set of subspaces in Algorithm 1.

For example, if the number of subspaces is known and the subspaces have equal and known dimensions (assume that there are M subspaces and each subspace has dimension d), then Algorithm 2 below combines Algorithm 1 and Theorem 2 as follows: First, the reduced row echelon form rref($\mathbf{W}$) of $\mathbf{W}$ is computed. Since the data is noisy, the non-pivot columns of rref($\mathbf{W}$) will most likely have all non-zero entries. The error in those entries will depend on the noise and the positions of the subspaces as in (3). Since each subspace is d-dimensional, the highest d entries of each non-pivot column is set to 1 and all other entries are set to 0. This determines the binary reduced row echelon form Brref($\mathbf{W}$) of $\mathbf{W}$ (note that, according to Theorem 2, each non-pivot column of Brref($\mathbf{W}$) is supposed to have d entries). M groups of the equivalent columns of Brref($\mathbf{W}$) are determined and used as the initial partition for Algorithm 1. This process is described in Algorithm 2. Note that a dimensionality reduction is also performed to speed up the process.

Remark 2. *In Step-5 of Algroithm 2, Brref($\mathbf{W}$) is computed by setting the highest d entries of each non-pivot columns to 1 and the others to 0. If we do not know the dimensions of the subspaces, we may need to determine a threshold from the noise characteristics and a priori knowledge of the relative position of subspaces using* (III.1) *and* (III.3).

Remark 3. *In Step-7 of Algorithm 2, we find the subspace $S_i^o(P)$ that minimizes the expression $e(\mathbf{W}_i, S) = \sum_{w\in\mathbf{W}_i} d^p(w, S)$ for each subset $\mathbf{W}_i$ in the partition P. For data with light-tailed noise (e.g. Gausian distributed noise) $p = 2$ is optimal and the minimum in Step-7 can be found using SVD. For heavy-tailed noise (e.g. Laplacian distributed noise), $p = 1$ is the better choice as described in the simulations section.*

Remark 4. *In order to reduce the dimensionality of the problem, we compute the SVD of $\mathbf{W} = U\Sigma V^t$. In Algorithm 2, each subspace is d-dimensional and there are M subspaces. Therefore, it replaces $\mathbf{W}$ by $(V^t)_r$, where $r = M \times d$ is known or estimated rank of $\mathbf{W}$.*

Algorithm 2 Combined Algorithm - Optimal Solution $\mathbf{S}^o$

Require: Normalized data matrix $\mathbf{W}$.

1: Set $r = M \times d$.
2: Compute the SVD of $\mathbf{W}$ and find $(V^t)_r$.
3: Replace the data matrix $\mathbf{W}$ with $(V^t)_r$.
4: Compute rref($\mathbf{W}$)
5: Compute Brref($\mathbf{W}$) by setting the highest d entries of each non-pivot column to 1 and all the others to 0.
6: Group the non-pivot equivalent columns of Brref($\mathbf{W}$) into M largest clusters $\{\mathbf{W}_1, \ldots, \mathbf{W}_M\}$ and set the initial partition $P = \{\mathbf{W}_1, \ldots, \mathbf{W}_M\}$.
7: For each subset $\mathbf{W}_i$ in the partition P find the subspace $S_i^o(P)$ that minimizes the expression $e(\mathbf{W}_i, S) = \sum_{w \in \mathbf{W}_i} d^p(w, S)$.
8: **while** $\sum_{i=1}^{M} e(\mathbf{W}_i, S_i^o(P)) > e(\mathbf{W}, \mathbf{S}^o(P))$ **do**
9: **for all** i from 1 to M **do**
10: Update $\mathbf{W}_i = \{w \in \mathbf{W} : d(w, S_i^o(P)) \leq d(w, S_k^o(P)),\ k = 1, \ldots, M\}$
11: Update $S_i^o(P) = \underset{S}{\operatorname{argmin}}\, e(\mathbf{W}_i, S)$
12: **end for**
13: Update $P = \{\mathbf{W}_1, \ldots, \mathbf{W}_M\}$
14: **end while**
15: $\mathbf{S}^o = \{S_1^o(P), \ldots, S_M^o(P)\}$

V. Experimental Results

We used the Hopkins 155 Dataset [6] to evaluate our algorithm. The RREF-based algorithm is extremely fast and works well with two-motion video sequences. The average and median errors for all two-motion sequences are 11.45% and 6.78%, respectively (8.81% and 5.44% for checker, 16.04% and 11.94% for traffic, and 17.25% and 12.69% for articulated motion). However, the error is very high for three-motion sequences and obviously it does not work well with such video sequences. We believe that this is due to the fact that the noise is correlated, and the minimum of Problem 1 does not give the correct clustering for this case. The best clustering method to date for clustering in this case is based on similarity between trajectory vectors computed from local subspace estimations [26].

VI. Conclusion

This paper introduces a simple and very fast approach for subspace segmentation for data drawn from a union of subspaces. In absence of noise, our approach can find the number of subspaces, their dimensions, and an orthonormal basis for each subspace. We provide an analysis of our theory and determine its limitations and strengths in presence of outliers and noise.

Acknowledgment

The research of Akram Aldroubi is supported in part by NSF Grant DMS-1108631. The research of Ali Sekmen is supported in part by NASA Grant NNX12AI14A.

References

[1] Y. C. Eldar and M. Mishali, "Robust recovery of signals from a structured union of subspaces," *IEEE Transactions on Information Theory*, vol. 55, no. 11, pp. 5302–5316, 11 2009.
[2] E. Elhamifar and R. Vidal, "Sparse subspace clustering," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 2790–2797.
[3] R. Vidal and E. Elhamifar, "Clustering disjoint subspaces via sparse representation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2010, pp. 1926–1929.
[4] G. Liu, Z. Lin, and Y. Yu, "Robust subspace segmentation by low-rank representation," in *International Conference on Machine Learning*, 2010, pp. 663–670.
[5] R. Vidal, Y. Ma, and S. Sastry, "Generalized principal component analysis (GPCA)," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 12, pp. 1945–1959, 12 2005.
[6] R. Tron and R. Vidal, "A benchmark for the comparison of 3-d motion segmentation algorithms," in *Computer Vision and Pattern Recognition*, 2007, pp. 1–8.
[7] K. Kanatani and Y. Sugaya, "Multi-stage optimization for multi-body motion segmentation," in *IEICE Trans. Inf. and Syst*, 2003, pp. 335–349.
[8] A. Aldroubi and K. Zaringhalam, "Nonlinear least squares in $\mathbb{R}^n$," *Acta Applicandae Mathematicae*, vol. 107, no. 1-3, pp. 325–337, 7 2009.
[9] A. Aldroubi, C. Cabrelli, and U. Molter, "Optimal non-linear models for sparsity and sampling," *Journal of Fourier Analysis and Applications*, vol. 14, no. 5, pp. 793–812, 12 2009.
[10] P. Tseng, "Nearest q-flat to m points," *Journal of Optimization Theory and Applications*, vol. 105, no. 1, pp. 249–252, 2000.
[11] M. Fischler and R. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 6 1981.
[12] N. Silva and J. Costeira, "Subspace segmentation with outliers: a grassmannian approach to the maximum consensus subspace," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2008.
[13] U. V. Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, pp. 395–416, 2007.
[14] G. Chen and G. Lerman, "Spectral curvature clustering (SCC)," *International Journal of Computer Vision*, vol. 81, pp. 317–330, 2009.
[15] F. Lauer and C. Schnorr, "Spectral clustering of linear subspaces for motion segmentation," in *IEEE International Conference on Computer Vision*, 2009.
[16] J. Yan and M. Pollefeys, "A general framework for motion segmentation: Independent, articulated, rigid, non-rigid, degenerate and nondegenerate," in *9th European Conference on Computer Vision*, 2006, pp. 94–106.
[17] A. Goh and R. Vidal, "Segmenting motions of different types by unsupervised manifold clustering," in *Computer Vision and Pattern Recognition*, 6 2007, pp. 1–6.
[18] R. Vidal, "A tutorial on subspace clustering," *IEEE Signal Processing Magazine*, vol. 79, no. 1, pp. 52–68, 2010.
[19] G. Chen and G. Lerman, "Foundations of a multi-way spectral clustering framework for hybrid linear modeling," vol. 9, no. 5, pp. 517–558, 2009.
[20] C. Gear, "Multibody grouping from motion images," *International Journal of Computer Vision*, vol. 29, no. 2, pp. 133–150, 1998.
[21] G. Lerman and T. Zhang, "Robust recovery of multiple subspaces by geometric lp minimization," vol. 39, no. 5, pp. 2686–2715, 2011.
[22] A. Aldroubi and R. Tessera, "On the existence of optimal unions of subspaces for data modeling and clustering," *Foundation of Computational Mathematics*, vol. 11, no. 3, pp. 363–379, 6 2011.
[23] A. Aldroubi and A. Sekmen, "Reduced row echelon form and non-linear approximation for subspace segmentation and high-dimensional data clustering," *Preprint*.
[24] R. Latala, "Some estimates of norms of random matrices," *Proc. Amer. Math. Soc.*, vol. 133, 2005.
[25] G. H. Golub and C. F. V. Loan, *Matrix computations, 3rd Edition*. Johns Hopkins University Press, 1996.
[26] A. Aldroubi and A. Sekmen, "Nearness to local subspace algorithm for subspace and motion segmentation," *arXiv.org*, 2010. [Online]. Available: http://arxiv.org/abs/1010.2198

Missing Entries Matrix Approximation and Completion

Gil Shabat
School of Electrical Engineering
Tel Aviv University
gil@eng.tau.ac.il

Yaniv Shmueli
School of Computer Science
Tel Aviv University
yaniv.shmueli@cs.tau.ac.il

Amir Averbuch
School of Computer Science
Tel Aviv University
amir@math.tau.ac.il

***Abstract*—We describe several algorithms for matrix completion and matrix approximation when only some of its entries are known. The approximation constraint can be any whose approximated solution is known for the full matrix. For low rank approximations, similar algorithms appear recently in the literature under different names. In this work, we introduce new theorems for matrix approximation and show that these algorithms can be extended to handle different constraints such as nuclear norm, spectral norm, orthogonality constraints and more that are different than low rank approximations. As the algorithms can be viewed from an optimization point of view, we discuss their convergence to global solution for the convex case. We also discuss the optimal step size and show that it is fixed in each iteration. In addition, the derived matrix completion flow is robust and does not require any parameters. This matrix completion flow is applicable to different spectral minimizations and can be applied to physics, mathematics and electrical engineering problems such as data reconstruction of images and data coming from PDEs such as Helmholtz's equation used for electromagnetic waves.**

I. Introduction

Matrix completion and matrix approximation are important problems in a variety of fields such as statistics [1], biology [2], statistical machine learning [3], signal processing and computer vision/image processing [4]. Rank reduction by matrix approximation is important, for example, in compression where low rank indicates the existence of redundant information and matrix completion is important in collaborative filtering, such as the Netflix problem and different reconstruction problems. Usually, the matrix completion problem, is defined as finding a matrix, with smallest possible rank, that satisfy the existence of certain entries.

$$\begin{aligned} &\text{minimize rank } (\mathbf{X}) \\ &\text{subject to } X_{i,j} = M_{i,j}, \quad (i,j) \in \Omega. \end{aligned} \tag{I.1}$$

Since Eq. I.1 is an NP-hard problem, some relaxations methods have been proposed. The most popular relaxation is one that replaces the rank by the nuclear norm:

$$\begin{aligned} &\text{minimize } \|\mathbf{X}\|_* \\ &\text{subject to } X_{i,j} = M_{i,j}, \quad (i,j) \in \Omega, \end{aligned} \tag{I.2}$$

where $\|\mathbf{X}\|_*$ denotes the nuclear norm of $\mathbf{X}$ that is equal to the sum of the singular values of $\mathbf{X}$. A small value of $\|\mathbf{X}\|_*$ is related to the property of having a low rank [5]. An iterative solution, which is based on a singular value thresholding, is given in [6]. A completion algorithm, based on the local information of the matrix, is proposed in [7]. In this work, a more robust and simple approach for solving a variety of matrix approximation of certain entries by approximating the full matrix is discussed. We approximate problems of the form

$$\begin{aligned} &\text{minimize } \|\mathcal{P}_\Omega \mathbf{X} - \mathcal{P}_\Omega \mathbf{M}\|_F \\ &\text{subject to } f(\mathbf{X}) \leq 0, \end{aligned} \tag{I.3}$$

given that the solution for

$$\begin{aligned} &\text{minimize } \|\mathbf{X} - \mathbf{M}\|_F \\ &\text{subject to } f(\mathbf{X}) \leq 0 \end{aligned} \tag{I.4}$$

is known. Here, $\{\mathcal{P}_\Omega \mathbf{X}\}_{i,j} = X_{i,j}$ if $(i,j) \in \Omega$ and 0 otherwise. If $f(\mathbf{X})$ is convex and satisfies some condition (which is explained in the next sections), the algorithm finds the global solution. Nevertheless, convergence is guaranteed, but to a local solution. Then, we show how this algorithm can be used for solving a variety of matrix completion problems as well, such as spectral norm completion:

$$\begin{aligned} &\text{minimize } \|\mathbf{X}\|_2 \\ &\text{subject to } X_{i,j} = M_{i,j}, \quad (i,j) \in \Omega, \end{aligned} \tag{I.5}$$

Ky-Fan norm completion:

$$\begin{aligned} &\text{minimize } \|\mathbf{X}\|_{(k)} \\ &\text{subject to } X_{i,j} = M_{i,j}, \quad (i,j) \in \Omega, \end{aligned} \tag{I.6}$$

where $\|\mathbf{X}\|_{(k)} = \sum_{i=1}^{k} \sigma_i$ (sum of largest k singular values). Note that the spectral norm and the nuclear norm are a special case of the Ky-Fan norm. We also discuss approximation problems such as:

$$\begin{aligned} &\text{minimize } \|\mathcal{P}_\Omega \mathbf{X} - \mathcal{P}_\Omega \mathbf{M}\|_F \\ &\text{subject to } \mathbf{X}^{\mathbf{T}}\mathbf{X} = \mathbf{I}. \end{aligned} \tag{I.7}$$

II. Theorems on full matrix approximation

The algorithm that approximates a matrix at certain points requires from us to be able to approximate the matrix when taking into account all its entries. Therefore, we review some theorems on full matrix approximation theorems in addition to the well known Eckart-Young theorem mentioned in the introduction. The low rank approximation problem can be modified to approximate a matrix under the Frobenius norm

while having the Frobenius norm as a constraint as well instead of having low rank. Formally,

$$\begin{aligned} &\text{minimize } \|\mathbf{X}-\mathbf{M}\|_F \\ &\text{subject to } \|\mathbf{X}\|_F \leq \lambda. \end{aligned} \tag{II.1}$$

A solution for Eq. II.1 is given by $\mathbf{X} = \frac{\mathbf{M}}{\|\mathbf{M}\|_\mathbf{F}} \min(\|\mathbf{M}\|_\mathbf{F}, \lambda)$.

Proof: The expression $\|\mathbf{X}\|_F^2 \leq \lambda^2$ can be thought of as an $m \times n$ dimensional ball with radius λ centered at the origin. $\mathbf{M}$ is an $m \times n$ dimensional point. We are looking for a point $\mathbf{X}$ on the ball $\|\mathbf{X}\|_F^2 = \lambda^2$ that has a minimal Euclidean distance (Frobenius norm) from $\mathbf{M}$. If $\|\mathbf{M}\|_F \leq \lambda$ then $\mathbf{X} = \mathbf{M}$ and it is inside the ball having a distance of zero. If $\|\mathbf{M}\|_F > \lambda$, then the shortest distance is given by the line going from the origin to $\mathbf{M}$ whose intersection with the sphere $\|\mathbf{X}\|_F^2 \leq \lambda^2$ is the closest point to $\mathbf{M}$. This point is given by $\mathbf{X} = \frac{\mathbf{M}}{\|\mathbf{M}\|_\mathbf{F}}\lambda$. ■

An alternative approach uses the Lagrange multiplier in a brute-force manner. This leads to a non-linear system of equations, which are difficult to solve. Note that this problem can be easily extended to the general case

$$\begin{aligned} &\text{minimize } \|\mathcal{P}\mathbf{X}-\mathcal{P}\mathbf{M}\|_F \\ &\text{subject to } \|\mathbf{X}\|_F \leq \lambda. \end{aligned} \tag{II.2}$$

Proof: The proof is similar to the previous one but here we are looking for a point $\mathbf{X}$ on the sphere that is the closest to a line whose points $\mathbf{X}' \in \mathcal{H}$ satisfy $\mathcal{P}\mathbf{X}' = \mathcal{P}\mathbf{M}$. By geometrical considerations, this point is given by $\mathbf{X} = \frac{\mathcal{P}\mathbf{M}}{\|\mathcal{P}\mathbf{M}\|_F}\lambda$. ■

Hence, we showed a closed form solution for the problem in Eq. II.2.

Another example is the solution to the problem:

$$\begin{aligned} &\text{minimize } \|\mathbf{X}-\mathbf{M}\|_F \\ &\text{subject to } \mathbf{X}^T\mathbf{X} = \mathbf{I}. \end{aligned} \tag{II.3}$$

This is known as the orthogonal Procrustes problem ([8]) and the solution is given by $\mathbf{X} = \mathbf{U}\mathbf{V}^*$, where the SVD of $\mathbf{M}$ is given by $\mathbf{M} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$. The solution can be extended to a matrix $\mathbf{X}$ satisfying $\mathbf{X^T X} = \mathbf{D^2}$, where $\mathbf{D}$ is a known or unknown diagonal matrix. When $\mathbf{D}$ is unknown, the solution is the best possible orthogonal matrix. When $\mathbf{D}$ is known, the problem can be converted to become the orthonormal case (Eq. II.3) by substituting $\mathbf{X} = \mathbf{V}\mathbf{D}$ where $\mathbf{V^T V} = \mathbf{I}$. When $\mathbf{D}$ is unknown, the problem can be solved by applying an iterative algorithm that is described in [9].

We now examine the following problem:

$$\begin{aligned} &\text{minimize } \|\mathbf{X}-\mathbf{M}\|_F \\ &\text{subject to } \|\mathbf{X}\|_2 \leq \lambda. \end{aligned} \tag{II.4}$$

A solution to this problem uses the Pinching theorem ([10]):

Lemma II.1 (Pinching theorem)**.** *For every matrix* $\mathbf{A}$ *and a unitary matrix* $\mathbf{U}$ *and for any norm satisfying* $\|\mathbf{U}\mathbf{A}\mathbf{U}^*\| = \|\mathbf{A}\|$ *then* $\|diag(\mathbf{X})\| \leq \|\mathbf{X}\|$.

A proof is given in [12]. An alternative proof is given in [14].

Lemma II.2 (Minimization of the Frobenius norm under the spectral norm constraint)**.** *Assume the SVD of* $\mathbf{M}$ *is given by* $\mathbf{M} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ *where* $\mathbf{\Sigma} = diag(\sigma_1, .., \sigma_n)$*. Then, the matrix* $\mathbf{X}$*, which minimizes* $\|\mathbf{X}-\mathbf{M}\|_F$ *such that* $\|\mathbf{X}\|_2 \leq \lambda$*, is given by* $\mathbf{X} = \mathbf{U}\tilde{\mathbf{\Sigma}}\mathbf{V}^*$ *where* $\tilde{\sigma}_i$ *are the singular values of* $\tilde{\Sigma}$ *and* $\tilde{\sigma}_i = \min(\sigma_i, \lambda), i = 1, \dots k,\ k \leq n$.

Proof: $\|\mathbf{X}-\mathbf{M}\|_F = \|\mathbf{X}-\mathbf{U}\mathbf{\Sigma}\mathbf{V}^*\|_F = \|\mathbf{U}^*\mathbf{X}\mathbf{V}-\mathbf{\Sigma}\|_F$. Since $\mathbf{\Sigma}$ is diagonal, $\|\text{diag}(\mathbf{U}^*\mathbf{X}\mathbf{V})-\mathbf{\Sigma}\|_F \leq \|\mathbf{U}^*\mathbf{X}\mathbf{V}-\mathbf{\Sigma}\|_F$. From Lemma II.1 we know that $\|\text{diag}(\mathbf{U}^*\mathbf{X}\mathbf{V})\|_2 \leq \|\mathbf{U}^*\mathbf{X}\mathbf{V}\|_2$. Therefore, $\mathbf{U}^*\mathbf{X}\mathbf{V}$ has to be diagonal and the best minimizer under the spectral norm constraint is achieved by minimizing each element separately yielding $\mathbf{U}^*\mathbf{X}\mathbf{V} = \text{diag}(\min(\sigma_i, \lambda)),\ i = 1, \dots k, k \leq n$. Hence, $\mathbf{X} = \mathbf{U}\tilde{\mathbf{\Sigma}}\mathbf{V}^*$. ■

The same argument that states that $\mathbf{U}^*\mathbf{X}\mathbf{V}$ has to be diagonal, can also be applied when the constraint is given by the nuclear norm. Define $\tilde{\mathbf{\Sigma}} = \mathbf{U}^*\mathbf{X}\mathbf{V}$. We wish to minimize $\|\tilde{\mathbf{\Sigma}}-\mathbf{\Sigma}\|_F = \sum_i (\tilde{\sigma}_i - \sigma_i)^2$ s.t. $\|\mathbf{X}\|_* = \|\tilde{\mathbf{\Sigma}}\|_* = \sum_i |\tilde{\sigma_i}| \leq \lambda, i = 1, \dots k, k \leq n$. Note that $\tilde{\sigma}_i$ has to be nonnegative otherwise it will increase the Frobenius norm but will not change the nuclear norm. Hence, the problem can now be formulated as:

$$\begin{aligned} &\text{minimize } \textstyle\sum_i (\tilde{\sigma}_i - \sigma_i)^2 \\ &\text{subject to } \textstyle\sum_i \tilde{\sigma}_i \leq \lambda, \\ &\qquad \tilde{\sigma}_i \geq 0. \end{aligned} \tag{II.5}$$

This is a standard convex optimization problem that can be solved by methods such as semidefinite programming [11]. The exact same can be done to the Ky-Fan norm.

III. Approximation of certain entries

Suppose we wish to approximate only certain entries of the matrix, under different constraints, i.e. we are interested in solving Eq. I.3, given that the solution of Eq. I.4 is known and given by $\mathcal{D}\mathbf{M}$, where $\mathcal{D}$ is the solution operator. For example, if the constraint is $\text{rank}(\mathbf{X}) \leq k$ $\mathcal{D}\mathbf{X}$ is the truncated SVD of $\mathbf{X}$ containing the first k singular values. Note that $\mathcal{D}$ is not necessarily convex. We examine the following iterative algorithm:

$$\mathbf{X}_{n+1} = \mathcal{D}(\mathbf{X}_n - \mathcal{P}(\mathbf{X}_n - \mathbf{M})). \tag{III.1}$$

Eq. III.1 can be considered as a projected gradient algorithm with unit step size, where the projection is given by $\mathbf{D}$.

Theorem III.1 (Local Convergence)**.** *: Let* $\epsilon(\mathbf{X}_n) = \|\mathcal{P}\mathbf{X}_n - \mathcal{P}\mathbf{M}\|_F$ *be the error at the* n*th iteration, then* $\epsilon(\mathbf{X}_n)$ *is monotonically decreasing, and because it is bounded the algorithm converges.*

The proof for Theorem III.1 is given in [14]. Theorem III.1 does not say anything about convergence to the global solution. However, when the projection $\mathcal{D}$ is convex and self adjoint ($\mathcal{D} = \mathcal{D}^*$) and the algorithm is modified to have adaptive step size, that is:

$$\mathbf{X}_{n+1} = \mathcal{D}(\mathbf{X}_n - \mu_n\mathcal{P}(\mathbf{X}_n - \mathbf{M})), \tag{III.2}$$

and $\mu_n = \tilde{\mu}2^{-l[n]}$ is computed by Armijo rule in a greedy form, minimizing the error in every iteration:

$$\begin{aligned} l[n] &= \min\{j \in \mathbb{Z}_{\geq 0} : f(\mathbf{X}_{n,j}) \\ &\qquad \leq f(\mathbf{X}_n) - \sigma \text{trace}(\nabla f(\mathbf{X}_n)^T(\mathbf{X}_n - \mathbf{Z}_{n,j}))\}, \\ \text{and } \mathbf{Z}_{n,j} &= \mathcal{D}(\mathbf{X}_n - \tilde{\mu}2^{-j}\nabla f(\mathbf{X}^n)), \end{aligned} \tag{III.3}$$

where $f(X) = \frac{1}{2}\|\mathcal{P}\mathbf{X} - \mathcal{P}\mathbf{M}\|_F^2$, $\tilde{\mu} > 0$ and $\sigma \in (0, 1)$, Then the algorithm is guarantee to achieve the global solution [13]. This approach has two major problems:

- For the cases of interest, the operators for truncating the nuclear and spectral norm, are not self-adjoint ($\mathcal{D} \neq \mathcal{D}^*$)
- This approach requires applying the Armijo rule in every iteration. This means several applications of the operator $\mathcal{D}$ in each iteration which is usually computationally expensive.

As for the first point, requiring the projection $\mathcal{D}$ to be self-adjoint can be slightly more than needed for the global convergence proof in [13]. This requirement is needed in order to satisfy $\langle X - Y, \mathcal{D}X - X\rangle \geq 0$ for $Y = \mathcal{D}Y$, which always holds when $\mathcal{D} = \mathcal{D}^*$, but also when $\mathcal{D}$ is as we defined in Lemma II.2 and Eq. II.5.

Theorem III.2. *Let $\mathcal{D}$ be the following projection (defined as in Lemma II.2): Given the SVD of X is $\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^*$, we define $\mathcal{D}_\lambda X = U\tilde{S}V^*$ where $\tilde{s}_i = min(s_i, \lambda)$. Then, for every matrices $\mathbf{X}$ and $\mathbf{Y}$ such that $Y = \mathcal{D}\mathbf{Y}$, $\langle \mathbf{X} - \mathbf{Y}, \mathcal{D}\mathbf{X} - \mathbf{X}\rangle \geq 0$*

Proof: The condition $\langle \mathbf{X} - \mathbf{Y}, \mathcal{D}\mathbf{X} - \mathbf{X}\rangle \geq 0$ can be reformulated as

$$\langle \mathbf{X}, \mathbf{X} - \mathcal{D}\mathbf{X}\rangle \geq \langle \mathbf{Y}, \mathbf{X} - \mathcal{D}\mathbf{X}\rangle, \tag{III.4}$$

where $\|Y\|_2 \leq \lambda$.

First, note that the value of the right hand side is maximal when $\mathbf{Y}$ and $\mathbf{X} - \mathcal{D}\mathbf{X}$ have the same angle (Cauchy-Schwartz inequality). Hence, we define: $\mathbf{X} = \mathbf{U}\mathbf{S}_X\mathbf{V}^*$, $\mathbf{Y} = \mathbf{U}\tilde{\mathbf{S}}_Y\mathbf{V}^*$ and $\mathcal{D}\mathbf{X} = \mathbf{U}\tilde{\mathbf{S}}_X\mathbf{V}^*$. The tilde is for indicating that the singular values of $\tilde{\mathbf{S}}$ are smaller or equal to λ.

We start by evaluating the left side of Eq.III.4:

$$\langle \mathbf{X}, \mathbf{X} - \mathcal{D}\mathbf{X}\rangle = \text{trace}[\mathbf{S}_X(\mathbf{S}_X - \tilde{\mathbf{S}_X})] = \sum_i s_{x_i}(s_{x_i} - \tilde{s}_{x_i}). \tag{III.5}$$

Now, for $s_{x_i} \leq \lambda$ we get $(s_{x_i} - \tilde{s}_{x_i}) = 0$. Hence, only when $s_{x_i} > \lambda$ the sum grows and the expression can be rewritten as: $\langle \mathbf{X}, \mathbf{X} - \mathcal{D}\mathbf{X}\rangle = \sum_{s_{x_i} > \lambda} s_{x_i}(s_{x_i} - \tilde{s}_{x_i})$

We now observe the right side of Eq. III.4:

$$\langle \mathbf{Y}, \mathbf{X} - \mathcal{D}\mathbf{X}\rangle = \text{trace}[\tilde{\mathbf{S}}_Y(\mathbf{S}_X - \tilde{\mathbf{S}}_X)] = \sum_i \tilde{s}_{y_i}(s_{x_i} - \tilde{s}_{x_i}). \tag{III.6}$$

Again, the elements that contribute to the sum are those for which $s_{x_i} > \lambda$. Hence, on the right side we obtained: $\langle \mathbf{Y}, \mathbf{X} - \mathcal{D}\mathbf{X}\rangle = \sum_{s_{x_i} > \lambda} \tilde{s}_{y_i}(s_{x_i} - \tilde{s}_{x_i})$.

Both expressions can be thought of as a sum of the positive elements $(s_{x_i} - \tilde{s}_{x_i})$ with different coefficients. Both series have the same length ($s_{x_i} > \lambda$) but the coefficient on the left side is s_{x_i} for i's that give $s_{x_i} > \lambda$ and the right hand series coefficients are by definition (since $\|\mathbf{Y}\|_2 \leq \lambda$) smaller than λ. Therefore, the sum of the left side is bigger than the sum of the right side. This completes the proof. ■

This means that for the spectral norm, the algorithm converges to the global solution. The exact same proof can be done for the nuclear norm and Ky-Fan norm as well, showing the algorithm converges to global solution.

Theorem III.3 (Optimal step size)**.** *For the matrix approximation problem (Eq. I.3) with convex $\mathcal{D}$, the optimal step size is given by $\mu_n = 1$.*

The proof of Theorem III.3 is given in [14]. Note that this holds for any case of projected gradient involving orthogonal axes. Theorem III.3 states that in our case, when having a convex constraint and projection, then Eq. III.1 converges to the global solution. This means, that now we can solve a variety of matrix approximation problem with reasonable computation rate. Note, that we have shown that in some cases, global solution is achieved even when the projection is not self-adjoint (orthogonal). The next section shows, how this very simple algorithm, can be applied to matrix completion problems as well.

IV. Matrix Completion

Matrix completion is an important problem that has been investigated extensively. The matrix completion problem differs from the matrix approximation problem by the fact that the known entries must remain fixed while changing their role from the objective function to be minimized to the constraint part. A well investigated matrix completion problem appears in the introduction as the rank minimization problem. Because rank minimization is not convex and NP-hard, it is usually relaxed for the nuclear norm minimization. Since for the convex case, we have seen that Eq. III.1 converges to the global solution, matrix completion can be achieved simply by using binary search. The advantage of this approach over other different approaches, which minimize the nuclear norm for example, is that it is general and can be applied to other problems that were not addressed such as minimizing the spectral norm. Moreover, some algorithms such as the Singular Value Thresholding (SVT) [6] require additional parameters τ and δ that affect the convergence and the final result, where in this approach no external parameters are required (except for tolerance for determining convergence).

This approach is detailed in Algorithm IV.1, which is robust and does not require any tuning, other than tolerance threshold for determining convergence. Algorithm IV.1 can be used for a matrix completion under a variety of constraints.

Fig. IV shows Algorithm IV.1 results over a corrupted image. In the corrupted image, squares of size 3×3 were randomly removed from the image, destroying 18% of it. The reconstruction is more difficult, since the damage is in squares and not just irregular points. The original image nuclear norm is $51,625$, the corrupted nuclear norm is $96,500$ and the norm of the completed matrix is $50,418$. Minimizing nuclear norm for image reconstructing is a well known method, as images

Algorithm IV.1: Matrix Completion using Nuclear Norm / Spectral Norm Minimization

Input: $\mathbf{M}$ - matrix to complete, $\mathcal{P}$ - projection operator that specifies the important entries, tol - admissible approximation error, λ_{tol} - admissible constraint accuracy
Output: $\mathbf{X}$ - Completed matrix

$\mathbf{M} \leftarrow \mathcal{P}\mathbf{M}$
$\lambda_{min} \leftarrow 0$
$\lambda_{max} \leftarrow \|\mathbf{M}\|_*$ (or $\|\mathbf{M}\|_2$ for the spectral norm)
$\lambda \leftarrow 0$
repeat
  $\lambda_{prev} \leftarrow \lambda$
  $\lambda \leftarrow (\lambda_{min} + \lambda_{max})/2$
  $\mathbf{X} \leftarrow$ Approximate $\mathcal{P}\mathbf{M}$ s.t. $\|\mathbf{X}\|_* \leq \lambda$ (or $\|\mathbf{X}\|_2 \leq \lambda$ for the spectral norm case)
  $error \leftarrow \|\mathcal{P}\mathbf{X} - \mathcal{P}\mathbf{M}\|_F$
  if $error > tol$ **then**
    $\lambda_{min} \leftarrow \lambda$
  else
    $\lambda_{max} \leftarrow \lambda$
  end if
until $error < tol$ and $|\lambda - \lambda_{prev}| < \lambda_{tol}$
return $\mathbf{X}$

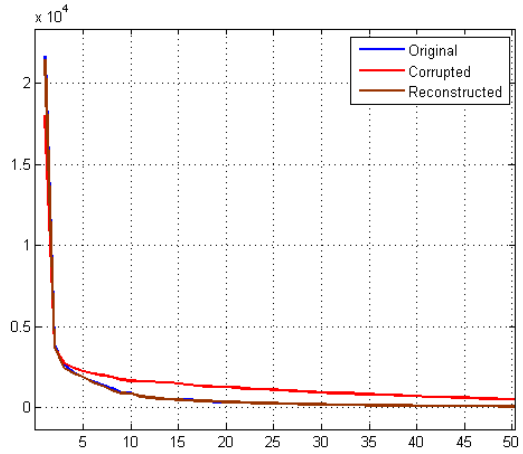


Fig. IV.1. Singular values comparison between the different images.

usually have a low numerical rank as the singular values decay very fast. It can be seen in Fig. IV that the singular values of the reconstructed image, are almost identical to the original.

ACKNOWLEDGMENT

This research was partially supported by the Israel Science Foundation (Grant No. 1041/10) and by the Israeli Ministry of Science & Technology 3-9096.

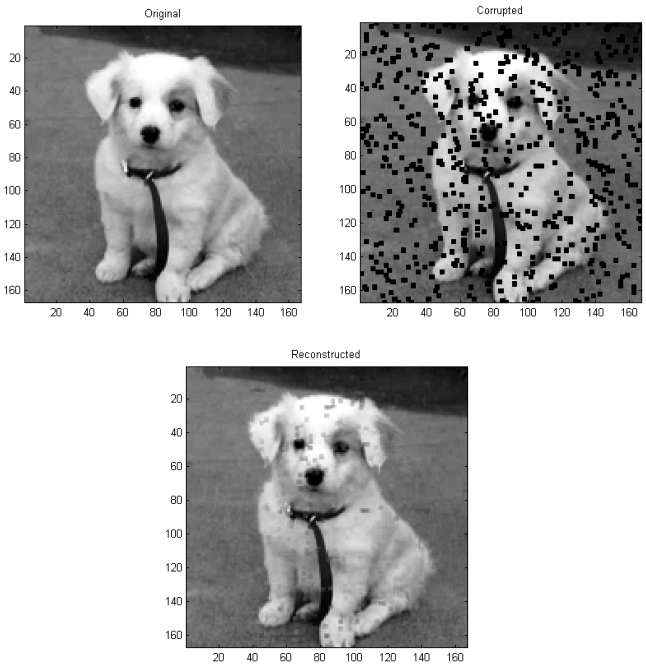


Fig. IV.2. Corrupted dog image and the reconstructed image.

REFERENCES

[1] T.A. Louis, *Finding the observed information matrix when using the EM algorithm*, Journal of the Royal Statistical Society, Series B. (Methodological), Vol. 44, No. 2, pp. 226-233, 1982.
[2] T. Hastie, R. Tibshirani, G. Sherlock, M. Eisen, P. Brown, and D. Botstein, *Imputing missing data for gene expression arrays*. Technical report; Division of Biostatistics, Stanford University, 1999.
[3] N. Srebro and T. Jaakkola, *Weighted low-rank approximations*, Preceeding of the 20th International Conference on Machine Learning (ICML-2003), Washington DC, 2003.
[4] J. Mairal, M. Elad, G. Sapiro, *Sparse representation for color image restoration*, IEEE Transactions on Image Processing, Vol. 17, No. 1, pp.53-69, 2008.
[5] M. Fazel, *Matrix Rank Minimization with Applications. PhD thesis*, Stanford University, 2002.
[6] J.F. Cai, E.J. Candes and Z. Shen, *Singular Value Thresholding Algorithm for Matrix Completion*, SIAM Journal on Optimization, Vol. 20, No. 4, pp. 1956-1982, 2010.
[7] Feng Nan, *Low Rank Matrix Completion*, Master thesis, Massachusetts Institute of Technology, 2009.
[8] P. H. Schonemann, *A generalized solution of the orthogonal procrustes problem*, Psychometrika, Vol. 31, No. 1, pp. 1-10, 1966.
[9] R. Everson, *Orthogonal but not orthonormal Procrustes problem*, 1997.
[10] R. Bhatia, *Matrix Analysis*, Graduate Texts in Mathematics, Springer 1996.
[11] S. Boyd, L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.
[12] , I.C. Gohberg, M.G. Krein, *Introduction to the theory of linear and selfadjoint operators*, Translations of Mathematical Monographs, Vol. 18, pp. 94-95, 1969.
[13] A.N. Iusem, *On the convergence properties of the projected gradient method for convex optimization*, Computational and Applied Mathematics, Vol. 22, No. 1, pp. 37-52, 2003.
[14] G. Shabat, A. Averbuch, *Interest Zone Matrix Approximation*, Electronic Journal of Linear Algebra, Vol. 23, pp. 678-702, 2012.

Using Affinity Perturbations to Detect Web Traffic Anomalies

Yaniv Shmueli
School of
Computer Science
Tel Aviv University
yaniv.shmueli@cs.tau.ac.il

Tuomo Sipola
Department of
Mathematical Information Technology
University of Jyväskylä
tuomo.sipola@jyu.fi

Gil Shabat
School of
Electrical Engineering
Tel Aviv University
gil@eng.tau.ac.il

Amir Averbuch
School of
Computer Science
Tel Aviv University
amir@math.tau.ac.il

***Abstract*—The initial training phase of machine learning algorithms is usually computationally expensive as it involves the processing of huge matrices. Evolving datasets are challenging from this point of view because changing behavior requires updating the training. We propose a method for updating the training profile efficiently and a sliding window algorithm for online processing of the data in smaller fractions. This assumes the data is modeled by a kernel method that includes spectral decomposition. We demonstrate the algorithm with a web server request log where an actual intrusion attack is known to happen. Updating the kernel dynamically using a sliding window technique, prevents the problem of single initial training and can process evolving datasets more efficiently.**

***Index Terms*—perturbation theory, eigenvalue problem, diffusion maps, dimensionality reduction, anomaly detection, web traffic**

I. Introduction

Evolving data that requires frequent updates to the training is a challenging target when extracting constructive information. The computational complexity of the training phase increases with such datasets because an earlier profile may not accurately represent the behavior of current data. Therefore, the extracted profile has to be updated frequently. A straightforward approach for updating the training profile is to repeat the entire computational process that generated the original profile. This paper summarizes a method for efficiently updating the evolving profile.

A common practice in kernel methods is to extract features from a high dimensional dataset, and to form a similarity graph between the features in the dataset. In this research we apply the Diffusion Maps (DM) methodology [1] to a web traffic log. DM finds the embedded coordinates for a low-dimensional representation of the data. This embedding is accomplished by eigenvectors computation of the graph affinity matrix. Changes in the affinity matrix will result in changes in the eigenvectors, and thus will force us to compute them frequently. We use a solution based on the Recursive Power Iteration algorithm combined with the first-order approximation of the perturbed eigenvectors and eigenvalues (eigenpairs) [2]. This enables us to update the dataset profile by considering only the changes in the original dataset, which also requires less computational effort.

Since network data is dynamic and evolving, the embedded low-dimensional space has to be updated as the training data does not adequately represent the incoming data that did not participate in the initial training phase. Even if most of the network log lines in our interest window are unchanged, we will still need to perform the entire computation since we cannot determine the effect of such a change on the embedded space. Therefore, the goal of the paper is to provide an efficient method for updating the embedding coordinates without the need to re-compute the entire SVD again and again over time. We treat the log line feature changes as perturbations from the original network log profile of the feature affinity matrix. By applying a sliding window technique to the incoming network data, we are able to process the data online, and keep embedding it in the low-dimensional space. We test this method on real web traffic data and compare our results to the true classification.

II. Related Work

Traditional computational methods such as the power iteration, inverse iteration and Lanczos methods operate in the same way and compute the eigenpairs of each update of the perturbed matrix. Here, the computation is performed with a random guess as the initial input without taking the unperturbed matrix and its eigenpairs into consideration.

Incremental versions of low-dimensional embedding algorithms have been tailored specifically to fit local linear embeddings (LLE) [3] and ISOMAP [4]. These algorithms use modified manifold learning methods to process the data iteratively. When a new data point arrives, these algorithms add it to the embedding and then efficiently update all the existing data points in the low-dimensional space.

Network security has been one focus among the machine learning community. Kruegel and Vigna studied the parameters of HTTP queries using a training step with unlabeled data with various methods. Their character distribution analysis uses similar feature extraction as our current study [5]. Hubballi et al. described an n-gram approach to detect intrusions from network packets [6]. Ringberg et al. studied IP packets using principal component analysis-based dimensionality reduction [7]. Callegari et al. analyzed similar low-level packet data [8].

Diffusion maps have been also used for network security problems. David studied the use of diffusion map methodology for detecting intrusions in network traffic [9]. Network server logs have also been studied with diffusion maps with an offline approach using n-gram features and spectral clustering [10]. In these works, data analysis was performed in a batch fashion, processing all recordings as a single, offline dataset.

III. Finding a Low-Dimensional Embedded Space

A. Diffusion Maps

Finding a low-dimensional embedded space is an important step in understanding high-dimensional data more profoundly. To better understand the proposed algorithm, we review the DM methodology [1] that performs non-linear dimensionality reduction. Given our web log feature matrix X, we define a weighted graph over the log lines, where the weight between lines i and j is given by the kernel $k(i,j) \triangleq e^{-\frac{\|x_i - x_j\|}{\varepsilon}}$. The degree of a log line (vertex) i in this graph is $d(i) \triangleq \sum_j k(i,j)$. Normalizing the kernel with this degree produces an $n \times n$ row stochastic transition matrix whose cells are $[P]_{ij} = p(i,j) = k(i,j)/d(i)$ for log lines i and j. This defines a Markov process over the network log features.

The dimensionality reduction achieved by this diffusion process is a result of the spectral analysis of the kernel. Thus, it is preferable to work with a symmetric conjugate to P that we denote by A and its cells are denoted by

$$[A]_{ij} = a(i,j) = \frac{k(i,j)}{\sqrt{d(i)}\sqrt{d(j)}} = \sqrt{d(i)}p(i,j)\frac{1}{\sqrt{d(j)}}. \quad (1)$$

The eigenvalues $1 = \lambda_1 \geq \lambda_2 \geq \ldots$ of P and their corresponding eigenvectors v_k $(k = 1, 2, \ldots)$ are derived from the eigenvectors u_k of A. The v_k are used to obtain the desired dimensionality reduction by mapping each i onto the data point $\Psi(i) = (\lambda_2 v_2(i), \lambda_3 v_3(i), ..., \lambda_\delta v_\delta(i))$ for a sufficiently small δ, which depends on the decay of the spectrum of A [1].

In matrix notation, the operator A is defined as $A = D^{-\frac{1}{2}} K D^{-\frac{1}{2}} = D^{\frac{1}{2}} P D^{-\frac{1}{2}}$ where D is the diagonal matrix containing the $d(i)$ value in cell D_{ii}. To retrieve the eigenvectors in columns V of P from the eigenvactors of A, we use the transformation $V = D^{-\frac{1}{2}} U$ where U is the eigenvector column matrix of A. The eigenvectors V obtained for P are scaled by dividing each one by the first value of the first eigenvector.

B. Updating the Embedding

Once we have the DM embedding of the initial matrix A, we need to keep updating the embedding for the next arriving samples. By replacing the oldest samples with the newly arriving ones, we can model such a change as a perturbation matrix $\tilde{A}$ of the matrix A. We assume that the perturbations are sufficiently small, that is, $\|\tilde{A} - A\| < \varepsilon$ for some small ε. Note that $\tilde{A}$ is symmetric since it represents the operator defined in 1. We wish to update the eigenpairs of $\tilde{A}$ based on A and its eigenpairs. We now present the problem in mathematical terms.

Given a symmetric $n \times n$ matrix A where its k dominant eigenvalues are $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_k$ and its eigenvectors are $\phi_1, \phi_2, ..., \phi_k$, respectively, and a perturbed matrix $\tilde{A}$ such that $\|\tilde{A} - A\| < \varepsilon$, find the perturbed eigenvalues $\tilde{\lambda}_1 \geq \tilde{\lambda}_2 \geq ... \geq \tilde{\lambda}_k$ and its eigenvectors $\tilde{\phi}_1, \tilde{\phi}_2, ..., \tilde{\phi}_k$ of $\tilde{A}$ in the most efficient way [2].

In the next section, we explain how such processing can be done using the recursive power iteration (RPI) algorithm.

IV. The Recursive Power Iteration (RPI) Algorithm

A. Eigenpair First-Order Approximation

To efficiently update each eigenpair of the perturbed matrix $\tilde{A}$, we first compute the first-order approximation of each eigenpair. Later, it will be used in our algorithm as the initial guess for the RPI algorithm.

Given an eigenpair (ϕ_i, λ_i) of a symmetric matrix A where $A\phi_i = \lambda_i \phi_i$, we compute the first-order approximation of the eigenpair of the perturbed matrix $\tilde{A} = A + \Delta A$. We assume that the change ΔA is sufficiently small, which results in a small perturbation in ϕ_i and λ_i. We look for $\Delta\lambda_i$ and $\Delta\phi_i$ that satisfy the equation

$$(A + \Delta A)(\phi_i + \Delta\phi_i) = (\lambda_i + \Delta\lambda_i)(\phi_i + \Delta\phi_i). \quad (2)$$

Using the methods described by Shmueli et al. [2], we can obtain the following first-order approximations for the eigenvalues and eigenvectors of $\tilde{A}$

$$\tilde{\lambda}_i = \lambda_i + \phi_i^T [\Delta A] \phi_i \quad (3)$$

and

$$\tilde{\phi}_i = \phi_i + \sum_{j \neq i} \frac{\phi_j^T [\Delta A] \phi_i}{\lambda_i - \lambda_j} \phi_j. \quad (4)$$

B. The Recursive Power Iteration Method

The power iteration method has proved to be effective when calculating the principal eigenvector of a matrix [11]. However, this method cannot find the other eigenvectors of the matrix. In general, an initial guess of the eigenvector is also important to guarantee fast convergence of the algorithm. In Algorithm IV.1, which we call recursive power iteration (RPI), the first-order approximations of the perturbed eigenvectors of $\tilde{A}$ are the initial guess for each power iteration. Once the eigenvector $\tilde{\phi}_i$ is obtained in step i, we transform $\tilde{A}$ into a matrix that has $\tilde{\phi}_{i+1}$ as its principal eigenvector. We iterate this step until we recover the k dominant eigenvectors of $\tilde{A}$.

The correctness of the RPI algorithm and its complexity analysis are given in the original article [2].

The justification for this approach is that the first-order approximation of the perturbed eigenvector is inexpensive, and each RPI step guarantees that this approximation converges to the actual eigenvector of $\tilde{A}$. The first-order approximation should be close to the actual solution we seek and therefore requires fewer iteration steps to converge.

Algorithm IV.1: Recursive Power Iteration Algorithm

Input: Perturbed symmetric matrix $\tilde{A}_{n\times n}$, number of eigenvectors to calculate k, initial eigenvectors guesses $\{v_i\}_{i=1}^{k}$, admissible error *err*

Output: Approximated eigenvectors $\left\{\tilde{\phi}_i\right\}_{i=1}^{k}$, approximated eigenvalues $\left\{\tilde{\lambda}_i\right\}_{i=1}^{k}$

for $i = 1 \rightarrow k$ **do**
  $\phi \leftarrow v_i$
  repeat
    $\phi_{next} \leftarrow \frac{\tilde{A}\phi}{\|\tilde{A}\phi\|}$
    $err_\phi \leftarrow \|\phi - \phi_{next}\|$
    $\phi \leftarrow \phi_{next}$
  until $err_\phi \leq err$
  $\tilde{\phi}_i \leftarrow \phi$
  $\tilde{\lambda}_i \leftarrow \frac{\tilde{\phi}_i^T \tilde{A} \tilde{\phi}_i}{\tilde{\phi}_i^T \tilde{\phi}_i}$
  $\tilde{A} \leftarrow \tilde{A} - \tilde{\phi}_i \tilde{\lambda}_i \tilde{\phi}_i^T$
end for

V. Sliding Window Diffusion Map

Using DM to embed high volumes of data can be computationally intensive. It is even more challenging when the data is generated online and needs to be processed continuously. Therefore, we try to process the incoming data with iterative methodology by using the sliding window model. A sliding window X takes into account the n latest measurements. In practice, it is an $n \times m$ matrix with features on the columns and samples on the rows. The samples are high-dimensional, so the dimensionality of the sliding window is reduced from m to d using DM. This $n \times d$ matrix X_r now contains the low-dimensional representation of the data. This reduction is done each time a new sample appears and the window moves. However, the consecutive update of the DM is a time-consuming process that requires singular value decomposition during each window.

When updating the window, we can replace the oldest measurement with a new one in the matrix X, therefore changing a single row in X. This means that one line and one column of the K matrix in the DM algorithm change. This change can be interpreted as a perturbation to the matrix K, and furthermore to the matrix A, which is defined using the K matrix. The RPI algorithm with first-order approximation solves the eigenvectors for perturbed matrices. This leads us to use the RPI algorithm instead of time-consuming SVD.

Algorithm V.1 outlines the sliding window DM method. First, it solves the eigenvectors for the initial window using SVD. Then the algorithm iteratively process the following windows until no new samples are available.

There are, some practical problems with this approach. First, the RPI algorithm might not be able to solve the eigenvectors for some low-rank matrices. It is possible to prevent this with standard SVD when a low-rank (or otherwise unsuitable) matrix is encountered. Second, the window size itself has to be decided. The changing scales of the data over time introduce a challenge to the sliding window algorithm. The initial window still determines the profile and scale for the beginning of the analysis. Big windows cover a larger representation of the data and thus include a more varied overview of the normal behavior. With smaller windows, the percentage of anomalies within the data might get too big, and detecting the normal state becomes more difficult. Small windows, however, require less computational time since they induce smaller matrices. Optimal window size would therefore be the smallest possible that contains a small enough percentage of anomalies within the data, enabling it to capture the normal samples correctly.

Algorithm V.1: Sliding Window Diffusion Map with RPI

Input: Dataset X, window width n, embedded dimension k, admissible error err.

Output: Anomaly score for points in X.

$\epsilon \leftarrow$ estimate kernel parameter for first window of size n.
$[K]_{ij} \leftarrow \exp\left(-\frac{||x_i - x_j||^2}{\epsilon}\right)$, where $i, j = 1 \dots n$
$D \leftarrow \mathrm{diag}(\sum_{i=1}^{n}[K]_{ij})$
$A \leftarrow D^{-\frac{1}{2}} K D^{-\frac{1}{2}}$
$U, \Lambda, U^T \leftarrow \mathrm{SVD}(A)$
while new sample x_t available, where $t > n$ **do**
  $l \leftarrow t \bmod n$
  Replace the row l in X with the new sample x_t.
  Update both row l and column l of the affinity matrix K.
  $D \leftarrow \mathrm{diag}(\sum_{i=1}^{n}[K]_{ij})$
  $\tilde{A} \leftarrow D^{-\frac{1}{2}} K D^{-\frac{1}{2}}$
  $U, \Lambda \leftarrow$ RPI with first-order approximation $(\tilde{A}, A, k, U, \Lambda, err)$
  $V \leftarrow D^{-\frac{1}{2}} U$
  $V \leftarrow \frac{V}{V_{1,1}}$
  $\Psi \leftarrow V\Lambda$
  Find anomalies in Ψ and rate all samples in X.
  $A \leftarrow \tilde{A}$
end while
Return aggregated anomaly scores for each sample in X.

Detecting the anomalies in the low-dimensional representation can be done in various ways. A straightforward approach is to calculate distances between the embedded samples and find the ones that deviate too far from the center of the dataset. This and other spectral clustering methods give good results for datasets that contain clear separation [12], [10]. Similarly, k-means or any other clustering algorithm can find possible normal as well as anomalous behavior in the data. The density of points in the low-dimensional space tells how far they are from the more clustered areas. These methods calculate the distances to neighboring points [9]. All these methods usually need a threshold value for the anomalous region.

In each iteration, we evaluate the anomaly level of the samples within the window. Each sample gets a score if it is classified as an anomaly according to the selected anomaly

detection method. The scores of each sample are added as the window moves. This cumulative anomaly score histogram may be used to determine the anomaly level of a point. Scoring is used because locally inside a window some samples might appear anomalous but globally, considering the whole dataset, they are not. Even if the sample looks like an anomaly in some windows, it still gets only a few scores globally.

VI. Experimental Results

For the experiment, we use a labeled proprietary dataset of queries to a web server, which is known to contain some network attacks. These web queries are in Apache combined log text file. To extract numerical features from this text file, only the changing parameter values are used. The frequencies of 2-grams in these parameters are calculated to a matrix. In this matrix, the rows represent the log lines, and the columns represent the different 2-grams we found. The entries in this matrix count how many times each specific 2-gram appeared in the parameters of a log line. See [10] for more information about this dataset and the feature extraction.

The web log we use has 4292 lines and contains 480 different 2-grams. Thus, the feature matrix has dimensions 4292×480. The experiment simulates the initial state when n samples, or log lines, have arrived. When a new line arrives, it is added to the current window, while the oldest sample is removed from the matrix. This is continued until no new samples are available. The algorithm tracks only the samples within the window so that the dynamically changing nature of the data can be followed. As the size of the window does not change, the eigenpair problem stays reasonably sized.

Anomaly detection with Euclidian distances finds the most deviating samples within a window. This leads to false alarms when using simple normalized anomaly metrics because inside a window a point might look anomalous. Its local abnormality might be evident, but it should not be classified as one since globally it is just a small deviation from the normal state. This fact promotes thresholding the non-normalized but centered low-dimensional representation $d_k = |\Psi_k - \mathrm{mean}(\Psi_k)|$ within one window using statistical threshold $\theta_k = c \cdot \mathrm{std}(d_k)$, where the parameter c has to be adjusted empirically, for each dimension k in the embedded space.

Figure 1 illustrates the scores each point gets as the sliding window moves. The number of times the data points are classified anomalous are plotted against time. The window width is set to $n = 1000$. This experiment uses only the second eigenpair, $k = 2$, Ψ_2 for the low-dimensional presentation. In our analysis, we use a value of $c = 10$ for the anomaly threshold calculation. These scores themselves indicate in how many windows each sample is considered anomalous: the data points that are considered attacks are clearly seen from 2500 to 3500. Notice that a sample might be considered anomalous in several windows, but in the global view it is not an anomaly. Therefore, we use another threshold, which is the horizontal red line in the figure. With this setup, we manage to reach an accuracy of 92.5% and a precision of 99.7% after tuning the parameters of the algorithm.

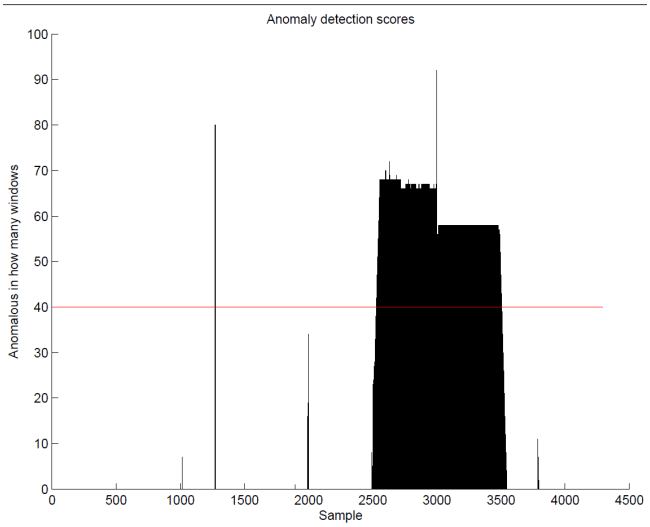


Fig. 1. The scores for each point with window size 1000 using the second eigenvector. The more times the data point is classified anomalous, the higher the score.

Acknowledgments

This research was supported by the Israel Science Foundation (Grant No. 1041/10) and by the Foundation of Nokia Corporation. The authors thank Antti Juvonen for assisting with the data analysis and Tapani Ristaniemi for guidance and support.

References

[1] R. R. Coifman and S. Lafon, "Diffusion maps," *Applied and Computational Harmonic Analysis*, vol. 21, no. 1, pp. 5–30, 2006.

[2] Y. Shmueli, G. Wolf, and A. Averbuch, "Updating kernel methods in spectral decomposition by affinity perturbations," *Linear Algebra and its Applications*, vol. 437, no. 6, pp. 1356–1365, 2012.

[3] O. Kouropteva, O. Okun, and M. Pietikäinen, "Incremental locally linear embedding algorithm," *Image Analysis*, pp. 145–159, 2005.

[4] M. Law and A. Jain, "Incremental nonlinear dimensionality reduction by manifold learning," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, no. 3, pp. 377–391, 2006.

[5] C. Kruegel and G. Vigna, "Anomaly detection of web-based attacks," in *Proceedings of the 10th ACM conference on Computer and communications security*. ACM, 2003, pp. 251–261.

[6] N. Hubballi, S. Biswas, and S. Nandi, "Layered higher order n-grams for hardening payload based anomaly intrusion detection," in *Availability, Reliability, and Security, 2010. ARES'10 International Conference on*. IEEE, 2010, pp. 321–326.

[7] H. Ringberg, A. Soule, J. Rexford, and C. Diot, "Sensitivity of PCA for traffic anomaly detection," *ACM SIGMETRICS Performance Evaluation Review*, vol. 35, no. 1, pp. 109–120, 2007.

[8] C. Callegari, L. Gazzarrini, S. Giordano, M. Pagano, and T. Pepe, "A novel PCA-based network anomaly detection," in *Communications (ICC), 2011 IEEE International Conference on*. IEEE, 2011, pp. 1–5.

[9] G. David, "Anomaly Detection and Classification via Diffusion Processes in Hyper-Networks," Ph.D. dissertation, Tel-Aviv University, 2009.

[10] T. Sipola, A. Juvonen, and J. Lehtonen, "Anomaly detection from network logs using diffusion maps," in *Engineering Applications of Neural Networks*, ser. IFIP Advances in Information and Communication Technology, L. Iliadis and C. Jayne, Eds. Springer Boston, 2011, vol. 363, pp. 172–181.

[11] A. Langville and C. Meyer, "Updating markov chains with an eye on google's pagerank," *SIAM journal on matrix analysis and applications*, vol. 27, no. 4, pp. 968–987, 2006.

[12] U. von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, pp. 395–416, 2007.

Finite Rate of Innovation Signals: Quantization Analysis with Resistor-Capacitor Acquisition Filter

Srikanth Tenneti
Electrical Engineering
California Institute of Technology
Pasadena, CA 91125
Email: stenneti@caltech.edu

Animesh Kumar and Abhay Karandikar
Department of Electrical Engineering
Indian Institute of Technology Bombay
Mumbai, India 400076
Email: animesh,karandi@ee.iitb.ac.in

***Abstract*—Sampling and perfect reconstruction of Finite rate of innovation (FRI) signals, which are usually not bandlimited, was introduced by Vetterli, Marziliano, and Blu [1].**

A typical FRI reconstruction algorithm requires solving for FRI signal parameters from a power-sum series. This in turn requires annihilation filters and polynomial root-finding techniques. These steps complicate the analysis of FRI signal reconstruction in the presence of *quantization*. In this work, we introduce a *three-channel* resistor-capacitor filter bank for the acquisition and reconstruction of FRI signals consisting of stream of Diracs and nonuniform splines. The effect of quantization error is derived for our three-channel filter-bank scheme. However, the sampling-rate required for our scheme is larger than the minimum sampling-rate of FRI signals.

I. Introduction

Parametric signals with finite degrees of freedom per unit time can be nonbandlimited [1]. E.g., for an integer $K_0 > 0$

$$x(t) = \sum_{k=0}^{K_0-1} c_k \delta(t - t_k), \quad c_k, t_k \in \mathbb{R}, \tag{1}$$

for all $0 \leq k \leq K_0 - 1$ is a parametric signal specified by the $2K_0$ real-valued parameters $\{(t_0, c_0), (t_1, c_1), \ldots, (t_{K_0-1}, c_{K_0-1})\}$. However, the Fourier bandwidth of $x(t)$ is infinite. If a signal is formed by the superposition of shifted and scaled versions of a known pulse, then the shifts and amplitudes of this pulse constitute its degrees of freedom rather than its Fourier bandwidth. Parametric signal class is large and it includes piecewise polynomials and non-uniform splines [1]. Signals, which can be specified by a finite number (or finite rate) of parameters are finite rate of innovation (FRI) signals. For an FRI signal, the degrees of freedom per unit time is the fundamental quantity to be used for determining the sampling rate [1].

The stream of Dirac delta signals in (1) has been widely studied due to its applicability in biomedical signal modeling, ultra-wideband communications, and global positioning system (e.g., see [2], [3]). Typically a power sum series has to be solved to obtain the parameters of an FRI signal. The solution involves annihilation filter and polynomial root finding [1], [4], [5], [6]. This approach is not amenable to quantization error analysis [7]. To the best of our knowledge, closed-form upper-bounds for quantization error in FRI signals are not known.

This work has been supported by IRCC, IIT Bombay by grant number P09IRCC039.

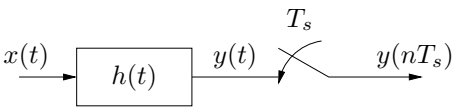


Fig. 1. *The FRI signal acquisition setup of Vetterli et al. [1] is illustrated. The filter $h(t)$ spreads the spikes thereby making $y(t)$ suitable for sampling. Typically $h(t)$ is a Gaussian or a sinc-filter.*

The acquisition filter $h(t)$ in Fig. 1 is a design choice. Consider FRI signals consisting of a stream of Dirac delta signals or nonuniform splines. For these FRI signals, a new sample acquisition setup consisting of a *three channel* resistor-capacitor (RC) filter-bank is proposed in this work. For this setup, *closed-form upper bounds* on error in FRI signal parameters due to quantization will be derived. Two channel or multi-channel resistor-capacitor filter banks for the reconstruction of stream of Diracs or other FRI signals, respectively, have been considered in the past for perfect reconstruction [8], [9]. However, these works do not study quantization and a detailed scheme for the sampling of nonuniform splines have not been suggested in them. The FRI signal reconstruction method proposed in this work *does not* involve a power sum series. This simplification, though, comes at a faster sampling rate than the minimum required by FRI signal sampling. [1]

Organization: Section II describes the signal model and our acquisition filter. Related work is reviewed in Section III. Perfect reconstruction and quantized reconstruction are discussed in Section IV. Conclusions are presented in Section V.

II. Modeling Assumptions

A finite-duration FRI signal is completely characterized by a finite number of parameters. Within the wide class of FRI signals, we consider the following signal model in this work,

$$x(t) = \sum_{k=0}^{K_0-1} c_{k,0}\delta(t - t_{k,0}) + \ldots + \sum_{k=0}^{K_p-1} c_{k,p}\delta^{(p)}(t - t_{k,p}) \tag{2}$$

[1] See C1 in Sec. IV-A for the exact condition.

where $\delta^{(r)}(t)$ denotes the r^{th} order derivative of the Dirac delta signal. The time epochs $t_{k,j} \neq t_{l,i}$ if $i \neq j$. Apart from modeling neural signals, it is also known that non-uniform splines can be reduced to the form of (2) after differentiation operations [1]. Due to space constraints, the signal model will be limited to stream of Dirac delta signal and its first derivative. The analysis extends in an analogous fashion to stream of Dirac delta signal and its higher order derivatives. Denote $x_l^m := (x_l, x_{l+1}, \ldots, x_m)$ for $m > l$. Given this signal, the parameters $\{(c_i)_0^{K_i-1}, (t_i)_0^{K_i-1}\}_{i=0}^{p-1}$ are to be (approximately) obtained from a set of sampled and quantized values obtained after filtering $x(t)$.

An ideal first-order RC filter will be used to facilitate the sampling of FRI signal in (2). Its impulse response is

$$h_{\text{rc}}(t) = e^{-\lambda t}u(t),$$

where $\lambda > 0$ is the *decay-rate* and $u(t)$ is the unit-step function. This filter is *causal* and can be implemented by a circuit consisting of single resistance of value R and single capacitor of value C. The decay-rate is $\lambda = 1/(RC)$. This passive filter is one of the simplest to implement in practice.

III. Prior Art

Sampling and perfect reconstruction of FRI signals with Gaussian and ideal lowpass acquisition filters was recently studied by Vetterli, Marziliano, and Blu [1]. These filters transform the problem of unknown Dirac delta signal (or its derivative) locations t_0^{K-1} to that of frequency estimation of a power sum series; the frequencies are estimated using annihilation filters. This method works well for perfect reconstruction.

FRI signal in (1) has been studied in application areas such as biomedical signal processing, ultra wideband communications, and global positioning system (e.g., see [2], [3]). Quantization noise analysis of FRI sampling and reconstruction has not been addressed [1], [5], [6] since the annihilation filter and polynomial root finding technique are complicated. Some quantization and oversampling results pertaining to FRI signal sampling are known in the literature [10]. Any *closed-form error analysis* due to quantization is mostly unsolved to the best of our knowledge.

In the presence of statistical sensing noise, the estimation of FRI signal parameters has also been studied in the literature. A qualitative analysis related to the numerical stability of some of these algorithms is presented in [5]. Related work includes the derivation of Cramer-Rao lower bounds for estimated poles of the power-sum series under additive Gaussian noise in [11]. This analysis, however, is restricted to a maximum of two delta functions.

IV. FRI signal reconstruction and quantization

Our sampling scheme for FRI signals in (2) is presented in two parts. Perfect reconstruction is presented first and quantization analysis is discussed in the later section.

A. Perfect reconstruction of Dirac delta signals with RC filters

Conceptually, a term of the form $c_k\delta^{(p)}(t - t_k)$, with $p = 0, 1$ has three degrees of freedom, namely, the constant c_k, the unknown order p, and the time instant t_k. Three RC-filters in parallel will be used to identify these three parameters. The general signal model is given by

$$x(t) = \sum_{k=0}^{K_0-1} c_{k,0}\delta(t - t_{k,0}) + \sum_{k=0}^{K_1-1} c_{k,1}\delta^{(1)}(t - t_{k,1}). \quad (3)$$

where the constants K_0, K_1 are positive integers. The time epochs $t_{k,j} \neq t_{l,i}$ if $i \neq j$. This signal class is obtained when piecewise linear signals are subjected to two differentiation operations. Consider the acquisition system shown in Fig. 2. There are three parallel RC filters with distinct decay-rate λ_1, λ_2, and λ_3. These filter outputs will be used to reconstruct the three degrees of freedom associated with every Dirac delta signal or its derivative present in $x(t)$.

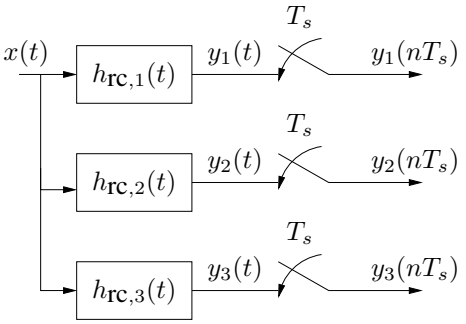


Fig. 2. *The three RC filters in parallel can be used to sample the signal in (3), provided T_s satisfies Condition C1 and $\lambda_1, \lambda_2, \lambda_3$ are distinct.*

Consider $\delta^{(1)}(t - t_0)$ as the input to an RC filter. Since delay and differentiation are linear time-invariant operations, the output of an RC filter with decay-rate λ_1 is given by

$$\begin{aligned}\frac{\mathrm{d}h(t-t_0)}{\mathrm{d}t} &= \frac{\mathrm{d}}{\mathrm{d}t}\left[\exp(-\lambda_1(t-t_0))u(t-t_0)\right] \\ &= \delta(t-t_0) - \lambda_1\exp[-\lambda_1(t-t_0)]u(t-t_0). \quad (4)\end{aligned}$$

Observe that, except at $t = t_0$ and a proportionality constant dependent on λ_1, these outputs are the same as the response of an RC filter to a Dirac delta signal at $t = t_0$. Define $\mathcal{T} := \{t_{0,0}, \ldots, t_{0,K_0-1}, t_{1,0}, \ldots, t_{1,K_1-1}\}$. The set $\mathcal{T}$ consists of all the points where Dirac delta signal or its derivative is present in the signal $x(t)$. Using linearity and the derivation in (4), it is straightforward to show that if $x(t)$ in (3) is the input to the system in Fig. 2, the output of the first filter is given by

$$y_1(t) = \sum_{k=0}^{K_0-1} c_{k,0}h_1(t - t_{k,0}) + \sum_{k=0}^{K_1-1} c_{k,1}(-\lambda_1)h_1(t - t_{k,1}).$$

The elements of $\mathcal{T}$ will be reordered for clarity in the analysis. Reorder the elements of set $\mathcal{T}$ in an ordered set $\{t_0, t_1, \ldots, t_{K-1}\}$ where $K = K_0K_1$. Due to re-ordering,

$t_i = t_{l_i,j_i}$ for some unique (l_i, j_i) pair for each i. Thus,

$$y_1(t) = \sum_{k=0}^{K-1} c_k h_1(t - t_k) = e^{-\lambda_1 t} \sum_{k=0}^{K-1} c_k u(t - t_k)$$

for $t \notin \mathcal{T}$, where $c_k = c_{l_k,j_k}(-\lambda_1)^{j_k}$.[2] The parameter c_k depends on λ_1. The decay-rate λ_1 is known but c_{l_k,j_k} and j_k are parameters to be obtained or approximated. After sampling at nT_s and multiplication by $e^{\lambda_1 nT_s}$, the following readings are obtained provided $nT_s \notin \mathcal{T}$:

$$e^{\lambda_1 nT_s} y_1(nT_s) = \sum_{k=0}^{K-1} c_k e^{\lambda_1 t_k} u(nT_s - t_k). \tag{5}$$

Now the following condition is assumed:

C1: $T_s < \min_i \{t_i - t_{i-1}\}$ and $nT_s \neq t_i$ for any i and n.

Under Condition C1, the different levels of the piecewise constant discrete-time signal in (5) reveal the product $c_k e^{\lambda_t t_k} = c_{l_k,j_k}(-\lambda_1)^{j_k} \exp(\lambda_1 t_{l_k,j_k})$ one by one for different values of $k = 0, 1, \ldots, K-1$. Under Condition C1, there is at least one sample between consecutive shifted Dirac or its derivative in $x(t)$; thus, for each Dirac or its derivative at t_i an integer $N_i \in \mathbb{Z}$ exists such that

$$\begin{aligned} e^{\lambda_1 N_i T_s} y_1(N_i T_s) - e^{\lambda_1 N_{i-1} T_s} y_1(N_{i-1} T_s) \\ = c_{l_i,j_i}(-\lambda_1)^{j_i} e^{\lambda_1 t_{l_i,j_i}}. \end{aligned} \tag{6}$$

The value of λ_1 is known. The following result of interest is stated and proved next. All the logarithms have base e.

Proposition 4.1: Assume that three RC-filters in parallel operate with distinct $(\lambda_1, \lambda_2, \lambda_3)$ and sampling rate T_s satisfies Condition C1. Then there exist indices $N_i, i = 0, 1, \ldots, K-1$ such that $e^{\lambda_1 N_i T_s} y_1(N_i T_s) = \sum_{k=0}^{i} c_k \exp(\lambda_1 t_k)$. Define $d_m(i) := e^{\lambda_m N_i T_s} y_m(N_i T_s) - e^{\lambda_m N_{i-1} T_s} y_m(N_{i-1} T_s)$ for $m = 1, 2, 3$. Choose $(\lambda_2)^2 = \lambda_1 \lambda_3$. Then the parameters of $x(t)$ in (3) are given by the following set of equations,

$$t_{l_i,j_i} = \frac{1}{\lambda_1 + \lambda_3 - 2\lambda_2} \log \left[\frac{d_1(i) d_3(i)}{d_2^2(i)} \right], \tag{7}$$

$$j_i = \frac{1}{\log\left(\frac{\lambda_1}{\lambda_2}\right)} \left[\log \left[\frac{d_1(i)}{d_2(i)} \right] + (\lambda_2 - \lambda_1) t_{l_i,j_i} \right], \tag{8}$$

$$\text{and } c_i = (-\lambda_1)^{j_i} c_{l_i,j_i} = \frac{d_1(i)}{e^{\lambda_1 t_{l_i,j_i}}}. \tag{9}$$

Proof: The existence of N_i has been argued while deriving (6); it follows from the definition of the unit-step function and (5). In the following equations, m takes the values $1, 2, 3$. The output of the three channels in Fig. 2 are given by,

$$y_m(N_i T_s) = e^{-\lambda_m N_i T_s} \sum_{k=0}^{i} c_k e^{\lambda_m t_k},$$

$$\text{or} \quad e^{\lambda_m N_i T_s} y_j(N_i T_s) = \sum_{k=0}^{i} c_k e^{\lambda_m t_k}.$$

[2] At locations mentioned in this set $\mathcal{T}$, Dirac delta signal and its derivatives are present.

Upon successive subtraction, we get

$$\begin{aligned} d_m(i) &= e^{\lambda_m N_i T_s} y_m(N_i T_s) - e^{\lambda_m N_{i-1} T_s} y_m(N_{i-1} T_s) \\ &= c_k \exp(\lambda_m t_k). \end{aligned} \tag{10}$$

The equations in (7), (8), and (9) follow from (10) by simple algebraic manipulations and using $\lambda_2^2 = \lambda_1 \lambda_3$. Since $\lambda_1, \lambda_2, \lambda_3$ are in geometric progression. The arithmetic mean of two unequal numbers is great than their geometric mean. Hence, $\lambda_1 + \lambda_3 > 2\lambda_2$. This ensures that the expression in (7) is well defined. ■

B. Quantization error in RC filter sampling scheme

In this section, it is assumed that $y_m(nT_s)$ values are quantized. Bounds on approximated parameters obtained through Proposition 4.1 will be derived. To work with scalar quantizers and maximum pointwise error, it is assumed that $|y_m(t)|$ is bounded. Without loss of generality, $|y_m(t)| \leq 1$ for all $t \in \mathbb{R}$. A uniform scalar quantizer will be assumed for analysis, where the quantizer precision is L-bits [12]. Let $\widehat{y}_m(nT_s)$ be the quantized value of $y_m(nT_s)$. Define $e_m(nT_s) := \widehat{y}_m(nT_s) - y_m(nT_s)$ Then, the following pointwise bound

$$|e_m(nT_s)| = |\widehat{y}_m(nT_s) - y_m(nT_s)| \leq 2^{-L}$$

holds for uniform scalar quantizer [12]. With quantized samples $\widehat{y}_m(nT_s)$, the approximate variables $d_m(i)$ have an error

$$\begin{aligned} &|\widehat{d}_m(i) - d_m(i)| \\ &= |e^{\lambda_m N_i T_s} e_m(N_i T_s) - e^{\lambda_m N_{i-1} T_s} e_m(N_{i-1} T_s)|. \end{aligned}$$

The FRI signal parameters can be approximated as follows:

$$\widehat{t}_{l_i,j_i} = \frac{1}{\lambda_1 + \lambda_3 - 2\lambda_2} \log \left[\frac{\widehat{d}_1(i) \widehat{d}_3(i)}{\widehat{d}_2^2(i)} \right], \tag{11}$$

$$\widehat{j}_i = \frac{1}{\log\left(\frac{\lambda_1}{\lambda_2}\right)} \left[\log \left(\frac{\widehat{d}_1(i)}{\widehat{d}_2(i)} \right) + (\lambda_2 - \lambda_1) \widehat{t}_{l_i,j_i} \right], \tag{12}$$

$$\text{and } \widehat{c}_{l_i} = (-\lambda_1)^{\widehat{j}_i} \widehat{c}_{l_i,j_i} = \frac{\widehat{d}_1(i)}{\exp(\lambda_1 \widehat{t}_{l_i,j_i})}. \tag{13}$$

Note that $\widehat{d}_m(N_i T_s) = c_i(m) e^{\lambda_m t_i} + e_m(N_i T_s) e^{\lambda_m N_i T_s} - e_m(N_{i-1} T_s) e^{\lambda_m N_{i-1} T_s}$. The constant c_i depends on m through λ_m for $m = 1, 2, 3$. The main result is stated next.

Theorem 4.1: Let $\widehat{y}_m(nT_s)$ be available with T_s satisfying Condition C1 and $m = 1, 2, 3$. Let $\lambda_1, \lambda_2, \lambda_3$ be distinct. Define the approximations for FRI signal parameters as in (11) and (13). Denote $\Delta := \lambda_1 + \lambda_3 - 2\lambda_2$. Then,

$$|\widehat{t}_i - t_i| \leq -\frac{4}{\Delta} \min_m \log \left[1 - \frac{2^{-L}(1 + e^{\lambda_m T_s})}{|c_i(m)|} \right]$$

$$\text{and } \left| \frac{\widehat{c}_i(m) - c_i(m)}{c_i(m)} \right|$$

$$\leq \left[1 + \frac{2^{-L}(1 + e^{\lambda_m T_s})}{|c_i(m)|} \right] \left[1 - \frac{2^{-L}(1 + e^{\lambda_{m^*} T_s})}{|c_i(m^*)|} \right]^{\frac{-4\lambda_m}{\Delta}} - 1,$$

where m^* is obtained by maximization as discussed in proof.

Proof: The proof omits algebraic steps for brevity. Define

$$\beta_{m,i} := \frac{e_m(N_iT_s)e^{\lambda_m[N_iT_s-t_i]} - e_m(N_{i-1}T_s)e^{\lambda_m[N_{i-1}T_s-t_i]}}{c_i(m)}.$$

Then

$$\begin{aligned}|\beta_{m,i}| &\leq \frac{2^{-L}|e^{\lambda_m[N_iT_s-t_i]}| + 2^{-L}|e^{\lambda_m[N_{i-1}T_s-t_i]}|}{|c_i(m)|} \\ &\leq \frac{2^{-L}(1+e^{\lambda_m T_s})}{|c_i(m)|} \qquad (14)\end{aligned}$$

since N_i can be chosen such that $0 < N_iT_s - t_i < T_s$. Using quantized estimates $\widehat{d}_m(i)$, we get

$$\frac{\widehat{d}_1(N_iT_s)}{\widehat{d}_2(N_iT_s)} = \frac{c_i(1)}{c_i(2)}e^{(\lambda_1-\lambda_2)t_i}\left[\frac{1+\beta_{1,i}}{1+\beta_{2,i}}\right].$$

Similarly,

$$\frac{\widehat{d}_3(N_iT_s)}{\widehat{d}_2(N_iT_s)} = \frac{c_i(3)}{c_i(2)}e^{(\lambda_3-\lambda_2)t_i}\left[\frac{1+\beta_{3,i}}{1+\beta_{2,i}}\right].$$

Note that $c_i(1)c_i(3) = (c_i(2))^2$. Therefore,

$$\frac{\widehat{d}_1(N_iT_s)\widehat{d}_3(N_iT_s)}{\widehat{d}_2^2(N_iT_s)} = e^{(\lambda_1+\lambda_3-2\lambda_2)t_i}\left[\frac{(1+\beta_{1,i})(1+\beta_{3,i})}{(1+\beta_{2,i})^2}\right].$$

Taking logarithms on both sides, we get

$$\begin{aligned}|\widehat{t}_i - t_i| &= \frac{1}{\Delta}\left|\log\left[\frac{(1+\beta_{1,i})(1+\beta_{3,i})}{(1+\beta_{2,i})^2}\right]\right|. \\ &\leq \frac{-4}{\Delta}\min_m \log\left[1 - \frac{2^{-L}(1+e^{\lambda_m T_s})}{|c_i(m)|}\right]. \\ &= \frac{-4}{\Delta}\log\left[1 - \frac{2^{-L}(1+e^{\lambda_{m^*} T_s})}{|c_i(m^*)|}\right]\end{aligned}$$

The last inequality utilizes the inequality $|\log(1+x)| \leq -\log(1-x_0)$ for all $|x| \leq x_0$. For very large values of L, note that the error is *decaying exponentially* in L as $\log(1-x) \approx -x$ for very small values of x.

The error in $\widehat{c}_i(m)$ will be derived now. From the definition of $\beta_{m,i}$

$$\widehat{c}_i(m) = c_i(m)e^{\lambda_m(t_i-\widehat{t}_i)}[1+\beta_{m,i}].$$

$$\text{or} \quad \frac{\widehat{c}_i(m) - c_i(m)}{c_i(m)} = e^{\lambda_m(t_i-\widehat{t}_i)}[1+\beta_{m,i}] - 1.$$

Thus,

$$\begin{aligned}\left|\frac{\widehat{c}_{m,i} - c_{m,i}}{c_{m,i}}\right| &= |e^{\lambda_m(t_i-\widehat{t}_i)}[1+\beta_{m,i}] - 1| \\ &\leq |e^{\lambda_m(t_i-\widehat{t}_i)} - 1| + |\beta_{m,i}e^{\lambda_m(t_i-\widehat{t}_i)}|\end{aligned}$$

Now we note that $|e^\theta - 1| \leq e^{|\theta|} - 1$ for any θ. Therefore,

$$\begin{aligned}&\left|\frac{\widehat{c}_i(m) - c_i(m)}{c_i(m)}\right| \\ &\leq e^{\lambda_m|t_i-\widehat{t}_i|} - 1 + |\beta_{m,i}|e^{\lambda_m|t_i-\widehat{t}_i|} \\ &\leq (1+|\beta_{m,i}|)\left[1 - \frac{2^{-L}(1+e^{\lambda_{m^*}T_s})}{|c_i(m^*)|}\right]^{-4\lambda_m/\Delta} - 1.\end{aligned}$$

Substituting the upper-bound on $\beta_{m,i}$ from (14),

$$\begin{aligned}&\left|\frac{\widehat{c}_i(m) - c_i(m)}{c_i(m)}\right| \\ &\leq \left[1 + \frac{2^{-L}(1+e^{\lambda_m T_s})}{|c_i(m)|}\right]\left[1 - \frac{2^{-L}(1+e^{\lambda_{m^*}T_s})}{|c_i(m^*)|}\right]^{\frac{-4\lambda_m}{\Delta}} - 1.\end{aligned}$$

As for $\widehat{t}_i$, if L is very large, then the error in $\widehat{c}_i(m)$ is proportional to 2^{-L}. ■

It must be noted that $\widehat{j}_i$ is either 0 or 1. For large-enough L, this parameter can be recovered exactly since j_i is discrete. Due to space constraints the derivations for $\widehat{j}_i$ and condition on L under which it can be recovered exactly is omitted.

V. Conclusions

In this work, a new sample acquisition method for sampling and reconstruction of an important class of FRI signals was explored. The new method, consisting of RC filters in parallel, studied in this work does not require solving a power-sum series, and ensuing annihilation filters or polynomial root finding, to obtain the FRI signal parameters. The effect of quantization error, in terms of upper bound on parameter reconstruction error, was addressed for our setup. Quantization error bounds are not available with the power-sum series approach. If L bits are used for quantizing each sample, then the reconstruction error was shown to be eventually decreasing as 2^{-L}. However, the sampling-rate required for our scheme is larger than the minimum sampling-rate of FRI signals.

References

[1] M. Vetterli, P. Marziliano, and T. Blu, "Sampling Signals with Finite Rate of Innovation," *IEEE Trans. Signal Proc.*, vol. 50, no. 6, pp. 1417–1428, June 2002.

[2] Y. Hao, P. Marziliano, M. Vetterli, and T. Blu, "Compression of ECG as a signal with finite rate of innovation," in *Proceedings of the 27th Annual International Conference in Engineering Medicine Biology Society*. New York, NY: IEEE-EMBS, 2006, pp. 7564–7567.

[3] J. Kusuma, I. Maravić, and M. Vetterli, "Sampling with finite rate of innovation: Channel timing and estimation for UWB and GPS," in *Proceedings of the International Conference on Communications*. New York, NY: IEEE, May 2003, pp. 3540–3544.

[4] I. Maravić and M. Vetterli, "Exact sampling results for some classes of parametric nonbandlimited 2-D signals," *IEEE Trans. Signal Proc.*, vol. 52, no. 1, pp. 175–189, Jan. 2004.

[5] ——, "Sampling and reconstruction of signals with finite rate of innovation in the presence of noise," *IEEE Trans. Signal Proc.*, vol. 53, no. 8, pp. 2788–2805, Aug. 2005.

[6] J. Kusuma and V. Goyal, "Multichannel sampling for parametric signals with a successive approximation property," in *Proc. of the Intl. Conf. in Image Processing*. New York, NY: IEEE, Oct. 2006, pp. 1265–1268.

[7] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Boston: Kluwer Academic, 1992.

[8] C. S. Seelamantula and M. Unser, "A generalized sampling method for finite-rate-of-innovation-signal reconstruction," *IEEE Signal Processing Letters*, vol. 15, pp. 813–816, 2008.

[9] H. Olkkonen and J. T. Olkkonen, "Measurement and reconstruction of impulse train by parallel exponential filters," *IEEE Signal Processing Letters*, vol. 15, pp. 241–244, 2008.

[10] I. Jovanović and B. Beferull-Lozano, "Oversampled a/d conversion and error-rate dependence of nonbandlimited signals with finite rate of innovation," *IEEE Trans. Signal Proc.*, vol. 54, no. 6, Jun. 2006.

[11] J. Kusuma and V. Goyal, "On the accuracy and resolution of powersum-based sampling methods," *IEEE Transactions on Signal Processing*, vol. 57, no. 1, pp. 182 –193, Jan. 2009.

[12] R. M. Gray and D. L. Neuhoff, "Quantization," *IEEE Transactions on Information Theory*, vol. IT-44, pp. 2325–2383, Oct. 1998.

Tangent space estimation bounds for smooth manifolds

Hemant Tyagi*, Elif Vural† and Pascal Frossard†
*Institute of Theoretical Computer Science
ETH Zurich, Switzerland
Email: htyagi@inf.ethz.ch
†School of Electrical Enginnering, EPFL, Switzerland
Emails: {elif.vural, pascal.frossard}@epfl.ch

***Abstract*—Many manifold learning methods require the estimation of the tangent space of the manifold at a point from locally available data samples. Local sampling conditions such as (i) the size of the neighborhood and (ii) the number of samples in the neighborhood affect the performance of learning algorithms. In this paper, we propose a theoretical analysis of local sampling conditions for the estimation of the tangent space at a point P lying on an m-dimensional Riemannian manifold S in $\mathbb{R}^n$. Assuming a smooth embedding of S in $\mathbb{R}^n$, we estimate the tangent space $T_P S$ by performing a Principal Component Analysis (PCA) on points sampled from the neighborhood of P on S. Our analysis explicitly takes into account the second order properties of the manifold at P, namely the *principal curvatures* as well as the higher order terms. Considering a random sampling framework, we leverage recent results from random matrix theory to derive local sampling conditions for an accurate estimation of tangent subspace. Our main results state that the width of the sampling region in the tangent space guaranteeing an accurate estimation is inversely proportional to the manifold dimension, curvature, and the square root of the ambient space dimension. At the same time, we show that the number of samples increases quadratically with the manifold dimension and logarithmically with the ambient space dimension.**

I. Introduction

A data set that resides in a high-dimensional Euclidean space and that is locally homeomorphic to a lower-dimensional Euclidean space constitutes a manifold. For example, a set of signals that is representable by a parametric model, such as parametrizable visual signals or acoustic signals forms a manifold. Data manifolds are however rarely given in an explicit form. The recovery of low-dimensional structures underlying a set of data, also known as manifold learning, has thus been a popular research problem in the recent years [1], [2], [3]. Importantly, most manifold learning methods rely on the assumption that the data has a locally linear structure. Of course, for such an assumption to be valid at some reference point on the manifold, one has to take into account (i) the size of the neighborhood from which the samples are chosen and also, (ii) the number of neighborhood points. For instance, if the manifold is a linear subspace, then the neighborhood can be chosen to be arbitrarily large and the number of samples needs to be simply greater than the dimension of the manifold. However, most manifolds are typically nonlinear, which prevents the selection of an arbitrarily large neighborhood size. Hence, one might expect the existence of an upper bound on the neighborhood size for a given estimation accuracy. Furthermore, the number of necessary samples is likely to vary according to the local characteristics of the manifold. In this work, we present an analysis of the sampling problem and derive conditions on the size of the sampling region and the number of samples for an accurate estimation of the tangent space.

This work has been performed while the first author was with the Signal Processing Laboratory LTS4 at EPFL.

The work has been partly funded by the Swiss National Science Foundation under Grant 200020-132772.

There are many examples of dimensionality reduction algorithms such as [3], [4], [5], [6], which apply a local Principal Component Analysis (PCA) for the computation of the tangent space of the manifold. The performance of Singular Value Decomposition (SVD) or PCA under noise is a well-studied topic (see [7], [8], [9] and references within). However these studies do not involve the geometric structure of the data. Only a few recent works have studied the relation between PCA performance and data geometry. The work in [10] generalizes the idea of diffusion maps in dimensionality reduction [11] to vector diffusion maps. As part of their analysis, the authors have shown in particular that when the size ε of the local area for tangent space estimation is set to $\varepsilon = O(K^{-\frac{1}{m+2}})$ with K being the number of samples on the *whole* manifold and m being the dimension of the manifold, then the deviation between the estimated and the true tangent space is typically $O(\varepsilon^{3/2})$. Their work however considers a global sampling from a compact manifold while we focus here on the local manifold geometry. Finally, the accuracy of tangent space estimation from noisy manifold samples is analyzed in a work parallel to ours [12]. This study focuses on manifolds that are embedded with exactly quadratic forms and poses the sampling problem as the selection of a subset of samples from a set of noisy samples given a priori. On the contrary, we analyze more generic embeddings with arbitrary smooth functions and we aim at characterizing a sampling strategy in terms of the sampling width and density for noiseless manifold samples.

Our contribution in this paper can be summarized as follows. Firstly, we determine a suitable upper bound on the size of the neighborhood in the tangent space within which the manifold can be sampled randomly. In the derivation of this bound, we consider the asymptotic case $K \to \infty$ with arbitrarily many manifold samples so that the neighborhood size purely depends on the manifold geometry. In particular, our analysis depends on (i) the maximum principal curvature of the manifold and (ii) the deviation of the manifold from its second-order approximation. Secondly, we compute a bound on the *minimum* number of samples for accurate tangent space estimation, given that the sampling is performed randomly in a neighborhood whose size conforms with the aforementioned bound. To this end, we utilize recent results from random matrix theory [13], [14]. Combining the two above results, we give a complete characterization of the local sampling conditions in Theorem 1.

The rest of the paper is organized as follows. Section II contains the formal outline of the problem. In Section III we present the main results along with a discussion. In Section IV we provide concluding remarks and possible directions for future work.

II. Problem setup

We consider an m-dimensional submanifold S of $\mathbb{R}^n$ with a smooth embedding in $\mathbb{R}^n$, $n \geq m+1$. Let $P \in S$ be a manifold point and $\mathcal{N}_\varepsilon(P)$ denote an ε-neighbourhood of P on S for some

$\varepsilon > 0$

$$\mathcal{N}_\varepsilon(P) = \{M \in S : \| M - P \|_2 \leq \varepsilon\}$$

where $\| \cdot \|_2$ stands for the ℓ_2-norm in $\mathbb{R}^n$. Let $T_P S$ denote the tangent space at P.

In our analysis, we represent the points in $\mathcal{N}_\varepsilon(P)$ via tangent space parameterization using local functions $f_l : T_P S \to \mathbb{R}$. There exists an ε such that all points $M \in \mathcal{N}_\varepsilon(P)$ can be uniquely represented in the form

$$[\bar{x}^T \ f_1(\bar{x}) \dots f_{n-m}(\bar{x})]^T. \tag{II.1}$$

Here $\bar{x} = [x_1 \dots x_m]^T$ denotes the coordinates of the orthogonal projection of manifold points onto $T_P S$. Note that, in (II.1), the coordinates are given with respect to the reference manifold point P, which is taken as the local origin. Furthermore, aligning the coordinate system with the tangent space at P, $T_P S$ can be represented as

$$T_P S = \text{span}\{\bar{e}_1, \dots, \bar{e}_m\},$$

where $\bar{e}_j \in \mathbb{R}^n$ denote the canonical vectors. Now, we further assume the smoothness of the embedding to be $\mathcal{C}^r$, $r > 2$, implying that each

$$f_l : T_P S \to \mathbb{R}, \quad l = 1, \dots, n-m,$$

is a $\mathcal{C}^r$-smooth function in the variables $(x_1, \dots, x_m)$. Since $\nabla f_l(\bar{0}) = \bar{0}$, we have the following identity by the Taylor expansion of f_l around the origin (P)

$$f_l(\bar{x}) = f_{q,l}(\bar{x}) + R_l(\bar{x}); \quad l = 1, \dots, n-m \tag{II.2}$$

where $f_{q,l}$ is a quadratic form and $R_l(\bar{x})$ is the remainder term of $O(\| \bar{x} \|_2^3)$. The Hessian of f_l at the local origin P can be represented as

$$\nabla^2 f_l(\bar{0}) = V_l \Lambda_l V_l^T,$$

where $\Lambda_l = \text{diag}(\mathcal{K}_{l,1}, \dots, \mathcal{K}_{l,m})$ and $\mathcal{K}_{l,1}, \dots, \mathcal{K}_{l,m}$ are the principal curvatures of the hypersurface

$$\mathcal{S}_l = \left\{[\bar{x}^T \ f_l(\bar{x})] : \bar{x} \in T_P S\right\} \subset \mathbb{R}^{m+1}$$

defined by f_l. We then define the maximum principal curvature at P as $\mathcal{K}_{max} := \mathcal{K}_{l',j'}$ where $(l', j') = \underset{l,j}{\text{argmax}} |\mathcal{K}_{l,j}|$. We consider that the tangent space is estimated from sample points in $\mathcal{N}_\varepsilon(P)$ through a PCA decomposition. More precisely, let us consider K points $\{P_i\}_{i=1}^K$ sampled from $\mathcal{N}_\varepsilon(P)$. Let $M^{(K)}$ denote the local covariance matrix where

$$M^{(K)} = \sum_{i=1}^{K} \frac{1}{K} P_i P_i^T = U \Lambda U^T.$$

The matrices U and $\Lambda \in \mathbb{R}^n$ represent respectively the eigenvector and eigenvalue matrices of $M^{(K)}$ where

$$U = [\bar{u}_1 \dots \bar{u}_m \dots \bar{u}_n]; \quad \Lambda = \text{diag}(\lambda_1, \dots \lambda_m, \dots \lambda_n),$$

with the ordering $\lambda_1 \geq \dots \geq \lambda_m \geq \dots \geq \lambda_n$. The optimal m-dimensional linear subspace at P in the least squares sense is then given by the span of the m largest eigenvectors of $M^{(K)}$, i.e.,

$$\widehat{T}_P S := \text{span}\{\bar{u}_1, \dots, \bar{u}_m\}.$$

Hence, $\widehat{T}_P S$ is the estimation of the true tangent space $T_P S$ at P with PCA. This is illustrated in Fig. 1. Finally, we characterize the accuracy of the tangent space estimation with the angle between $\widehat{T}_P S$ and $T_P S$, where we use the angle definition given in [15].

We can now state our problem formally. Given the above setting, we want to describe the conditions on the manifold samples $\{P_i\}_{i=1}^K$ such that for a given error bound $\phi \in (0, \frac{\pi}{2})$ on the tangent space estimation,

$$|\angle \widehat{T}_P S, T_P S| < \phi < \frac{\pi}{2}$$

is ensured. In particular, for a given error bound ϕ, we would like to answer the following questions:

- *Question 1:* What would be a suitable upper bound on the *sampling distance*; i.e., the distance from P_i to P?
- *Question 2:* Given that the points $\{P_i\}_{i=1}^K$ are sampled such that the sampling distance satisfies the sampling distance bound, what would be a suitable lower bound on the *sampling density* K?

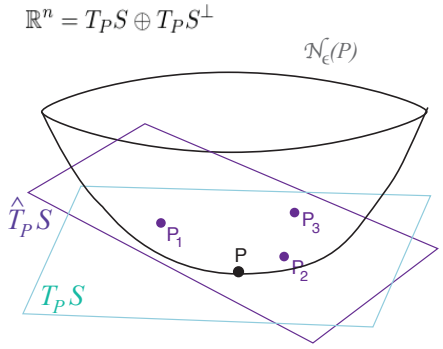


Fig. 1. The true tangent space $T_P S$ and the estimated tangent space $\widehat{T}_P S$ at point P.

In particular, for large embeddding dimensions n, we would like to determine the the dependency of the above bounds on n, m and $\mathcal{K}_{max}$. In order to address these questions, we consider a random sampling framework where we assume that the coordinates of the orthogonal projections of manifold samples on $T_P S$ are distributed uniformly in the region $[-\nu, \nu]^m \in T_P S$. In other words, we assume that

$$x_j^{(i)} \sim \mathcal{U}[-\nu, \nu] \quad \text{i.i.d.} \quad i = 1, \dots, K; j = 1, \dots, m$$

where $\mathcal{U}$ denotes the uniform distribution. Therefore, we characterize the *sampling distance* in Question 1 by the parameter ν, which we shall refer to as the sampling width in our analysis.

III. Main results

We now present our main results regarding the sampling of a smooth manifold. First, since we consider the sampling of $f_l(\bar{x})$ over the compact region $[-\nu, \nu]^m$, $R_l(\bar{x})$ is bounded over this region. Therefore, for each l there exists a constant $C_{s,l} > 0$ such that

$$|R_l(\bar{x})| < C_{s,l} \| \bar{x} \|_2^3 \quad l = 1, \dots, n-m,$$

where $C_{s,l}$ depends on the magnitude of the third order derivatives of f_l in $\mathcal{N}_\varepsilon(P)$. We denote

$$C_s = \max_l C_{s,l}, \quad l = 1, \dots, n-m.$$

The empirical covariance matrix $M^{(K)}$ corresponding to the samples $\{P_i\}_{i=1}^K$ is in the form $M^{(K)} = M_q^{(K)} + \Delta^{(K)}$

$$M_q^{(K)} = \begin{bmatrix} A^{(K)} & B^{(K)} \\ B^{(K)^T} & D^{(K)} \end{bmatrix}, \qquad \Delta^{(K)} = \begin{bmatrix} 0 & B_1^{(K)} \\ B_1^{(K)^T} & D_1^{(K)} \end{bmatrix}.$$

Here $M_q^{(K)}$ is the covariance matrix associated with the quadratic components $f_{q,l}(\bar{x})$ of the embeddings. The $m \times m$ matrix $A^{(K)}$ gives the covariance of the tangential components $\bar{x}_i$ of data points

P_i. As $K \to \infty$, the matrix $A^{(K)} \to \frac{\nu^2}{3} I_{m\times m}$ approaches a scaled version of the identity matrix; and therefore, the column space of $A^{(K)}$ approaches the true tangent space $T_P S$. The submatrices $B^{(K)}$ and $D^{(K)}$ represent the error on account of the nonzero manifold curvature at P, which stems from the second-order terms $f_{q,l}$. Meanwhile, $\Delta^{(K)}$ is an additional error term resulting from the higher-order Taylor terms R_l of the mappings f_l. We give the explicit formulation of $B_1^{(K)}$ and $D_1^{(K)}$ in [16, Section 4.4] and show that their Frobenius norms $\|B_1^{(K)}\|_F$ and $\|D_1^{(K)}\|_F$ can be bounded as

$$\begin{aligned} \| B_1^{(K)} \|_F &< \| B_1 \|_{F,bound} := \sqrt{m(n-m)} C_s m^{3/2} \nu^4, \\ \| D_1^{(K)} \|_F &< \| D_1 \|_{F,bound} \\ &:= (n-m) C_s m^{5/2} \nu^5 (C_s m^{1/2} \nu + |\mathcal{K}_{max}|). \end{aligned}$$

Now let us denote

$$B_1 = \mathbb{E}[B_1^{(K)}], \quad D_1 = \mathbb{E}[D_1^{(K)}], \text{ and } \Delta = \mathbb{E}[\Delta^{(K)}].$$

Due to the ergodicity of the sampling process, we have $B_1 = \lim_{K\to\infty} B_1^{(K)}$, $D_1 = \lim_{K\to\infty} D_1^{(K)}$, and $\Delta = \lim_{K\to\infty} \Delta^{(K)}$. Consequently, one can show that [16]

$$\|B_1\|_F < \|B_1\|_{F,bound}, \qquad \|D_1\|_F < \|D_1\|_{F,bound}.$$

Equipped with the above definitions and properties, we are now ready to state our main results about the sampling of smooth manifolds. We characterize the sampling conditions for accurate tangent space estimation by first defining a region of sampling in the tangent space and then determining the number of samples to be chosen from this region. We begin with the region of sampling and present in Lemma 1 the conditions on the sampling width ν that guarantee an upper bound on the angle between $\widehat{T}_P S$ and $T_P S$, provided that the number of samples is arbitrarily large.

Lemma 1: Let the sampling width satisfy

$$\nu < \frac{1}{[3((\beta_2 + RL) + \beta_3\alpha + \beta_4\alpha^2)]^{1/2}}$$

where $\beta_2 = 4C_s m^2 (n-m)^{1/2}$, $\beta_3 = 2(n-m)C_s m^{5/2}|\mathcal{K}_{max}|$, $\beta_4 = 2(n-m)m^3 C_s^2$,

$$R = n - m, \qquad L = \frac{m(5m+4)|\mathcal{K}_{max}|^2}{180}$$

and

$$\alpha = \min\left\{(3(\beta_2 + RL))^{-1/2}, (3\beta_3)^{-1/3}, (3\beta_4)^{-1/4}\right\}.$$

Then, as $K \to \infty$,

$$\mathbb{P}\left(|\angle \widehat{T}_P S, T_P S| > \cos^{-1}\sqrt{(1 - m\sigma_\infty^2)^m}\right) \to 0$$

where

$$\sigma_\infty = \frac{\| B_1 \|_{F,bound}}{\frac{\nu^2}{3} - RL\nu^4 - 2(\| B_1 \|_{F,bound} + \| D_1 \|_{F,bound})}.$$

The proof of Lemma 1 is presented in [16, Appendix A.4]. The stated result is derived from the condition that the spectrum associated with $\frac{\nu^2}{3} I_m$, whose corresponding eigenvectors give the true tangent space $T_P S$, is separated from the spectrum of the error. There are two sources of error here; namely, the curvature components $f_{q,l}$ which give rise to the correlation matrix $D = \lim_{K\to\infty} D^{(K)}$ and the higher-order terms R_l yielding the perturbation matrix Δ. The lemma states that the angle $|\angle \widehat{T}_P S, T_P S|$ between the estimated and true tangent spaces converges to the residual bound $\cos^{-1}\sqrt{(1 - m\sigma_\infty^2)^m}$ as the number of samples tends to infinity. The error term $m\sigma_\infty^2$ can be interpreted as the bias error resulting from the fact that a smooth embedding has a non-symmetric structure around the origin in general. In particular, it is easily verifiable that $\sigma_\infty \to 0$ as $\nu \to 0$; i.e., the bias approaches zero as the sampling width shrinks to 0. Also note that, when the f_l's are quadratic forms, this bias term vanishes to yield $\sigma_\infty = 0$, which is due to the symmetry of quadratic forms around the origin [16].

We now proceed to the finite sampling case $K < \infty$ and give our complete main result. In Theorem 1, we state the sufficient conditions on the sampling width ν and the number of samples K, such that the deviation $|\angle \widehat{T}_P S, T_P S|$ is suitably upper bounded with high probability.

Theorem 1: Let $s_1 \in (0,1)$ and $s_2 > e$ be fixed constants. Assume that the sampling width ν is such that

$$\nu < \left(\frac{s_1}{3[(\beta_2 + s_2 RL) + \beta_3\alpha + \beta_4\alpha^2]}\right)^{1/2}.$$

For some $\tau \in (0,1)$, let $s_3 > 0$ be chosen such that

$$\begin{aligned} s_3 <& [(s_1 \frac{\nu^2}{3} - s_2 RL\nu^4) - 2(\| B_1 \|_{F,bound} + \| D_1 \|_{F,bound})] \\ & \left(\frac{\tau^2}{m} + \sigma_f^2\right)^{1/2} - \| B_1 \|_{F,bound} \end{aligned}$$

where

$$\sigma_f = \frac{\| B_1 \|_{F,bound}}{(s_1 \frac{\nu^2}{3} - s_2 RL\nu^4) - 2(\| B_1 \|_{F,bound} + \| D_1 \|_{F,bound})}.$$

Finally, let $0 < p_1, p_2, p_3 < 1$. Assume that the number of samples satisfies $K > K_{bound}$, where $K_{bound} = \max\{K_{bound}^{(1)}, K_{bound}^{(2)}, K_{bound}^{(3)}\}$

$$\begin{aligned} K_{bound}^{(1)} &= \frac{6R_M}{(1-s_1)^2} \log\left((n-m+1)/p_1\right), \\ K_{bound}^{(2)} &= \frac{R_D}{s_2 RL} \frac{\log((n-m)/p_2)}{\log(s_2/e)}, \\ K_{bound}^{(3)} &= \frac{\nu^6 R_\sigma + \frac{R_B \nu^3 s_3}{3}}{s_3^2/2} \log(n/p_3) \end{aligned}$$

and

$$\begin{aligned} R_M &= m + \frac{1}{4}(n-m)m^2\nu^2|\mathcal{K}_{max}|^2 \\ R_D &= \frac{1}{4}(n-m)m^2|\mathcal{K}_{max}|^2, \quad R_B = \frac{1}{2}m^{3/2}\sqrt{n-m}|\mathcal{K}_{max}| \end{aligned}$$

$$R_\sigma = \frac{m^2|\mathcal{K}_{max}|^2}{12} \max\left\{(n-m), \frac{R(5m+4)}{15}\right\}.$$

Then, with probability larger than $1 - p_1 - p_2 - p_3$,

$$|\angle \widehat{T}_P S, T_P S| < \cos^{-1}\sqrt{(1 - \tau^2 - m\sigma_f^2)^m}.$$

The proof of Theorem 1 is presented in [16, Appendix A.5]. The theorem builds on Lemma 1 that considers the case $K \to \infty$. In the proof of the theorem, in order to account for finite K, we use the results of [13], [14] in order to probabilistically bound how much the tangent space estimated with K samples deviates from the tangent space in Lemma 1 estimated with infinitely many samples. The parameters s_i are used to make the link between the estimation error and the sampling conditions (sampling width and sampling density), whereas the probability constants p_i establish the relation between the error probability and the sampling density K.

Note that the tangent space estimation error in this case consists of two terms - the variance term τ due to finite sampling and the bias

term σ_f resulting from the asymmetric manifold geometry, which is the probabilistic counterpart of the bias term σ_∞ in Lemma 1. In particular, the number of samples K is related to the variance term τ through the parameter s_3, such that a larger number of samples decreases the variance, bringing thus the estimation error closer to its asymptotic value given in Lemma 1.

Remark: Let us now interpret our results with respect to the variation of the sampling conditions with the manifold parameters. As shown in [16], the results of our analysis translate into the fact that the choices $\nu = O(n^{-1/2}m^{-1}|\mathcal{K}_{max}|^{-1})$ and $K = O(\tau^{-2}m^2 \log n)$ ensure for large n that $|\angle\widehat{T}_P S, T_P S| < \cos^{-1}\sqrt{(1-\tau^2-O(n^{-1}m|\mathcal{K}_{max}|^{-4}))^m}$ holds w.h.p. In this work, the sampling width ν is measured on the tangent space $T_P S$. However, using the estimation $\|.\|_{ambient\,space} \approx O(\|.\|_{tangent\,space}\sqrt{n/m})$, we see that the stated bound on ν implies that the sampling width measured in the ambient space must change at the rate $O(\nu\sqrt{n/m}) = O(m^{-3/2}|\mathcal{K}_{max}|^{-1})$. This practically means that, when applying PCA, the size of the neighborhood around a reference point in the ambient space must get smaller as the intrinsic dimension m or the curvature $\mathcal{K}_{max}$ of the manifold increases, whereas it is not affected by the ambient space dimension n. On the other hand, the number of samples K increases quadratically with m and logarithmically with n.

Let us now briefly discuss the usage of our results with regards to two important application areas, namely (i) the discretization of a manifold with a known parametric model - *manifold sampling* and (ii) the recovery of the tangent space of a manifold from a given set of data samples - *manifold learning*. In order to use our results in a real application, the intrinsic dimension m of the manifold, the curvature parameter $\mathcal{K}_{max}$, and the higher-order deviation term C_s have to be known or estimated. First, in a manifold sampling application, m is already known and it is possible to estimate $\mathcal{K}_{max}$ in the following ways. If the manifold conforms to a known analytic model, it is easy to compute the values of the principal curvatures and the higher-order terms from the Taylor expansion of the model. If an analytic model is not known for the manifold, the curvature of a manifold of known parameterization can be estimated using results from Riemannian geometry such as [17, Section V] and [18, Proposition 2]. On the other hand, in a manifold learning application where only data samples are available, m, $\mathcal{K}_{max}$ and C_s are unknown and need to be estimated. The estimation of the intrinsic dimension of a data set has been studied in several works such as [19], [20] and [21]. It is also possible to obtain an estimate of the curvature and the deviation term C_s from data samples using results such as in [22].

IV. Conclusions

We have presented a theoretical analysis of the tangent space estimation at a point on a submanifold of $\mathbb{R}^n$ from a set of manifold samples that are selected locally at random. We have considered a setting where the manifold is embedded smoothly in $\mathbb{R}^n$ and the tangent space is estimated with local PCA. We have derived relations between the accuracy of the tangent space estimation and the sampling conditions. In particular, we have examined the effect of the local curvature of the manifold in tangent space estimation and shown that the size of the sampling neighborhood shall be inversely proportional to the manifold curvature. The presented study can be used for obtaining performance guarantees in the discretization of parametrizable data and in manifold learning applications. Finally, our analysis assumes that the data samples are noiseless, i.e., the data lies exactly on the manifold. A future research direction resides therefore in the extension of the current results to a scenario where data samples are corrupted with noise.

Acknowledgment

The authors would like to thank Prof. Daniel Kressner and Dr. Bart Vandereycken for the helpful discussions and comments on the manuscripts.

References

[1] J.B. Tenenbaum, V.D. Silva, and J.C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, pp. 2319–2323, 2000.

[2] S.T. Roweis and L.K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, pp. 2323–2326, 2000.

[3] D. L. Donoho and C. E. Grimes, "Hessian eigenmaps: Locally linear embedding techniques for highdimensional data," *Proc. Natl. Acad. Sci. USA*, vol. 100, pp. 5591–5596, 2003.

[4] Z. Zhang and H. Zha, "Principal manifolds and nonlinear dimension reduction via local tangent space alignment," *SIAM Journal of Scientific Computing*, vol. 26, pp. 313–338, 2002.

[5] Y. Yang, F. Nie, S. Xiang, Y. Zhuang, and W. Wang, "Local and global regressive mapping for manifold learning with out-of-sample extrapolation," in *Proc. of the 24th AIII Conf. on Art. Intel.*, 2010.

[6] Y. Zhan, J. Yin, G. Zhang, and E. Zhu, "Incremental manifold learning algorithm using PCA on overlapping local neighborhoods for dimensionality reduction," in *Advances in Computation and Intelligence*, vol. 5370 of *Lecture Notes in Computer Science*, pp. 406–415. Springer Berlin/Heidelberg, 2008.

[7] C. Davis and W. M. Kahan, "The rotation of eigenvectors by a perturbation, III," *SIAM J. Numer. Anal.*, vol. 7, Mar. 1970.

[8] P.A. Wedin, "Perturbation bounds in connection with singular value decomposition," *BIT Numerical Mathematics*, vol. 12, pp. 99–111, 1972, 10.1007/BF01932678.

[9] V. Vu, "Singular vectors under random perturbation," *Random Struct. Algorithms*, vol. 39, no. 4, pp. 526–538, Dec. 2011.

[10] A. Singer and H. Wu, "Vector Diffusion Maps and the Connection Laplacian," *Comm. on Pure and App. Math.*, 2012.

[11] R.R. Coifman, S. Lafon, A.B. Lee, M. Maggioni, F. Warner, and S. Zucker, "Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps," in *Proceedings of the National Academy of Sciences*, 2005, pp. 7426–7431.

[12] D. Kaslovsky and F.G. Meyer, "Optimal tangent plane recovery from noisy manifold samples," Submitted to the Annals of Statistics, available at http://arxiv.org/abs/1111.4601v2.

[13] A. Gittens and J. Tropp, "Tail bounds for all eigenvalues of a sum of random matrices," *Preprint*, 2011.

[14] J. Tropp, "User-friendly tail bounds for sums of random matrices," *Preprint*, 2011.

[15] H. Gunawan, O. Neswan, and W. Setya-Budhi, "A formula for angles between subspaces of inner product spaces," *Contributions to Algebra and Geometry*, vol. 46, pp. 311–320, 2005.

[16] H. Tyagi, E. Vural, and P. Frossard, "Tangent space estimation for smooth embeddings of Riemannian manifolds," [Online]. Available: http://arxiv.org/pdf/1208.1065.pdf.

[17] E. Kokiopoulou, D. Kressner, and P. Frossard, "Optimal image alignment with random projections of manifolds: algorithm and geometric analysis," *IEEE Transactions on Image Processing*, vol. 20, no. 6, pp. 1543–1557, 2011.

[18] L. Jacques and C. De Vleeschouwer, "A geometrical study of matching pursuit parametrization," *IEEE Transactions on Signal Processing*, vol. 56, no. 7, pp. 2835–2848, July 2008.

[19] M. Hein, "Intrinsic dimensionality estimation of submanifolds in Euclidean space," in *Proceedings of the* 22^{nd} *International Conference on Machine Learning*, 2005, pp. 289–296.

[20] E. Levina and P.J. Bickel, "Maximum likelihood estimation of intrinsic dimension," in *Advances in Neural Information Processing Systems*, 2005.

[21] G. Chen, A.V. Little, M. Maggioni, and L. Rosasco, "Some recent advances in multiscale geometric analysis of point clouds," *Wavelets and Multiscale Analysis: Theory and Applications*, March 2011.

[22] A.V. Little, J. Lee, Y.M. Jung, and M. Maggioni, "Estimation of intrinsic dimensionality of samples from noisy low-dimensional manifolds in high dimensions with multiscale SVD," in *Proc. of S.S.P.*, 2009.

A null space property approach to compressed sensing with frames

Xuemei Chen
Department of Mathematics
University of Maryland, College Park
Email: xuemeic@math.umd.edu

Haichao Wang
Department of Mathematics
U C Davis
Email: hchwang@ucdavis.edu

Rongrong Wang
Department of Mathematics
University of Maryland, College Park
Email: rongwang@math.umd.edu

***Abstract*—An interesting topic in compressive sensing concerns problems of sensing and recovering signals with sparse representations in a dictionary. In this note, we study conditions of sensing matrices A for the ℓ^1-synthesis method to accurately recover sparse, or nearly sparse signals in a given dictionary D. In particular, we propose a dictionary based null space property (D-NSP) which, to the best of our knowledge, is the first sufficient and necessary condition for the success of the ℓ^1 recovery. This new property is then utilized to detect some of those dictionaries whose sparse families cannot be compressed universally. Moreover, when the dictionary is of full spark, we show that AD being NSP, which is well-known to be only sufficient for stable recovery via ℓ^1-synthesis method, is necessary as well.**

I. Introduction

Compressed sensing concerns the problem of recovering a sparse signal $x_0 \in \mathbb{C}^d$ from its undersampled linear measurements $y = Ax_0 \in \mathbb{C}^m$, where the number of measurements m is usually much less than the ambient dimension d. A vector is said to be *k-sparse* if it has at most k nonzero entries. The following linear optimization algorithm, also known as the Basis Pursuit, can reconstruct x_0 efficiently from a perturbed observation $y = Ax_0 + w$ where $\|w\|_2 \leq \epsilon$ [8][4]:

$$\hat{x} = \arg\min_{x \in R^d} \|x\|_1, \quad \text{subject to } \|y - Ax\|_2 \leq \epsilon. \tag{1}$$

A primary task of compressed sensing is to choose appropriate sensing matrix A in order to achieve good performance of (1). A matrix A is said to have the *Restricted isometry property (RIP)* with order k if

$$(1-\delta)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1+\delta)\|x\|_2^2 \tag{2}$$

for any k-sparse vectors x. RIP is shown to provide stable reconstruction of approximately sparse signals via (1) [5][8]. Moreover, many random matrices satisfy RIP with high probability [6], [14]. A matrix A is said to have the *Null space property of order k (k-NSP)* if

$$\forall v \in \ker A \backslash \{0\},\ \forall |T| \leq k, \quad \|v_T\|_1 < \|v_{T^c}\|_1.$$

NSP is known as a characterization of uniqueness of problem (1) when there is no noise [10]. It has also been proven that the NSP matrices admit a similar stability result as RIP does except that the constants may be larger [1].

A recent direction of interest in compressed sensing concerns problems where signals are sparse in an overcomplete dictionary D instead of a basis, see [3], [13], [10], [11], [1], [12], [9]. This is motivated by the widespread use of overcomplete dictionaries in signal processing and data analysis. Many signals naturally possess sparse frame coefficients, such as images consisting of curves (curvelet frame). In addition, the greater flexibility and stability of frames make them preferable for practical purposes in order to compensate the imperfectness of measurements. In this setting, the signal $x_0 \in \mathbb{C}^d$ can be represented as $x_0 = Dz_0$, where z_0 is k-sparse and D is a $d \times n$ matrix with $n \geq d$. The columns of D may be thought of as an overcomplete frame or dictionary for $\mathbb{C}^d$. The linear measurements are $y = Ax_0$.

A natual way to recover x_0 from y is first solving

$$\hat{z} = \arg\min_{z \in R^n} \|z\|_1, \quad \text{subject to } y = ADz. \tag{3}$$

for the sparse coefficients $\hat{z}$, then synthesizing it to obtain $\hat{x}$, i.e., $\hat{x} = D\hat{z}$. The resulting method is therefore called ℓ^1-synthesis or synthesis based method [11], [13]. Since we are only seeking the recovery of x_0, we say the ℓ^1-synthesis method (3) is *successful* when every minimizer $\hat{z}$ of (3) satisfies $D\hat{z} = x_0$.

In the case when the measurements are perturbed, we naturally solve the following:

$$\hat{z} = \arg\min_{z \in R^n} \|z\|_1, \quad \text{subject to } \|y - ADz\| \leq \epsilon. \tag{4}$$

The work in [13] established conditions on A and D to make the compound AD satisfy RIP. However, as pointed in [3], [11], forcing AD to satisfy RIP or even the weaker NSP implies the exact recovery of both z_0 and x_0, which is unnecessary if we only care about obtaining a good estimate of x_0. In particular, if D is perfectly correlated (has two identical columns), then there are infinitely many minimizers of (3) that may be assigned to $\hat{z}$, but all of them lead to the true signal x_0. It seems reasonable to expect that similar result may hold in the case of highly correlated dictionaries, since they are only a small perturbation away from the perfectly correlated ones.

A. Overview and main results

In this paper, we generalize the ordinary null space property to the dictionary case (D-NSP), and prove in Theorem II.1 that this new condition is equivalent to the accurate recovery

of sparse signals in dictionaries via ℓ^1-synthesis. Moreover, a stability result is given in Theorem III.1. To the best of our knowledge, these results are the first characterization of compressed sensing with dictionaries via ℓ^1-synthesis approach.

Section IV studies more properties of D-NSP, and shows that A has D-NSP is equivalent to AD has NSP as long as D is of full spark (every d columns of D are linearly independent). As a consequence, under the full spark assumption, the ℓ^1-synthesis method cannot accurately recover the signals without accurate recoveries of their sparse representations, therefore an incoherent dictionary is needed under this circumstance.

All proofs of the theorems presented can be found in [7], while some proofs are provided here.

II. A SUFFICIENT AND NECESSARY CONDITION FOR NOISELESS SPARSE RECOVERY

In this section, we develop a sufficient and necessary condition for the success of ℓ^1-synthesis method (3).We show that the following property on A is a necessary and sufficient condition for successfully recovering all signals in $D\Sigma_s$ via (3), where $D\Sigma_k = \{x : \exists\ z,\ \text{such that } x = Dz, \|z\|_0 \le k\}$ is the set of signals that have k-sparse representations in D.

Definition 1 (Null space property of a dictionary D (D-NSP)). *Fix a dictionary $D \in \mathbb{C}^{d,n}$, a matrix $A \in \mathbb{C}^{m,d}$ is said to satisfy the D-NSP of order k (k-D-NSP) if for any index set T with $|T| \le k$, and any $v \in D^{-1}(\ker A \backslash \{0\})$, there exists $u \in \ker D$, such that*

$$\|v_T + u\|_1 < \|v_{T^c}\|_1. \tag{5}$$

Theorem II.1. *D-NSP is a necessary and sufficient condition for ℓ^1-synthesis* (3) *to successfully recover all signals in the set $D\Sigma_k$.*

Proof: Necessary part. We need to show that, if from measurements taken by a sensing matrix A, ℓ^1-synthesis is successful in recovering all signals in $D\Sigma_k$, then A must be k-D-NSP.

For any $v \in D^{-1}(\ker A/\{0\})$ and any index set T with $|T| = k$, we define $x_0 = Dv_T$ be a signal in $D\Sigma_k$, $y = Ax_0$ be its measurements, and let $\hat{x}$, $\hat{z}$ be the reconstructed signal and its coefficients from y via (3). If ℓ^1-synthesis is successful for all signals in $D\Sigma_k$, then we must have $\hat{x} = x_0$, and so $\hat{z} = v_T + u$ with some $u \in \ker D$.

Observe that $v_T - v$ is also feasible to (3), but it is not a minimizer since it cannot be representated in the form of $v_T + u$ with any $u \in \ker D$. Therefore, its ℓ_1 norm is strictly greater than that of $\hat{z}$:

$$\|v_T + u\|_1 < \|v_T - v\|_1 = \|v_{T^c}\|_1,$$

implying A is k-D-NSP.

Sufficient part. Assuming A is k-D-NSP, we will show that the ℓ_1 synthesis can recover all signals $x \in D\Sigma_k$ from $y = Ax$. Suppose to the contrary that there exists an $x_0 = Dz_0 \in D\Sigma_k$, such that its reconstruction $\hat{x} = D\hat{z}$ is wrong. Then we must have $v := z_0 - \hat{z} \in D^{-1}(\ker A/\{0\})$. Let T be the support of z_0, by D-NSP, therefore there exists a $u \in \ker D$, such that $\|v_T + u\|_1 < \|v_{T^c}\|_1$, i.e., $\|z_0 - \hat{z}_T + u\|_1 < \|\hat{z}_{T^c}\|_1$. Hence,

$$\|z_0+u\|_1 \le \|z_0-\hat{z}_T+u\|_1+\|\hat{z}_T\|_1 < \|\hat{z}_{T^c}\|_1+\|\hat{z}_T\|_1 = \|\hat{z}\|_1.$$

This is a contradicts to the assumption that $\hat{z}$ is a minimizer. □

Notice when D is the canonical basis of $\mathbb{C}^d$, the D-NSP is reduced to the normal NSP with the same order. In other words, D-NSP is a generalization of NSP for the dictionary case. It is, however, a nontrivial generalization.

The intuition of D-NSP rises from the fact that we are only interested in recovering x_0 instead of the representation z_0. As long as the minimizer $\hat{z}$ lies in the affine plane $z_0 + \ker D$, our reconstruction is a success.

III. D-NSP BASED STABILITY ANALYSIS

It is known that the NSP is a sufficient and necessary condition not only for the sparse and noiseless recovery, but also for compressible signals with noisy measurement [1], [15]. However, the stability analysis of NSP [1] cannot be easily generalized to our case because essentially we need the function $f(v) = (\|v_{T^c}\|_1 - \|v_T + u\|_1)/\|Dv\|_2$ to be bounded away from zero. In the basis case, we have knowledge of $f(v)$ on a compact set, and consequently the extreme value theorem can be applied to prove the exisitence of a positive lower bound. In our case we do not have a compact set, therefore other constructions to overcome this difficulty is necessary.

Definition 2 (Strong null space property of a dictionary D (D-SNSP)). *A sensing matrix A is said to have the strong null space property with respect to D of order k (k-D-SNSP) if for any index set T with $|T| \le k$, and any $v \in \ker(AD)$, there exists $u \in \ker D$, such that*

$$\|v_{T^c}\|_1 - \|v_T + u\|_1 \ge c\|Dv\|_2 \tag{6}$$

D-SNSP is a stronger assumption than D-NSP by definition. We prove that under this assumption, the ℓ^1-synthesis recovery is stable with respect to perturbations on the measurement vector y.

Theorem III.1. *If A is k-D-SNSP, then any solution $\hat{z}$ of problem* (4) *satisfies*

$$\|D\hat{z} - x_0\|_2 \le C_1\sigma_k(z_0) + C_2\epsilon.$$

where $\sigma_k(z_0)$ denotes the ℓ^1 residue of the best k-term approximation to z_0, C_1, C_2 are constants dependent on n, the c in (6)*, the minimum singular values of A and D, but not on x_0.*

Proof: Let $x_0 = Dz_0$ with z_0 being an k-sparse representation of x_0. Let $h = D(\hat{z} - z_0)$, and decompose it as $h = Dw + \eta$ where $Dw \in \ker A$, $\eta \in \ker A^\perp$. It is easy to show that $\|\eta\|_2 \le \frac{1}{\nu_A}\|Ah\|_2 \le \frac{2\epsilon}{\nu_A}$ with ν_A being the smallest singular value of A.

Define $\xi = D^T(DD^T)^{-1}\eta$, then $\eta = D\xi$, and

$$\|\xi\|_2 \le \frac{1}{\nu_D}\|\eta\|_2 \le \frac{2}{\nu_A\nu_D}\epsilon. \tag{7}$$

Moreover, by our setting, $D(\hat{z}-z_0)=h=D(w+\xi)$, and therefore $\hat{z}-z_0=w+\xi+u_1$ with some $u_1\in\ker D$.

Let $v=w+u_1$, then $\hat{z}-z_0=v+\xi$ and $v\in\ker(AD)$. By the assumption of D-SNSP, there exists a $u\in\ker D$ such that (6) holds for u and v. Therefore,

$$\begin{aligned}&\|v+z_{0,T}\|_1-\|-u+z_{0,T}\|_1\\ =&\|v_{T^c}\|_1+\|v_T+z_{0,T}\|_1-\|-u_T+z_{0,T}\|_1-\|u_{T^c}\|_1\\ \geq&\|v_{T^c}\|_1-\|v_T+u_T\|-\|u_{T^c}\|_1\\ =&\|v_{T^c}\|_1-\|v_T+u\|_1\geq c\|Dv\|_2 \end{aligned}\tag{8}$$

On the other hand, from the fact that $\hat{z}$ is a minimizer, we have

$$\begin{aligned}&\|-u+z_{0,T}\|_1+\|z_{0,T^c}\|_1\geq\|-u+z_0\|_1=\|\hat{z}\|_1\\ &\geq\|v+z_0+\xi\|_1\geq\|v+z_0\|_1-\|\xi\|_1\\ &\geq\|v+z_{0,T}\|_1-\|z_{0,T^c}\|_1-\|\xi\|_1.\end{aligned}$$

Rearrange the above inequality, we will obtain

$$\|v+z_{0,T}\|_1-\|-u+z_{0,T}\|_1\leq 2\|z_{0,T^c}\|_1+\|\xi\|_1.\tag{9}$$

Combining (8) and (9), we get

$$\|Dv\|_2\leq\frac{2}{c}\|z_{0,T^c}\|_1+\frac{1}{c}\|\xi\|_1\leq\frac{2}{c}\|z_{0,T^c}\|_1+\frac{\sqrt{n}}{c}\|\xi\|_2\tag{10}$$

In the end, using (10) and (7),

$$\begin{aligned}\|h\|_2=\|Dv+D\xi\|_2=\|Dv+\eta\|_2\leq\|Dv\|_2+\|\eta\|_2\\ \leq\frac{2}{c}\|z_{0,T^c}\|_1+\frac{\sqrt{n}}{c}\|\xi\|_2+\frac{1}{\nu_A}2\epsilon\\ \leq\frac{2}{c}\|z_{0,T^c}\|_1+\frac{2\sqrt{n}}{c\nu_A\nu_D}\epsilon+\frac{1}{\nu_A}2\epsilon.\end{aligned}$$

□

It is natural to ask how much stronger this new assumption is than D-NSP. We address this question partially in the next section.

IV. A further study of D-NSP and admissible dictionaries

This section explores the two assumptions D-NSP and D-SNSP further for the purpose of answering the following important questions: What kind of dictionaries will allow sensing matrices A with few measurements to satisfy D-NSP? How to find those sensing matrices given a dictionary?

We call a $d\times n$ dictionary D *k-admissible* if there exists a measurement matrix $A\in\mathbb{C}^{m,d}$ with $m<d$ such that A is k-D-NSP. We call D *inadmissible* if D is not k-admissible for any $k\geq 2$. Intuitively speaking, D is not k-admissible means that $D\Sigma_k$ cannot be universally compressed by any linear matrix A.

The following proposition shows that adding repeated columns to the dictionary D will not affect admissibility. This is quite intuitive since we do not change the set $D\Sigma_k$ during this procedure, and we only care about recovering the signal x_0 rather than the representation z_0.

Proposition IV.1. *Let $D\in\mathbb{C}^{d,n}$, and let I be any index set $I\subset\{1,...,n\}$. Define $\widetilde{D}=[D,D_I]$, then for any sensing matrix $A\in\mathbb{C}^{m,n}$, we have A is D-NSP if and only if A is $\widetilde{D}$-NSP.*

Proposition IV.1 states that a perfectly correlated dictionary D does not get in the way of the reconstruction of signals. It is only natural to ask whether this is still the case for a highly coherent dictionary. We answer this question partially by showing a class of highly correlated dictionaries is inadmissible. Moreover, equivalent conditions of D-NSP is given in Section IV-B under the assumption that D is of full spark.

A. A Class of inadmissible matrices

The following theorem constructs a class of inadmissible matrices with a one dimensional kernel.

Theorem IV.2. *Given an orthonormal basis $\Phi=[\phi_1,...,\phi_d]$. Let $H=\bigcup_{j=1}^{d}span\{\phi_i\}_{i=1,i\neq j}^{d}$ be the union of the hyperplanes spanned by every combination of $d-1$ columns of Φ. Then there exists a small constant r_0 such that for every $v\in B(\phi_1,r_0)\backslash H$ where $B(\phi_1,r_0)$ is the ball centered at ϕ_1 with radius r_0, $D=[\Phi,v]\in\mathbb{C}^{d,d+1}$ is inadmissible.*

We need the following lemma for the proof of this Theorem.

Lemma IV.3. *Suppose D is a $d\times(d+1)$ dictionary. If there exist $T\subset\{1,...,d+1\}$ with $|T|\geq 2$ such that any vector $u\in\ker D\backslash\{0\}$ satisfies*

1. *$\|u_T\|_1>\|u_{T^c}\|_1$, and*
2. *$T^c\subset supp(u)$,*

Then D cannot be $|T|$-admissible.

For any vector $w\in\mathbb{C}^n$, we define $\|w\|_{\min}=\min_{1\leq i\leq n}\{|w_i|\neq 0\}$ to be the minimum magnitude in w.

Proof: Assume that the dictionary D defined in Lemma IV.3 is $|T|$-admissible, we will show how this leads to a contradiction.

Since D is admissible, then there exists at least one A that is k-D-NSP. Pick one of them, and fix a $v_0\in D^{-1}(\ker(A)\backslash\{0\})$. Define $\alpha=2\|v_0\|_\infty/\|u\|_{\min}$. Now that $v_0+\alpha u,-v_0+\alpha u\in D^{-1}(\ker(A)\backslash\{0\})$, by the definition of D-NSP, there exist $c_1,c_2\in\mathbb{C}$ such that

$$\|v_T+\alpha u_T-c_1u\|_1<\|v_{T^c}+\alpha u_{T^c}\|_1,\tag{11}$$

and

$$\|-v_T+\alpha u_T-c_2u\|_1<\|-v_{T^c}+\alpha u_{T^c}\|_1.\tag{12}$$

Therefore,

$$\begin{aligned}&2\alpha\|u_{T^c}\|_1 &(13)\\ =\;&\|v_{T^c}+\alpha u_{T^c}\|_1+\|-v_{T^c}+\alpha u_{T^c}\|_1 &(14)\\ >\;&\|v_T+\alpha u_T-c_1u\|_1+\|-v_T+\alpha u_T-c_2u\|_1 &(15)\\ =\;&\|v_T+(\alpha-c_1)u_T\|_1+|c_1|\|u_{T^c}\|_1\\ +\;&\|-v_T+(\alpha-c_2)u_T\|_1+|c_2|\|u_{T^c}\|_1\\ \geq\;&|2\alpha-c_1-c_2|\|u_T\|_1+(|c_1|+|c_2|)\|u_{T^c}\|_1, &(16)\end{aligned}$$

where (14) follows from our assumption on α and Assumption 2, while (15) from adding (11) and (12). Combining (13) and (16) to get

$$\|u_T\|_1 < \|u_{T^c}\|_1.$$

This is a contradiction to Assumption 1 of Lemma IV.3. □

Proof of Theorem IV.2: Notice that $\ker(D) = \text{span}\{u\}$ with $u = (a^T, -1)$. Let T be an index set with $|T| \geq 2$ such that $\{1, n+1\} \in T$. First, since $v \notin H$, then $\langle v, \phi_i \rangle \neq 0$ for $i = 1, ..., d$. This means that all coordinates of u are nonzero, so Assumption 2 of Lemma IV.3 holds. Second, we can pick r_0 small enough such that whenever $v \in B(\phi_1, r)$, it holds $\|u_T\|_1 > \|u_{T^c}\|_1$, so Assumption 1 is satisfied.

Applying Lemma IV.3 completes the proof. □

We have constructed an example of inadmissible dictionaries of special sizes: $d \times (d+1)$. The following proposition asserts that this dictionary can be used to generate inadmissible dictionaries of arbitrary dimension by adding appropriate columns to it.

Proposition IV.4. *If $D = [B, v]$ where B is a full rank $d \times (n-1)$ matrix and $v = B\alpha$ with $\|\alpha\|_1 \leq 1$, then A has D-NSP implies that A has B-NSP with the same order k.*

B. The relation between D-NSP and NSP

It is obvious that AD satisfies NSP implies A satisfies D-NSP, which explains why imposing RIP or incoherence conditions on AD could be too strong and unnecessary. To explore how much room there is between these two conditions can possibly answer the question whether we can allow highly coherent dictionaries or not, since AD being NSP will inevitably leads to the incoherence of D. Surprisingly enough, we show that whenever D is of full spark, these two conditions are equivalent.

A dictionary is of full spark means every d columns of this matrix are linearly independent.

Theorem IV.5. *The following conditions are equivalent under the assumption that D is of full spark,*

- *A is k-D-NSP;*
- *AD is k-NSP;*
- *A is k-D-SNSP;*
- *For any $v \in \ker AD$, there exists a u such that*

$$\|v_T + u\|_1 < \|v_{T^c}\|_1.$$

Remark IV.1. *We comment that full spark is not a strong assumption on matrices. In fact, full spark matrices is dense in the space of matrices [2], and a large class of full spark Harmonic frames is also constructed in [2].*

Remark IV.2. *Earlier we mentioned that we only care about recovering the signals x and allow the recovery of their representations z to be wrong. Theorem IV.5 tells us that when the dictionary is of full spark this requirement is actually not any looser than requiring both signals and their representations to be recovered. In spite of being negative, this result is quite important, since it has been largely thought that the opposite is true.*

Like the RIP, NSP is essentially an incoherence property of a matrix. Hence a highly coherent dictionary D cannot be NSP, nor can the composite AD be, because whichever vector in $\ker D$ that fails to satisfy NSP, is also contained in $\ker(AD)$. Consequently, the equivalence of the first two items in Theorem IV.5 implies that if a highly coherent D is also full spark, then it must be inadmissible.

Perfectly coherent dictionaries are not full spark, so they can be and many of them are indeed admissible (Proposition IV.1). However, if these dictionaries are perturbed a little bit, then no matter how small the perturbations are, with probability one, they will turn into highly coherent and full spark dictionaries and therefore become inadmissible. We conclude that admissibility is not stable with respect to perturbations.

ACKNOWLEDGMENT

This research has been supported in part by Laboratory for Telecommunications Science (LTS) and by Defense Threat Reduction Agency HDTRA1-13-1-0015.

REFERENCES

[1] Akram Aldroubi, Xuemei Chen, and Alexander M. Powell. Perturbations of measurement matrices and dictionaries in compressed sensing. *Appl. Comput. Harmon. Anal.*, 33(2):282–291, 2012.

[2] Boris Alexeev, Jameson Cahill, and Dustin G. Mixon. Full Spark Frames. *J. Fourier Anal. Appl.*, 18(6):1167–1194, 2012.

[3] E. Candes, Y. C. Eldar, D. Needell, and P. Randall. Compressed sensing with coherent and redundant dictionaries. *Applied and Computational Harmonic Analysis*, 31(1):59–73, 2010.

[4] E. Candes, J. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math.*, 59:1207–1223, 2006.

[5] E. Candes and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, December 2005.

[6] E. Candes and T. Tao. Near optimal signal recovery from random projections and universal encoding strategies. *IEEE Transactions on Information Theory*, 52:5406–5425, 2006.

[7] Xuemei Chen, Haichao Wang, and Rongrong Wang. A null space analysis of the ℓ_1-synthesis method in frame-based compressed sensing. *in preperation*, 2013.

[8] S. Foucart and M. Lai. Sparsest solutions of underdetermined linear systems via l_q-minimization for $0 < q \leq 1$. *Applied and Computational Harmonic Analysis*, 26(3):395–407, 2009.

[9] R. Giryes and M. Elad. Can we allow linear dependencies in the dictionary in the sparse synthesis framework? *to appear in ICASSP*, 2013.

[10] R. Gribonval and M. Nielsen. Highly sparse representations from dictionaries are unique and independent of the sparseness measure. *Applied and Computational Harmonic Analysis*, 22(3):335–355, May 2007.

[11] Shidong Li, Tiebin Mi, and Yulong Liu. Performance analysis of ℓ_1-synthesis with coherent frames. *http://arxiv.org/abs/1202.2223*, 2012.

[12] D. Needell M. A. Davenport and M. B. Wakin. Signal space cosamp for sparse recovery with redundant dictionaries. *arXiv preprint arXiv:1208.0353*, 2012.

[13] H. Rauhut, K. Schnass, and P. Vandergheynst. Compressed sensing and redundant dictionaries. *IEEE Transactions on Information Theory*, 54(5):2210–2219, 2008.

[14] Mark Rudelson and Roman Vershynin. On sparse reconstruction from fourier and gaussian measurements. *Communications on Pure and Applied Mathematics*, 61:1025–1045, 2008.

[15] Qiyu Sun. Sparse approximation property and stable recovery of sparse signals from noisy measurements. *IEEE Trans. Signal Process.*, 59(10):5086–5090, 2011.

Irregular Sampling of the Radon Transform of Bandlimited Functions

Thomas Wiese
Associate Institute for Signal Processing
Technische Universität München, Germany
Email: thomas.wiese@tum.de

Laurent Demaret
Helmholtz Center Munich, Germany
Email: laurent.demaret@helmholtz-muenchen.de

***Abstract*—We provide conditions for exact reconstruction of a bandlimited function from irregular polar samples of its Radon transform. First, we prove that the Radon transform is a continuous L^2-operator for certain classes of bandlimited signals. We then show that the Beurling-Malliavin condition for the radial sampling density ensures existence and uniqueness of a solution. Moreover, Jaffard's density condition is sufficient for stable reconstruction.**

I. Introduction

In computed tomography (CT), a central question is the following: what kind of detail can be resolved from a particular CT scan (sinogram)? Notwithstanding the lack of a clear definition of the term detail, in many applications a satisfactory and useful answer is provided in terms of the Nyquist frequency connected to the sampling geometry.

The "pure" case of reconstructing a function from its possibly irregular samples has been solved nicely for classes of bandlimited functions in terms of density theorems as we will see in Section III of this paper. Due to difficulties that arise when defining the Radon transform for bandlimited functions, these results have not yet been used in the context of CT. Instead, efforts revolved around the analysis of quasi-bandlimited functions and the results bear the deficiency of only asymptotically controllable errors [1], [2].

This paper closes this apparent gap in the literature, namely, we show that the Radon transform can be defined in the usual sense as a continuous L^2-operator for certain classes of bandlimited functions. The Radon transform of such signals is itself bandlimited and it is shown that stable and exact reconstruction of these signals from their irregularly sampled Radon transforms is possible if the sampling set satisfies certain density requirements.

The remainder of this paper is organized as follows: In Section II we present current techniques in reconstruction of the sampled Radon transform and how they relate to spaces of bandlimited functions and sampling. In Section III we provide a dense overview of results from sampling theory for bandlimited functions. After showing continuity of the Radon transform and its inverse for certain bandlimited functions in Section IV, we will apply these results to the sampled Radon transform in Section V.

II. Motivation and Related Work

For efficient experimental design of CT scans, i.e. determination of a suitable sampling geometry or a posteriori choice of function spaces for reconstruction, it is essential to understand the discretization effects due to sampling. Furthermore, the emergence of CT acquisition procedures involving incomplete or irregular data calls for irregular sampling theory. Past research has focused on functions that are simultaneously essentially space- and band-limited, i.e., functions for which

$$\int_{\mathbb{R}^2\setminus B_R} |f(x)|^2\,dx \quad \text{and} \quad \int_{\mathbb{R}^2\setminus B_R} |(\mathcal{F}f)(\xi)|^2\,d\xi$$

decay exponentially with the radius R of the ball B_R. For these functions, interleaved sampling geometries are more efficient than regular sampling geometries [1], [2].

Bandlimitedness conditions also appear implicitly in reconstruction techniques based on discretizations of the inverse Radon transform. Filtered backprojection tacitly assumes that the Radon transform is bandlimited and periodic in the radial coordinate (for computation of a so-called absolute derivative operator) and that quadrature rules for the angular integral (for the backprojection) are exact—for example by assuming that the angular component of the Radon transform has a finite Fourier series representation [1].

Algorithms that are based on the Fourier slice theorem commonly use some sort of Fast Fourier Transform (FFT) for the radial variable to obtain the Fourier transform of the unknown function on a polar grid. This operation is either followed by interpolation onto a rectangular grid and application of the two-dimensional inverse FFT—a process known as gridding [3], [4]—or by using a version of the two-dimensional inverse FFT for non-rectangular grids [5], [6]. The assumptions are, again, that the Radon transform $\mathcal{R}f$ is bandlimited and periodic with respect to the radial coordinate and that f is bandlimited and periodic in both Cartesian variables.

Finally, algebraic reconstruction techniques can handle all sorts of irregular grids and are very efficient. However, it is difficult to analyze irregular sampling with such methods. First, it is hard to find function spaces for which the system matrix is injective and second, the combined effects of regularization, noise reduction, and early termination of iterative solvers are hard to quantify and isolate from sampling effects.

III. Review of Sampling Theory for Bandlimited Functions

We provide a brief review of the available theorems and techniques used for irregular sampling of bandlimited square-integrable functions in one dimension. These results can be extended to more dimensions when sampling on product grids.

Definition 1 (Paley-Wiener spaces)**.** *Let $\mathcal{F}$ denote the Fourier transform. The Paley-Wiener space of R-radially bandlimited and square-integrable functions is defined as*

$$\mathcal{PW}_R(\mathbb{R}^d) \coloneqq \{f \in L^2(\mathbb{R}^d) : \mathcal{F}f|_{\mathbb{R}^d \setminus B_R^d} = 0\},$$

where B_R^d is the d-dimensional ball with radius R. Similarly, for $r > 0$, we define the space of bandpass functions as

$$\mathcal{BP}_{rR}(\mathbb{R}^d) \coloneqq \{f \in L^2(\mathbb{R}^d) : \mathcal{F}f|_{\mathbb{R}^d \setminus (B_R^d \setminus B_r^d)} = 0\}.$$

Let $\Lambda \subset \mathbb{R}$ be a *uniformly discrete* set of sample positions, that is, $\inf_{\lambda,\mu} |\lambda - \mu| > 0$ for $\lambda, \mu \in \Lambda$ and $\lambda \neq \mu$. This condition ensures that the sampling operator $S_\Lambda \colon \mathcal{PW}_R(\mathbb{R}) \to l^2(\Lambda)$ is always a continuous linear operator into $l^2(\Lambda)$ [7]. For a fixed bandwidth R, sampling theory gives conditions in terms of densities on the sampling set Λ under which functions in $\mathcal{PW}_R(\mathbb{R})$ can be identified by and reconstructed from its values on Λ. In particular, one wishes to establish whether [8]

- Λ is a *set of uniqueness* for $\mathcal{PW}_R(\mathbb{R})$, i.e., the sampling operator S_Λ is injective or whether
- Λ is a *set of sampling* for $\mathcal{PW}_R(\mathbb{R})$, i.e., the sampling operator S_Λ is continuous and continuously invertible on its range.

Definition 2 (Densities)**.** *Let $\Lambda \subset \mathbb{R}$ be uniformly discrete with $0 \notin \Lambda$ and with signed counting function $N_\Lambda(t)$, which counts the number of points in the interval with endpoints 0 and t and has negative sign for $t < 0$.*

i) The Beurling-Malliavin density *is defined as*

$$D_{bm}(\Lambda) = \inf_{c \geq 0} c \text{ s.t. } \left\{ \begin{array}{l} \exists\, h \in C^1(\mathbb{R}),\, 0 \leq h'(t) \leq c, \\ \displaystyle\int_{\mathbb{R}} \frac{|N_\Lambda(t) - h(t)|}{1+t^2}\, dt < \infty \end{array} \right\}.$$

ii) The frame density *is defined as*

$$D_f(\Lambda) \coloneqq \sup_{\Gamma \subset \Lambda} \sup_{c \geq 0} c \text{ s.t. } N_\Gamma(t) - ct = O(1),$$

where the supremum is over all subsets Γ for which the asymptotics exist and $D_f(\Lambda) = 0$ if no such subset exists.

These densities apply to reasonably general sampling sets. In particular, the frame density is invariant under removals of finitely many points, i.e., one arbitrarily sized "hole" is allowed. In case of the Beurling-Malliavin density, it is possible to construct grids with $D_{bm}(\Lambda) = 1$ that have infinitely many "holes" of unbounded size [9]. If Λ is a set for which there exists $c > 0$ and for which the asymptotics $N_\Lambda(t) - ct = O(1)$ hold, one also says that Λ has *uniform density* c. Our definition of the Beurling-Malliavin density can be found in [10]. It is a simplification of the original *exterior density* $A_e(\, dN_\Lambda)$ used by Beurling and Malliavin [11], which also applies for sampling sets of complex numbers. The following theorem encapsulates several decades of research [11], [12], [13], [14].

Theorem 1 (Sampling theorems)**.** *For Λ to be a set of uniqueness for $\mathcal{PW}_\pi(\mathbb{R})$*

(i) it is necessary that $D_{bm}(\Lambda) \geq 1$,
(ii) it is sufficient that $D_{bm}(\Lambda) > 1$.

For Λ to be a set of sampling for $\mathcal{PW}_\pi(\mathbb{R})$

(i) it is necessary that $D_f(\Lambda) \geq 1$,
(ii) it is sufficient that $D_f(\Lambda) > 1$.

In the above theorem, it is possible to replace π with $R > 0$ and 1 with R/π on the right hand side of the density conditions. The last condition is known as Jaffard's sufficient condition.

Using the theory of frames, reproducing kernel Hilbert spaces (RKHS), and tensor products, one can generalize these results to two (and more) dimensions for Cartesian products of sampling grids [8], [15]. A RKHS $\mathcal{H}$ with domain $\mathbb{R}^d$ is a Hilbert space in which all point evaluations are continuous linear functionals, i.e, for all $x \in \mathbb{R}^d$, the map $f \mapsto f(x)$ is a bounded linear functional in $\mathcal{H}$ [16]. Paley-Wiener spaces and subspaces of $L^2(\mathbb{R}^d)$ spanned by a finite number of functions are examples of RKHS [17].

Theorem 2. *Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be RKHS of functions on $\mathbb{R}$. If for $i = 1, 2$, Λ_i is a set of uniqueness, resp. set of sampling, for $\mathcal{H}_i$, then $\Lambda_1 \times \Lambda_2$ is a set of uniqueness, resp. set of sampling, for $\mathcal{H}_1 \otimes \mathcal{H}_2$.*

The result is a consequence of the fact that tensor products of complete systems are complete in the tensor product space and that tensor products of frames are frames in the tensor product space [15].

IV. Radon Transform of Bandlimited Functions

In an effort to apply Theorem 1 to the irregularly sampled Radon transform, we first need to ensure *compatibility* between Paley-Wiener spaces and the Radon transform. Therefore, in this section, we establish conditions under which the Radon transform can be defined as a continuous and continuously invertible L^2-operator between subspaces of $\mathcal{PW}_R(\mathbb{R}^2)$ and $\mathcal{PW}_R(\mathbb{R}) \otimes L^2(S^1)$, where S^1 is the unit sphere in $\mathbb{R}^2$. Our approach contrasts with the conventional definition of the Radon transform as a continuous—but not continuously invertible—operator between $L^2(B_R(\mathbb{R}^2))$ and $L^2([-R, R] \times S^1)$. The advantage of our definition is that for irregular sampling grids of the form $\Lambda_s \times \Lambda_\omega$ with $\Lambda_s \subset \mathbb{R}$ and $\Lambda_\omega \subset [0, \pi]$, we can apply Theorems 1 and 2 to find conditions under which *exact* and *stable* reconstruction of a function from its sampled Radon transform is possible.

First, we present a counter example which highlights the difficulties that arise when defining the Radon transform for bandlimited functions. We will use the well-known Fourier slice theorem, which provides the following decomposition of the Radon transform for Schwartz functions $f \in \mathcal{S}(\mathbb{R}^2)$:

$$(\mathcal{R}f)(s, \omega) = (\mathcal{F}_s^{-1} \operatorname{id} \Phi \mathcal{F} f)(s, \omega).$$

Here, we denote the two-dimensional Fourier transform by $\mathcal{F}$, the one-dimensional Fourier transform with respect to the radial coordinate by $\mathcal{F}_s$, and the change from polar to Cartesian coordinates by $\Phi\colon L^2(\mathbb{R}^2) \to L^2(\mathbb{R}^+ \times S^1, \sigma\, d\sigma \otimes d\omega)$. By explicitly considering the *change of norms*,

$$\mathrm{id}\colon L^2(\mathbb{R}^+ \times S^1, \sigma\, d\sigma \otimes d\omega) \to L^2(\mathbb{R} \times S^1,\, d\sigma \otimes d\omega),$$

we retain the L^2-isometry property (up to some power of 2π) of the Fourier transforms and, hence, L^2-continuity of the overall operator is determined solely by that of the change of norms.[1] To show that the Radon transform is not L^2-continuous on $\mathcal{PW}_R(\mathbb{R}^2)$, consider the sequence $f_n \in \mathcal{PW}_R(\mathbb{R}^2)$ defined through its Fourier transform:

$$(\mathcal{F}f_n)(\xi) = \begin{cases} |\xi|^{-1/2} & \text{if } n^{-1} \le |\xi| \le R, \\ 0 & \text{otherwise.} \end{cases}$$

Each f_n is bandlimited, because $(\mathcal{F}f_n)(\xi) = 0$ for $|\xi| > R$, and $f_n \in L^2(\mathbb{R}^2)$, because $\mathcal{F}f_n \in L^2(\mathbb{R}^2)$ as

$$\begin{aligned} \int_{\mathbb{R}^2} d\xi |(\mathcal{F}f_n)(\xi)|^2 &= \int_{S^1} d\omega \int_{n^{-1}}^{R} \sigma\, d\sigma \left|\sigma^{-1/2}\right|^2 \\ &= \int_{S^1} d\omega \int_{n^{-1}}^{R} d\sigma \\ &= 2\pi(R - n^{-1}). \end{aligned}$$

As can be seen, the norm of $\mathcal{F}f_n$ tends to $\sqrt{2\pi R}$ and that of f_n to $\sqrt{R/(2\pi)}$. On the other hand, the norm of $(\mathrm{id}\,\Phi\mathcal{F})(f_n)$ with respect to the "flat" measure $d\sigma \otimes d\omega$ and, hence, that of $\mathcal{R}f_n$ in $L^2(\mathbb{R} \times S^1)$, is unbounded, because

$$\begin{aligned} \int_{S^1} d\omega \int_{\mathbb{R}} d\sigma |(\Phi\mathcal{F}f_n)(\sigma,\omega)|^2 &= 2\int_{S^1} d\omega \int_{n^{-1}}^{R} d\sigma \left|\sigma^{-1/2}\right|^2 \\ &= 2\int_{S^1} d\omega \int_{n^{-1}}^{R} d\sigma\sigma^{-1} \\ &= 4\pi(\ln R + \ln n), \end{aligned}$$

which tends to infinity as n grows.[2] Thus, the Radon transform cannot be continuous on $\mathcal{PW}_R(\mathbb{R}^2)$.

This defect of the Radon transform is a consequence of the fact that the Fourier transforms of functions in $\mathcal{PW}_R(\mathbb{R}^2)$ may have mild singularities at the origin. If we restrict the Paley-Wiener space to bandpass functions $f \in \mathcal{BP}_{rR}(\mathbb{R}^2)$, we can easily verify the boundedness of the Radon transform:

$$\begin{aligned} \int_{S^1} d\omega \int_{\mathbb{R}} d\sigma\, |(\mathcal{F}f)(\sigma,\omega)|^2 &= 2\int_{S^1} d\omega \int_{r}^{R} d\sigma\, |(\mathcal{F}f)(\sigma,\omega)|^2 \\ &\le 2\int_{S^1} d\omega \int_{r}^{R} d\sigma \frac{\sigma}{r} |(\mathcal{F}f)(\sigma,\omega)|^2 \\ &= 2r^{-1} \int_{\mathbb{R}^2} d\xi |(\mathcal{F}f)(\xi)|^2. \end{aligned}$$

The same calculation—replace r with R in the denominator and turn around the inequality—also yields the lower bound

$$2R^{-1}\|f\|^2 \le \|\mathcal{R}f\|^2 \le 2r^{-1}\|f\|^2.$$

[1] For $\sigma < 0$ we define $(\mathrm{id}\, g)(\sigma,\omega) = g(-\sigma,-\omega)$, which ensures that the new variables can still be interpreted as polar coordinates.

[2] The factor 2 is a consequence of id mapping $\mathbb{R}^+$ to the whole real line.

This implies (e.g., [18], Thm. 4.48) closedness of the range and existence of a continuous inverse of the Radon transform:

Theorem 3 (Radon isomorphism)**.** *The Radon transform is a Hilbert space isomorphism between the Hilbert spaces $\mathcal{BP}_{rR}(\mathbb{R}^2)$ and $\mathcal{R}(\mathcal{BP}_{rR}(\mathbb{R}^2)) \subset L^2(\mathbb{R} \times S^1,\, ds \otimes d\omega)$.*

We can also characterize the range of the Radon transform for bandpass functions. The theory of tensor products of separable L^2-spaces yields the decomposition [19]

$$L^2(\mathbb{R} \times S^1) \simeq L^2(\mathbb{R},\, ds) \otimes L^2(S^1,\, d\omega).$$

The Fourier slice theorem shows that if $(\mathcal{F}f)(\xi) = 0$ for $|\xi| < r$ and $|\xi| > R$, then $(\mathcal{F}_s\mathcal{R}f)(\sigma,\omega)$ also vanishes for σ outside of $[-R,-r] \cup [r,R]$. Hence, if $f \in \mathcal{BP}_{rR}(\mathbb{R}^2)$, then $\mathcal{R}f \in \mathcal{BP}_{rR}(\mathbb{R}) \otimes L^2(S^1)$. Similarly, for $g \in \mathcal{BP}_{rR}(\mathbb{R}) \otimes L^2(S^1)$ with $g(s,\omega) = g(-s,-\omega)$, we can go the inverse way of the Fourier slice theorem to define a function $f = \mathcal{F}^{-1}\Phi^{-1}\,\mathrm{id}^{-1}\,\mathcal{F}_s g$. The same calculations as above show that $f \in \mathcal{BP}_{rR}(\mathbb{R}^2)$ and since, by definition, $g = \mathcal{R}f$, we obtain:

Theorem 4 (Range theorem for bandpass functions)**.** *The Radon transform maps $\mathcal{BP}_{rR}(\mathbb{R}^2)$ isomorphically to $\mathcal{BP}_{rR}(\mathbb{R} \times S^1)$, where we define*

$$\mathcal{BP}_{rR}(\mathbb{R} \times S^1) := \left\{ \begin{array}{c} f \in \mathcal{BP}_{rR}(\mathbb{R}) \otimes L^2(S^1) \\ \textit{with } f(-s,-\omega) = f(s,\omega) \end{array} \right\}.$$

Remark that for $g \in \mathcal{BP}_{rR}(\mathbb{R} \times S^1)$, the Helgason-Ludwig moment conditions [20] are automatically satisfied, because $\mathcal{F}_s g$ and all of its derivatives vanish around the origin:

$$\int_{\mathbb{R}} g(s,\omega) s^k ds = \left(\frac{i}{2\pi}\right)^k \frac{d^k}{d\sigma^k}(\mathcal{F}_s g)(0,\omega) = 0 \;\forall\, k \in \mathbb{N}_0.$$

V. Sampling Theorems for the Radon Transform

The developments from the previous section allow us to apply the theory of bandlimited functions to the sampled Radon transform. Because of the isomorphism property established in Theorems 3 and 4, the sampled Radon transform operator $\mathcal{R}_\Lambda\colon \mathcal{BP}_{rR}(\mathbb{R}^2) \to l^2(\Lambda), f \mapsto ((\mathcal{R}f)(\lambda))_{\lambda\in\Lambda}$ is continuous and continuously invertible on its range exactly if the sampling operator $S_\Lambda\colon \mathcal{BP}_{rR}(\mathbb{R} \times S^1) \to l^2(\Lambda), g \mapsto (g(\lambda))_{\lambda\in\Lambda}$ is continuous and continuously invertible on its range; hence, we can concentrate our analysis on the latter.

One possible way of getting rid of the symmetry requirement is to consider sampling sets of the form $\Lambda \subset \mathbb{R} \times [0,\pi]$ and interpret the Radon transform as a map from $\mathcal{BP}_{rR}(\mathbb{R}^2)$ to $\mathcal{BP}_{rR}(\mathbb{R}) \otimes L^2([0,\pi])$. The inverse of the sampled Radon transform is then given as $\mathcal{R}^{-1}\mathcal{I}S_\Lambda^{-1}$, where $\mathcal{I} : \mathcal{BP}_{rR}(\mathbb{R}) \otimes L^2([0,\pi]) \to \mathcal{BP}_{rR}(\mathbb{R} \times S^1)$ is the isomorphism defined by

$$(\mathcal{I}f)(s,\omega) = \begin{cases} f(s, \arg(\omega)) & \text{if } 0 \le \arg(\omega) < \pi, \\ f(-s, \arg(\omega) - \pi) & \text{otherwise.} \end{cases}$$

It is equally possible, but slightly more technical, to allow for sampling grids $\Lambda \subset \mathbb{R}^+ \times [0,2\pi]$. However, due to space limitations, we will postpone comments on that case to an upcoming journal publication.

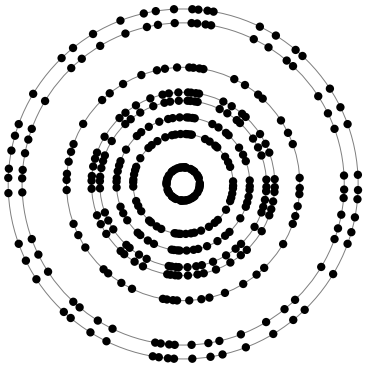


Fig. 1. Example of an irregular polar sampling grid in parallel geometry that is the Cartesian product of a radial and an angular irregular sampling grid. All circles may be unequally spaced, but the angular pattern must be the same on each circle. Only the first few circles are shown.

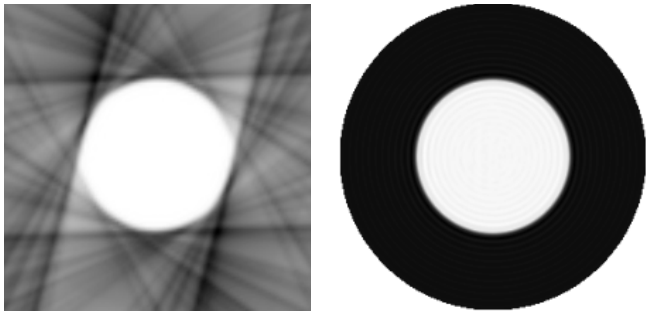


Fig. 2. Comparison of the reconstructions of a circle as obtained by Matlab's `iradon` function (Image Processing Toolbox), which uses filtered backprojection and a Ram-Lak filter, and a procedure that is based on sampling theory, where sinc-functions for radial components and complex exponentials for angular components were mapped through the inverse Radon transform.

Lastly, we need to restrict the angular behavior of admitted functions to some finite dimensional and, thus, automatically RKHS subspace $\mathcal{G} \subset L^2([0,\pi])$. The finiteness condition is a consequence of $[0,\pi]$ being bounded. One can then always find an angular sampling grid $\Lambda_\omega \subset [0,\pi)$ with $|\Lambda_\omega| = \dim(\mathcal{G})$ which is a set of uniqueness and a set of sampling for $\mathcal{G}$.

Theorem 5 (Sampling theorem for the Radon transform)**.** *Let Λ_ω be a set of uniqueness and, thus, set of sampling for some finite dimensional subspace $\mathcal{G} \subset L^2([0,\pi])$, $\Lambda_s \subset \mathbb{R}$ a uniformly discrete set and let $\Lambda = \Lambda_s \times \Lambda_\omega$. Let $\mathcal{H} \subset \mathcal{BP}_{rR}(\mathbb{R}^2)$ be defined as $\mathcal{H} = (\mathcal{R}^{-1} \circ \mathcal{I})(\mathcal{BP}_{rR}(\mathbb{R}) \otimes \mathcal{G})$ and let $\mathcal{R}_\Lambda : \mathcal{H} \to l^2(\Lambda)$ denote the sampled Radon transform.*

(i) For $\mathcal{R}_\Lambda$ to be injective it is sufficient that $D_{bm}(\Lambda_s) > R/\pi$.

(ii) For $\mathcal{R}_\Lambda$ to be continuous and continuously invertible it is sufficient that $D_f(\Lambda_s) > R/\pi$.

Proof: Due to Theorem 1, the conditions are sufficient for Λ_s being a set of uniqueness, resp. set of sampling, for $\mathcal{PW}_R(\mathbb{R})$ and thus also for the subspace $\mathcal{BP}_{rR}(\mathbb{R})$. With the assumptions on Λ_ω, we use Theorem 2 to see that Λ is a set of uniqueness, resp. set of sampling, for $\mathcal{BP}_{rR}(\mathbb{R}) \otimes \mathcal{G}$. Hence, the sampling operator $S_\Lambda \colon \mathcal{BP}_{rR}(\mathbb{R}) \otimes \mathcal{G} \to l^2(\Lambda)$ is injective, resp. continuous and continuously invertible on its range. These properties pass to the sampled Radon transform as all remaining operators are isomorphisms. ∎

Figure 1 shows an example of an irregular sampling grid in parallel geometry that is symmetric about the origin.

VI. Conclusion

As a consequence of the continuity of the inverse Radon transform of bandpass functions shown in Theorem 3, the reconstruction problem is not, strictly speaking, ill-posed, i.e., choosing $\mathcal{BP}_{rR}(\mathbb{R}^2)$ for reconstruction is stabilizing.

We will provide a reconstruction formula in an upcoming journal publication. For functions with finite angular Fourier series, the sinc function expansion is particularly suited for computation of the inverse Radon transform as all but a single one-dimensional integration can be carried out analytically using the theory of Bessel functions. To illustrate the practicality of our analytical reconstruction formula, we applied our method to the reconstruction of a non radially bandlimited image from its irregularly sampled Radon transform (Fig. 2).

References

[1] F. Natterer, *The Mathematics of Computerized Tomography*. Stuttgart: Teubner, 1986.

[2] P. Rattey and A. Lindgren, "Sampling the 2-d Radon transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 29, no. 5, pp. 994–1002, Oct. 1981.

[3] C. L. Epstein, *Introduction to the Mathematics of Medical Imaging*. Pearson Education, Inc., 2003.

[4] E. Margolis and Y. Eldar, "Nonuniform sampling in polar coordinates with applications to computerized tomography," in *Electrical and Electronics Engineers in Israel, 2004. Proceedings. 2004 23rd IEEE Convention of*, Sep. 2004, pp. 372–375.

[5] E. Grinberg and I. Pesenson, "Irregular sampling and the Radon transform," *Contemp. Math.*, vol. 251, pp. 255–268, 2000.

[6] D. Potts and G. Steidl, "Fourier reconstruction of functions from their nonstandard sampled Radon transform," *Journal of Fourier Analysis and Applications*, vol. 8, pp. 513–533, 2002.

[7] K. Gröchenig, "Reconstruction algorithms in irregular sampling," *Mathematics of Computation*, vol. 59, pp. 181–194, 1992.

[8] J. J. Benedetto and P. J. Ferreira, Eds., *Modern Sampling Theory: Mathematics and Applications*. Boston: Birkhäuser, 2001.

[9] N. Levinson, *Gap and Density Theorems*, ser. Colloquium Publication 26. New York: American Mathematical Society, 1940.

[10] J.-P. Kahane, "Travaux de Beurling et Malliavin," *Séminaire N. Bourbaki*, vol. 7, pp. 27–39, 1961-1962.

[11] A. Beurling and P. Malliavin, "On the closure of characters and the zeros of entire functions," *Acta Mathematica*, vol. 118, pp. 79–93, 1967.

[12] ——, "On fourier transforms of measures with compact support," *Acta Mathematica*, vol. 107, pp. 291–309, 1962.

[13] S. Jaffard, "A density criterion for frames of complex exponentials," *Michigan Math. J.*, vol. 38, no. 3, pp. 339–348, 1989.

[14] H. Landau, "Necessary density conditions for sampling and interpolation of certain entire functions," *Acta Mathematica*, vol. 117, pp. 37–52, 1967.

[15] A. Bourouihiya, "The tensor product of frames," *Sampling Theory in Signal and Image Processing*, vol. 7, no. 1, pp. 65–76, Jan. 2008.

[16] N. Aronszajn, "Theory of reproducing kernels," *Transactions of the American Mathematical Society*, vol. 68, no. 3, pp. 337–404, May 1950.

[17] K. Yao, "Applications of reproducing kernel Hilbert spaces – bandlimited signal models," *Information and Control*, vol. 11, no. 4, pp. 429–444, 1967.

[18] B. P. Rynne and M. A. Youngson, *Linear Functional Analysis*, 2nd ed. Springer, 2008, vol. 283.

[19] M. Reed and S. Barry, *Functional Analysis*. London, UK: AP, 1981.

[20] S. Helgason, *The Radon Transform*. Boston: Birkhäuser, 1980.

Spline-based frames for image restoration

Amir Averbuch
School of Computer Science
Tel Aviv University
Tel Aviv 69978, Israel
Email: amir@math.tau.ac.il

Pekka Neittaanmäki
Dept. of Mathematical Information Technology
University of Jyväskylä
P.O. Box 35 (Agora), Jyväskylä, Finland
Email: pekka.neittaanmaki@jyu.fi

Valery Zheludev
School of Computer Science
Tel Aviv University
Tel Aviv 69978, Israel
Email: zhel@post.tau.ac.il

***Abstract*—We present a design scheme to generate tight and semi-tight frames in the space of discrete-time periodic signals, which are originated from four-channel perfect reconstruction periodic filter banks. The filters are derived from interpolating and quasi-interpolating polynomial splines. Each filter bank comprises one linear phase low-pass filter (in most cases interpolating) and one high-pass filter, whose magnitude response mirrors that of a low-pass filter. In addition, these filter banks comprise two band-pass filters. In the semi-tight frames case, all the filters have linear phase and (anti)symmetric impulse response, while in the tight frame case, some of band-pass filters are slightly asymmetric. We introduce the notion of local discrete vanishing moments (LDVM). In the tight frame case, analysis framelets coincide with their synthesis counterparts. However, in the semi-tight frames, we have the option to swap LDVM between synthesis and analysis framelets. The design scheme is generic, and it enables to design framelets with any number of LDVM. The computational complexity of the framelet transforms, which consists of calculation of the forward and the inverse fast Fourier transforms and simple arithmetic operations, practically does not depend on the number of LDVM and on the size of the impulse response of filters. The designed frames are used for restoration of images, which are degraded by blurring, random noise and missing pixels. The images were restored by the application of the Split Bregman Iterations (SBI) method.**

I. Introduction

Restoration of corrupted and/or damaged and/or noised multidimensional signals is a major challenge that the signal/image processing community faces nowadays when rich multimedia content is the most popular data that is being transmitted over diverse networks types such as mobile. Quality degradation in multidimensional signals can come from sampling, acquisition, transmission through noisy channels, to name some. Restoration of multidimensional signals includes denoising, deblurring, recovering missing or damaged samples or fragments (inpainting in images), resolution enhancement and super resolution. Recent developments in wavelet frames (framelets) analysis provide innovative and powerful tools to meet faithfully and robustly the above challenges. Framelets produce redundant expansions whose valuable advantage is their ability to restore missing and incomplete information and to represent efficiently and compactly the data. In principle, only part of the samples/pixels is needed for (near) perfect object restoration. This approach, which is a variation of the *Compressive Sensing* methodology, proved to be extremely efficient for image restoration.

Practically, this approach is implemented via minimization of a parameterized functional where the sparse representation is reflected in the l_1 norm of the transform coefficients. The $\|\cdot\|_1$ minimization does not have an explicit solution and can be resolved only by iterative methods. The so-called *split Bregman iteration* (SBI) scheme, which was recently introduced in [1], provided a fast and stable algorithm for that. Variations of this scheme and its application to image restoration using wavelet frames are described in [2], [3], to mention a few. A variety of impressive results on image restoration were reported in the last couple of years. A survey is given in [4] while a recent development is described in [3].

Due to applications diversity, it is important to have a library of wavelet frames in order to select a frame that fits best a specific task. Forward and inverse transforms in iterative algorithms are repeated many times, therefore, members in this library must have fast and stable transforms implementation. Waveforms symmetry with the availability of vanishing moments are also important in order to avoid distortions when thresholding is used. To satisfy these requirements, most of the framelet systems that were designed so far operate with the compactly supported framelets and the transforms are implemented by finite (and short) impulse response (FIR) oversampled filter banks Thus, the number of framelet systems available for applications is very limited. This number is even smaller when the requirement is to have tight frames.

This limitation can be overcome by switching to a periodic setting, which is the subject of this presentation. A variety of four-channel PR filter banks, where the low-pass filters are derived from interpolating and quasi-interpolating polynomial splines, are designed. These filter banks generate a library of 4- framelet periodic tight and the so-called semi-tight frames with diverse properties. The transforms implementation is reduced to application of the direct and the inverse fast Fourier transforms (FFT) with simple arithmetic operations. While implementation of SBI in non-periodic setting requires multiple approximate solution of a system of equations by the conjugate gradient method, the periodic implementation makes it possible to avoid those procedures. This fact contributes significantly to reduction of the implementation cost.

The designed framelets libraries were tested for image restoration and demonstrated a high quality. Their diversity enabled us to select a frame, which best fits each specific application. In particular, in most of the experiments the semi-

tight frames outperformed tight frames.

II. Periodic filter banks and frame transform

We call the N-periodic real-valued sequences $\mathbf{x} \stackrel{\text{def}}{=} \{x[k]\},\ k \in \mathbb{Z},\ x[k+N] = x[k],\ N = 2^j$, the discrete-time periodic signals, which constitute an N-dimensional vector space $\Pi[N]$. We use the notation $\omega \stackrel{\text{def}}{=} e^{2\pi i/N}$. The circular convolution $y[k] = \sum_{k=0}^{N-1} h[k-l]\, x[l]$ of the signal $\mathbf{x}$ with a signal $\mathbf{h} \in \Pi[N]$ is called p-filtering and the signal $\mathbf{h}$ is called the p-filter. P-filtering results in multiplication of the DFT: $\hat{y}[n] = \hat{h}[n]\,\hat{x}[n]$.

It is well known that the perfect reconstruction (PR) oversampled filter banks generate frames in the signal space [5]. We deal with four channel analysis $\tilde{\mathbf{H}} \stackrel{\text{def}}{=} \left\{\tilde{\mathbf{h}}^s\right\}$, and synthesis $\mathbf{H} \stackrel{\text{def}}{=} \{\mathbf{h}^s\}\,,\ s = 0, ..., 3$ with downsampling factor of 2, which operate in the periodic signal space $\Pi[N]$. Either of $\tilde{\mathbf{H}}$, and $\mathbf{H}$ filter banks comprises one low-pass p-filter $\tilde{\mathbf{h}}^0$ and $\mathbf{h}^0$, one high-pass $\tilde{\mathbf{h}}^1$ and $\mathbf{h}^1$ and two band-pass $\tilde{\mathbf{h}}^s$ and $\mathbf{h}^s,\ s = 2, 3$, p-filters, respectively. The subsequent application of the time-reversed analysis and synthesis filter banks to an input signal $\mathbf{x} \in \Pi[N]$ restores the signal:

$$\begin{array}{ll} y^s[l] = & \sum_{k=0}^{N-1} \tilde{h}^s[k-2l]\, x[k],\ \ s = 0, \ldots, 3, \\ x[l] = & \sum_{s=0}^{S-1} \sum_{k=0}^{N/2-1} h^s[l-2k]\, y^s[k]. \end{array} \tag{1}$$

Denote by $\left\{\tilde{\psi}^s[k] = \tilde{h}^s[k]\right\}$ and $\{\psi^s[k] = h^s[k]\}$ the impulse responses of the analysis and synthesis p-filters, respectively. Equations (1) provide the frame expansion of a signal $\mathbf{x} \in \Pi[N]$:

$$x[l] = \sum_{s=0}^{3} \sum_{k=0}^{N/2-1} \psi^s[l-2k] \left\langle \mathbf{x},\, \tilde{\psi}^s[\cdot - 2k] \right\rangle. \tag{2}$$

The 2-sample shifts of the signals $\tilde{\psi}^s[k]$ and $\psi^s[k]$ form analysis and synthesis frames of the space $\Pi[N]$, respectively. Together they constitute a bi-frame $\left\{\tilde{\mathbf{F}},\, \mathbf{F}\right\}$. If the synthesis framelets can be chosen to be equal to the analysis framelets then the frame is tight.

III. Design of 4-channel PR filter banks

Denote by $\mathbf{x}_0 \stackrel{\text{def}}{=} \{x[2k]\} \in \Pi[N/2]$ and $\mathbf{x}_1 \stackrel{\text{def}}{=} \{x[2k+1]\}$ the even and odd polyphase components of a signal $\mathbf{x} \in \Pi[N]$. Then, the DFT of $\mathbf{x}$ is $\hat{x}[n] = \hat{x}_0[n] + \omega^n\, \hat{x}_1[n]$. Application of the 4-channel PR filter bank to a signal $\mathbf{x} \in \Pi[N]$ can be expressed in a matrix form. Denote $\vec{Y}[n] \stackrel{\text{def}}{=} (\hat{\tilde{y}}^0[n], ..., \hat{\tilde{y}}^3[n])^T$ and $\vec{X}[n] \stackrel{\text{def}}{=} (\hat{x}_0[n], \hat{x}_1[n])^T$. Then, we have

$$\vec{Y}[n] = \tilde{\mathbf{P}}[-n] \cdot \vec{X}[n], \quad \vec{X}[n] = \mathbf{P}[n] \cdot \vec{Y}[n],$$

where the 4×2 analysis and the 2×4 synthesis polyphase matrices are, respectively,

$$\begin{array}{ll} \tilde{\mathbf{P}}[n] \stackrel{\text{def}}{=} & \begin{pmatrix} \hat{\tilde{h}}_0^0[n] & ... & \hat{\tilde{h}}_0^3[n] \\ \hat{\tilde{h}}_1^0[n] & ... & \hat{\tilde{h}}_1^3[n] \end{pmatrix}^T, \\ \mathbf{P}[n] \stackrel{\text{def}}{=} & \begin{pmatrix} \hat{h}_0^0[n] & ... & \hat{h}_0^3[n] \\ \hat{h}_1^0[n] & ... & \hat{h}_1^3[n] \end{pmatrix}. \end{array}$$

The relations

$$\mathbf{P}[n] \cdot \tilde{\mathbf{P}}[-n] = \mathbf{I}_2, \tag{3}$$

is the condition for the pair $\left\{\tilde{\mathbf{H}},\, \mathbf{H}\right\}$ of filter banks to form a PR filter bank.

a) Design: The matrix product in Eq. (3) can be split into two products.

$$\mathbf{P}^{01}[n] \cdot \tilde{\mathbf{P}}^{01}[-n] + \mathbf{P}^{23}[n] \cdot \tilde{\mathbf{P}}^{23}[-n] = \mathbf{I}_2, \tag{4}$$

$$\mathbf{P}^{01}[n] \stackrel{\text{def}}{=} \begin{pmatrix} \hat{h}_0^0[n] & \hat{h}_0^1[n] \\ \hat{h}_1^0[n] & \hat{h}_1^1[n] \end{pmatrix},\ \tilde{\mathbf{P}}^{01}[n] \stackrel{\text{def}}{=} \begin{pmatrix} \hat{\tilde{h}}_0^0[n] & \hat{\tilde{h}}_1^0[n] \\ \hat{\tilde{h}}_0^1[n] & \hat{\tilde{h}}_1^1[n] \end{pmatrix},$$

$$\mathbf{P}^{23}[n] \stackrel{\text{def}}{=} \begin{pmatrix} \hat{h}_0^2[n] & \hat{h}_0^3[n] \\ \hat{h}_1^2[n] & \hat{h}_1^3[n] \end{pmatrix},\ \tilde{\mathbf{P}}^{23}[n] \stackrel{\text{def}}{=} \begin{pmatrix} \hat{\tilde{h}}_0^2[n] & \hat{\tilde{h}}_1^2[n] \\ \hat{\tilde{h}}_0^3[n] & \hat{\tilde{h}}_1^3[n] \end{pmatrix}.$$

A PR pair $\left\{\mathbf{H},\, \tilde{\mathbf{H}}\right\}$ of filter banks generate a tight frame if their polyphase matrices are linked as

$$\mathbf{P}[n] = \tilde{\mathbf{P}}[n]^T \Longleftrightarrow \mathbf{P}^{01}[n] = \tilde{\mathbf{P}^{01}}[n]^T \text{ and } \mathbf{P}^{23}[n] = \tilde{\mathbf{P}}^{23}[n]^T.$$

If the matrices $\mathbf{P}^{01}[n] = \tilde{\mathbf{P}}^{01}[n]^T$ and $\mathbf{P}^{23}[n] \neq \tilde{\mathbf{P}}^{23}[n]^T,\quad n \in \mathbb{Z}$, *then the frame* $\left\{\tilde{\mathbf{F}},\, \mathbf{F}\right\}$ *is called semi-tight.*

The design of four-channel (semi-)tight filter banks begins from a linear phase low-pass filter $\mathbf{h}^0 = \tilde{\mathbf{h}}^0$, whose frequency response (FR) $\hat{h}^0[n] = \hat{h}_0^0[n] + \omega^{-n}\hat{h}_1^0[n]$ is a rational function of $\omega^n = e^{2\pi i n/N}$ with real coefficients that has no poles for $n \in \mathbb{Z}$. Assume $\hat{h}^0[n]$ is symmetric about the swap $n \to -n$, which implies that $\hat{h}_0^0[n] = \hat{h}_0^0[-n]$ and $\omega^{-n}\hat{h}_1^0[n] = \omega^n\, \hat{h}_1^0[-n]$. The impulse response $\left\{h^0[k]\right\}$ is symmetric about $k = 0$.

In addition, assume that $\mathbf{P}^{01}[n] = \tilde{\mathbf{P}}^{01}[n]^T$ and the product

$$\mathbf{P}^{01}[n] \cdot \mathbf{P}^{01}[-n] = \begin{pmatrix} \alpha[n] & 0 \\ 0 & \beta[n] \end{pmatrix} \tag{5}$$

is a diagonal matrix. The assumption in Eq. (5) implies the condition $\hat{h}_0^0[n]\hat{h}_1^0[-n] + \hat{h}_0^1[n]\hat{h}_1^1[-n] = 0$. The simplest way to satisfy this condition is to define

$$\begin{array}{c} \hat{h}^1[n] = -\hat{h}_1^0[-n] + \omega^{-n}\hat{h}_0^0[-n] \\ \Longrightarrow \alpha[n] = \beta[n] = \left|\hat{h}_0^0[n]\right|^2 + \left|\hat{h}_1^0[n]\right|^2. \end{array}$$

The sequence $\omega^n\hat{h}^1[n] = \omega^{-n}\hat{h}^1[-n]$ and, consequently, the impulse response $\left\{h^1[k]\right\}$ is symmetric about $k = 1$. The product

$$\mathbf{P}^{23}[n] \cdot \tilde{\mathbf{P}}^{23}[-n] = \mathbf{Q}[n] \stackrel{\text{def}}{=} \begin{pmatrix} t[n] & 0 \\ 0 & t[n] \end{pmatrix}, \tag{6}$$

where $t[n] \stackrel{\text{def}}{=} 1 - \left|\hat{h}_0^0[n]\right|^2 + \left|\hat{h}_1^0[n]\right|^2$. Thus, the design of the PR filter bank is reduced to factorization of the matrix $\mathbf{Q}[n]$.

There are many ways to factorize the matrix $\mathbf{Q}[n]$. One way is to define the matrices $\mathbf{P}^{23}[n]$ and $\tilde{\mathbf{P}}^{23}[n]$ to be diagonal:

$$\mathbf{P}^{23}[n] = \begin{pmatrix} \hat{h}_0^2[n] & 0 \\ 0 & \hat{h}_1^3[n] \end{pmatrix},\ \tilde{\mathbf{P}}^{23}[n] \stackrel{\text{def}}{=} \begin{pmatrix} \hat{\tilde{h}}_0^2[n] & 0 \\ 0 & \hat{\tilde{h}}_1^3[n] \end{pmatrix}.$$

Consequently, we have to derive four sequences $\hat{h}_0^2[n]$, $\hat{\tilde{h}}_0^2[n]$, $\hat{h}_1^3[n]$ and $\hat{\tilde{h}}_1^3[n]$ such that

$$\hat{h}_0^2[n]\,\hat{\tilde{h}}_0^2[n] = \hat{h}_1^3[n]\,\hat{\tilde{h}}_1^3[n] = t[n]. \tag{7}$$

b) Tight frame filter banks: If the following inequality holds

$$\alpha[n] = \left|h_0^0[n]\right|^2 + \left|h_1^0[n]\right|^2 > 1, \quad n \in \mathbb{Z}, \tag{8}$$

then, due to the symmetry of the rational functions $\hat{h}_0^0[n]$ and $\hat{h}_0^0[n]$, the sequence $t[n]$ is strictly positive rational function of $\cos 2\pi n/N)$. Due to Riesz Lemma , it can be factorized $t[n] = T[n]\,T[-n]$, where T is a rational function of ω^n, which does not have roots for $n \in \mathbb{Z}$. Thus, we define

$$\hat{h}_0^2[n] = \hat{\tilde{h}}_0^2[n] = \hat{h}_1^3[-n] = \hat{\tilde{h}}_1^3[-n] = T[n]. \tag{9}$$

The PR filter bank, whose filters are

$$\begin{aligned} \hat{h}^0[n] &= \hat{h}_0^0[n] + \omega^{-n}\hat{h}_0^0[n] & \hat{h}^2[n] &= T[n], \\ \hat{h}^1[n] &= -\hat{h}_1^0[-n] + \omega^{-n}\hat{h}_0^0[-n], & \hat{h}^3[n] &= \omega^{-n}T[-n], \end{aligned}$$

generates a tight wavelet frame in the space $\Pi[N]$. Certainly, the symmetry of the FR $\hat{h}^0[n]$ does not guarantee the symmetry of the FR $\hat{h}^2[n]$ and $\hat{h}^3[n]$.

c) Semi-tight frame filter banks: If the condition Eq. (8) is not fulfilled then the sequence $t[n]$ can be factorized as $t[n] = T[n]\,\tilde{T}[-n]$, where $T[n] \neq \tilde{T}[n]$. Thus, we obtain the PR filter bank, whose filters are

$$\begin{aligned} \hat{h}^0[n] &= \hat{h}_0^0[n] + \omega^{-n}\hat{h}_0^0[n], & \hat{h}^1[n] &= -\hat{h}_1^0[-n] + \omega^{-n}\hat{h}_0^0[-n], \\ \hat{h}^2[n] &= T^2[n], & \hat{\tilde{h}}^2[n] &= \tilde{T}^2[n], \\ \hat{h}^3[n] &= \omega^{-n}T^3[n], & \hat{\tilde{h}}^3[n] &= \omega^{-n}\tilde{T}^3[n], \end{aligned} \tag{10}$$

where $T^2[n]\,\tilde{T}^2[-n] = T[n]\,\tilde{T}^3[-n] = t[n]$. The PR filter bank defined by Eq. (10) generates a semi-tight frame in the space $\Pi[N]$.

Remark 1: Since the rational function $t[n]$ of ω^n is symmetric about the change $n \to -n$, then it can be factorized into product of two symmetric rational functions $T[n]$ and $\tilde{T}[-n]$. An additional advantage of the semi-tight design is the option to swap approximation properties between the analysis and the synthesis framelets.

As usual, a multiscale frame transform is implemented by subsequent application of the frame transform to the low-frequency array of the transform coefficients.

IV. Spline p-filters

It was described above how to design a tight or a semi-tight frame comprising four framelets starting from a low-pass p-filter. A variety of such p-filters can be derived from the theory of periodic splines ([6], for example). The p-filters possess useful properties such as linear phase, flat spectra and well localised impulse responses. The idea is to design an $N/2$-periodic spline $S^p(t)$ of order p on the grid $\{k\}$, which interpolates the even polyphase component $\mathbf{x}_0$ of a signal $\mathbf{x}$: $S^p(k) = x[2k]$. Then, in order to derive the spline's values at the intermediate points $s_1[k] \stackrel{\text{def}}{=} S^p(k+1/2)$, which, in a sense predict the odd polyphase component $\mathbf{x}_1$ of $\mathbf{x}$, the signal $\mathbf{x}_0$ should be filtered with some "prediction" p-filter: $\hat{s}_1[n] = f^p[n]\,\hat{x}_0[n]$. Then, the interpolating low-pass p-filter is defined as $\hat{h}^0[n] \stackrel{\text{def}}{=} \left(1 + \omega^{-n} f^p[n]\right)/\sqrt{2}$. The corresponding high-pass p-filter is $\hat{h}^1[n] \stackrel{\text{def}}{=} \omega^{-n}\left(1 - \omega^n f^p[-n]\right)/\sqrt{2}$. The filters $f^p[n]$ can be explicitly calculated for any order of a spline. For all the orders except for $p = 2$ (piece-wise linear spline) the p-filters have infinite impulse response. This fact does not complicate the implementation, which consists of application of the forward and inverse fast Fourier transforms and simple arithmetic operations. The finite impulse response (up to periodization) p-filters can be derived from quasi-interpolating splines.

Because the conventional notion of vanishing moments is not applicable to the periodic discrete-time setting, we use the notion of the local discrete vanishing moments (LDVM). Loosely speaking, a framelet has m LDVM if, being convolved with a signal containing fragments of sampled polynomials of degree $m-1$, it eliminates these fragments. If the FR of a p-filter comprises either the factor $\left(1-\omega^{2n}\right)^r$ or the factor $\sin^{2r}\pi n/N$ then the corresponding framelet has $2r$ LDVM.

We designed a diverse collection of tight and semi-tight frames originating from interpolating and quasi-interpolating splines of different orders. Below are two examples.

d) Example 1: quadratic interpolating spline: $p = 3$: The frequency response of low- and high-pass p-filters are

$$\begin{aligned} \hat{h}^0[n] &= \sqrt{2}\,\frac{\cos^4 \pi n/N}{\cos^4 \pi n/N + \sin^4 \pi n/N} \\ \hat{h}^1[n] &= \omega^{-n}\sqrt{2}\,\frac{\sin^4 \pi n/N}{\cos^4 \pi n/N + \sin^4 \pi n/N}. \end{aligned}$$

In this case a symmetric factorization of the matrix $Q[n]$ defined in Eq. (6) is possible. Therefore the p-filters $\mathbf{h}^2$ and $\mathbf{h}^3$, which complete the p-filters $\mathbf{h}^0$ and $\mathbf{h}^1$ to the PR filter bank, have linear phase. Their frequency responses are

$$\hat{h}^2[n] = \frac{1}{2\sqrt{2}}\,\frac{\sin^2 2\pi n/N}{\cos^4 \pi n/N + \sin^4 \pi n/N} = -\omega^n\,\hat{h}^3[n].$$

The high-frequency framelet $\psi^1 = \mathbf{h}^1$ has four LDVM, while the framelets $\psi^l = \mathbf{h}^l$, $l = 2, 3$, have two.

e) Example 2: quadratic quasi-interpolating spline: In the tight frame derived from this spline the p-filters $\mathbf{h}^2$ and $\mathbf{h}^3$ are non-symmetric. However, the semi-tight frame, where those p-filters are antisymmetric proved to be highly efficient in applications. The frequency response of low- and high-pass p-filters are

$$\begin{aligned} \hat{h}^0[n] &= \tfrac{1}{\sqrt{2}}\,\cos^4 \tfrac{\pi n}{N}\left(3 - \cos \tfrac{2\pi n}{N}\right) \\ \hat{h}^1[n] &= \tfrac{\omega^{-n}}{\sqrt{2}}\,\sin^4 \tfrac{\pi n}{N}\left(3 + \cos \tfrac{2\pi n}{N}\right). \end{aligned}$$

Denote

$$\begin{aligned} \tilde{T}[n]_1 &= \tfrac{\omega^{4n} - 3\omega^{2n} + 3 - \omega^{-2n}}{8} \\ T[n]_1 &= \tfrac{\left(1-\omega^{2n}\right)\left(-\omega^{4n} - 12\omega^{2n} + 346 - 12\omega^{-2n} - \omega^{-4n}\right)}{1024}. \end{aligned}$$

Then,

$$\hat{h}^2[n] = T[n]_1 = -\frac{\hat{\tilde{h}}^3[-n]}{\omega^n},\ \hat{\tilde{h}}^2[n] = \tilde{T}[n]_1 = -\frac{\hat{h}^3[-n]}{\omega^n}.$$

Fig. 1. Top:FB generating tight frames: impulse and magnitude responses derived from quadratic quasi-interpolating spline. Bottom: analysis and synthesis band-pass filters for semi-tight frame

The high-frequency framelet ψ^1 has four LDVM. We assign three LDVM to the analysis framelet $\tilde{\psi}^2$ leaving only one LDVM to the synthesis framelet ψ^2 and vice versa for the framelets $\tilde{\psi}^3$ and ψ^3. Figure 1 displays impulse and magnitude responses of the p-filters derived from quadratic quasi-interpolating spline, which generate the tight and the semi-tight frames. We observe that the impulse responses of the band-pass p-filters for the tight frame are non-symmetric. The semi-tight frame band-pass p-filters have anti-symmetric impulse responses.

V. Application to Image Restoration

We designed a diverse library of tight and semi-tight frames, which were extended to two dimensions via the tensor products of 1D framelets. The frames were tested in multiple image restoration experiments where images were blurred, affected by random noise and a significant number of pixels were missing. For restoration, the SBI scheme [1] with the designed frames was utilized. Performance of different frames was compared. These experiments as well as the frame design are described in details in [7]. In most cases semi-tight frames were advantageous over respective tight frames. Especially successful was the semi-tight frame presented in Example 2. However in some experiments, the frames derived from higher order splines (thus having a big number of LDVM) outperformed the frame derived from low-order splines.

f) "Boats" and "Fingerprint" images: The "Boats" image was blurred by the motion kernel and its PSNR becomes 22.88 dB. Then, 70% of pixels were randomly removed. This reduces the PSNR to 7.37 dB. The image restored using the semi-tight frame from Example 2 with PSNR =30.28 dB. The "Fingerprint" image was affected by a strong zero-mean white noise with STD $\sigma = 20$ after being blurred by the Gaussian kernel (PSNR=19.75 dB). Then, 50% of its pixels were randomly removed and this produced PSNR=9.05 dB. The frame decomposition is implemented down to the fifth level. The PSNR=23.75 dB result was achieved by the application of the tight frame derived from the interpolating spline of fifth order. In this frame, the high-frequency framelet ψ^1 has six LDVM, while either of the framelets ψ^2 and ψ^3 has three LDVM. Results of the above experiments are displayed in Fig. 2. We observe that despite a strong degradation, which made images almost undistinguishable, they are successfully restored.

g) Restoration experiments for the "Window" image: This image was taken from [8]. The image was blurred by the motion kernel. In one experiment random noise presented (PSNR= 23.56 dB), while in the other white noise with STD $\sigma = 5$ was added (PSNR= 23.19 dB). Then, 30% of pixels were randomly removed. This reduces the PSNR to 10.22 dB and 10.20 dB, respectively. We compared the restoration results with the respective results reported in [8]. In the no-noise experiment the PSNR of the image restored by the semi-tight frame described in Example 2 was 43.78 dB versus 40.25 dB in [8]. For the noise experiment the PSNR was 28.81 dB versus 27. 76 dB in [8]. The restoration results are displayed in Fig. 3.

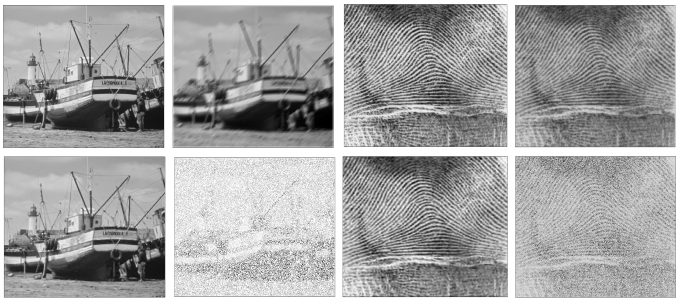


Fig. 2. "Boats" and "Fingerprint" images Top left in each quadruple – original images. Bottom left – restored images

Fig. 3. "Window" image. Left quadruple – no-noise experiment. Right quadruple – noise experiment

References

[1] T. Goldstein and S. Osher, "The split Bregman method for $L1$-regularized problems," *SIAM J. Imaging Sci.*, vol. 2, no. 2, pp. 323–343, 2009.

[2] J. Cai, S. Osher, and Z. Shen, "Split Bregman methods and frame based image restoration," *Multiscale Model. Simul.*, vol. 8, no. 2, pp. 337–369, 2009/10.

[3] J. Cai, B. Dong, S. Osher, and Z. Shen, "Image restoration: total variation, wavelet frames, and beyond," *J. Amer. Math. Soc.*, vol. 25, no. 4, pp. 1033–1089, 2012.

[4] Z. Shen, "Wavelet frames and image restorations," in *Proceedings of the International Congress of Mathematicians, Vol. IV*, R. Bhatia, Ed. New Delhi: Hindustan Book Agency, 2010, pp. 2834–2863.

[5] Z. Cvetković and M. Vetterli, "Oversampled filter banks," *IEEE Trans. Signal Process.*, vol. 46, no. 5, pp. 1245–1255, 1998.

[6] A. Averbuch and V. Zheludev, "Construction of biorthogonal discrete wavelet transforms using interpolatory splines," *Appl. Comput. Harmon. Anal.*, vol. 12, no. 1, pp. 25–56, 2002.

[7] A. Averbuch, P. Neittaanmäki, and V. Zheludev, "Spline-based frames in the space of periodic signals," *submitted*, 2012.

[8] H. Ji, Z. Shen, and Y. Xu, "Wavelet based restoration of images with missing or damaged pixels," *East Asian J. Appl. Math.*, vol. 1, no. 2, pp. 108–131, 2011.

On the Noise-Resilience of OMP with BASC-Based Low Coherence Sensing Matrices

Henning Zörlein, Dejan E. Lazich and Martin Bossert
Institute of Communications Engineering
Ulm University
89081 Ulm, Germany
{henning.zoerlein, dejan.lazich, martin.bossert}@uni-ulm.de

***Abstract*—In Compressed Sensing (CS), measurements of a sparse vector are obtained by applying a sensing matrix. With the means of CS, it is possible to reconstruct the sparse vector from a small number of such measurements. In order to provide reliable reconstruction also for less sparse vectors, sensing matrices are desired to be of low coherence. Motivated by this requirement, it was recently shown that low coherence sensing matrices can be obtained by Best Antipodal Spherical Codes (BASC) [1]. In this paper, the noise-resilience of the Orthogonal Matching Pursuit (OMP) used in combination with low coherence BASC-based sensing matrices is investigated.**

I. Introduction

In Compressed Sensing (CS), one is particularly interested in the sparsest solution to an underdetermined system of $\mathcal{M}$ linear equations:

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}.$$

This is commonly interpreted as acquiring a sufficiently k-sparse vector $\boldsymbol{x} \in \mathbb{R}^{\mathcal{N}}$ from a small number of measurements. A so called sensing matrix $\boldsymbol{A} \in \mathbb{R}^{\mathcal{M}\times\mathcal{N}}$ describes these measurements enlisted in $\boldsymbol{b} \in \mathbb{R}^{\mathcal{M}}$, where $\mathcal{M}$ is significantly smaller than $\mathcal{N}$.

However, for practical applications, measurement noise will always be present. Therefore, an additional noise term $\boldsymbol{n} \in \mathbb{R}^{\mathcal{M}}$ consisting of Gaussian distributed elements with zero mean is usually considered in the system model (e.g. [2]):

$$\boldsymbol{A}\boldsymbol{x} + \boldsymbol{n} = \boldsymbol{b} + \boldsymbol{n}.$$

There are multiple approaches to reconstruct the sparse vector $\hat{\boldsymbol{x}}$ out of its measurements $\boldsymbol{b}$, e.g. the Basis Pursuit (BP) and Basis Pursuit De-Noising (BPDN) algorithms [3] based on convex relaxation, or greedy algorithms like the Orthogonal Matching Pursuit (OMP) [4].

The selection of suitable sensing matrices $\boldsymbol{A}$ is crucial for a successful reconstruction. There are multiple properties providing conditions on sensing matrices, e.g. the worst-case coherence μ between columns of the sensing matrix [5]–[9]. The worst-case coherence is defined by the maximal absolute value of the inner product between two distinct columns of $\boldsymbol{A}$:

$$\mu = \max_{i\neq j} |\boldsymbol{a}_i \cdot \boldsymbol{a}_j|, \tag{1}$$

where $\boldsymbol{a}_i$ is the ith column. Motivated by these coherence properties, the construction of Best Antipodal Spherical Codes (BASC)-based sensing matrices with low worst-case coherence has been proposed in [1].

Other approaches for guarantees on successful reconstruction utilize the Restricted Isometry Property (RIP) [10]. Normalized Gaussian random matrices are often used as sensing matrices, because they fulfill the RIP with high probability [10]. Due to its combinatorial nature, the direct evaluation of a matrix for its RIP is not tractable. However, Monte Carlo experiments can be performed, which indicate the suitability of BASC-based sensing matrices with respect to the RIP [1].

The reason for using BASC-based sensing matrices and their construction is briefly summarized in Section II as given in [1]. An analysis on the noise-resilience based on numerical simulations is given in Section III. In Section IV, conclusions are provided as well.

II. BASC-based Sensing Matrices

A. Spherical Codes

Any finite set of M points placed on the surface of the N-dimensional unit sphere centered at the origin of the N-dimensional Euclidean space $\mathbb{R}^N$ is called a spherical code and denoted by $C_s(N, M)$ [11]–[13]. A point of $C_s(N, M) = \{\boldsymbol{s}_m\}_{m=1}^{M}$ is determined by its corresponding code word $\boldsymbol{s}_m = (s_{m1}, \ldots, s_{mn}, \ldots, s_{mN})$ representing a unit position vector $(|\boldsymbol{s}_m| = 1, m = 1, \ldots, M)$ whose components $s_{mn} \in \mathbb{R}$ are the coordinates of the point in some reference Cartesian coordinate system centered at the origin. Best Spherical Codes (BSC), $C_{\mathrm{bs}}(N, M)$, are spherical codes which maximize the minimal Euclidean (or angular) distance $d_{ml} = |\boldsymbol{s}_m - \boldsymbol{s}_l|$ between any two points (or equivalently, minimize the maximal inner product of the corresponding code words). All rotations of a BSC are usually regarded as the same, therefore, a BSC is characterized only by its distance distribution $\mathfrak{D} = \{d_{ml}\}_{m<l}$. For some specific (N, M) pairs $(N > 2, M > N)$, the corresponding BSC can be unique, however, there also exist (N, M) pairs with more than one corresponding BSC (these BSCs have different distance distributions but the same minimal distance). For BASC, $C_{\mathrm{bas}}(N, M)$, the antipodal of each code word is also a code word:

$$\boldsymbol{s}_m \in C_{\mathrm{bas}}(N, M) \iff -\boldsymbol{s}_m \in C_{\mathrm{bas}}(N, M).$$

This property can be used in order to construct low coherence sensing matrices [1].

B. Sensing Matrices Based on Spherical Codes

The $M = \mathcal{N}$ code words of a spherical code $C_s(N, M)$ can be regarded as $\mathcal{N}$ columns of a sensing matrix $\boldsymbol{A} \in \mathbb{R}^{\mathcal{M}\times\mathcal{N}}$. It can be easily shown that the squared Euclidean distance between code words is proportional to their inner product. However, the worst-case coherence is defined over the absolute value of the inner product, see Equation (1). Consequently, the search for sensing matrices with smallest worst-case coherence μ is transformed into the search for BASC with $N = \mathcal{M}$ and $M = 2\mathcal{N}$. The $\frac{M}{2}$ non-antipodal code words of $C_{bas}(N, M)$ are the columns of the desired sensing matrix.

C. Obtaining BSC

The points of spherical codes can be considered as M charged particles on the unit sphere acting in some field of repelling forces [14]. Starting from any initial position, such particles will move until the total potential energy of the system approaches some local minimum. In any one of these local minima the particles will settle causing a stable or unstable equilibrium of mutual repelling forces. In [15], such a generalized potential function, $g(\mathfrak{D})$, was introduced. For a specific form of $g(\mathfrak{D})$ given in [16] by

$$g(\mathfrak{D}) = \sum_{m<l} |\boldsymbol{s}_m - \boldsymbol{s}_l|^{-(\nu-2)}, \tag{2}$$

where $\nu \in \mathbb{N}$ $(\nu > 2)$, it was shown that the global minimum of $g(\mathfrak{D})$ is attained by a BSC if $\nu \to \infty$.

As further summarized in [1], the set of fixed points of two mappings can be regarded as the desired minima.

The first mapping $\boldsymbol{F}$ can be interpreted as collection of effective forces $\boldsymbol{f}_m$ acting on the code words $\boldsymbol{s}_m$ of a spherical code and is given by

$$\begin{aligned}\boldsymbol{F}[C_s(N,M)] &= \left\{\underline{\boldsymbol{f}}_m(C_s(N,M))\right\}_{m=1}^{M} \\ &= \left\{\boldsymbol{f}_m = \frac{\sum\limits_{l\neq m}\left[(\boldsymbol{s}_m - \boldsymbol{s}_l)/|\boldsymbol{s}_m - \boldsymbol{s}_l|^{\nu}\right]}{\left|\sum\limits_{l\neq m}\left[(\boldsymbol{s}_m - \boldsymbol{s}_l)/|\boldsymbol{s}_m - \boldsymbol{s}_l|^{\nu}\right]\right|}\right\}_{m=1}^{M}\end{aligned}$$

or, with the underlined denotation of unit vectors $\underline{\boldsymbol{u}} = \frac{\boldsymbol{u}}{|\boldsymbol{u}|}$, by

$$\left\{\underline{\boldsymbol{f}}_m = \underbrace{\sum_{l\neq m}\frac{\underline{\boldsymbol{s}}_m - \underline{\boldsymbol{s}}_l}{|\underline{\boldsymbol{s}}_m - \underline{\boldsymbol{s}}_l|^{\nu}}}_{=\sum\limits_{l\neq m}\boldsymbol{\delta}_{ml}}\right\}_{m=1}^{M}. \tag{3}$$

With the help of $\boldsymbol{F}$ a second mapping can be defined by

$$\boldsymbol{\Phi}[C_s(N,M)] = \left\{\underline{\boldsymbol{s}}_m + \alpha\underline{\boldsymbol{f}}_m\right\}_{m=1}^{M}, \tag{4}$$

where $\underline{\boldsymbol{f}}_m$ is given by (3) and $\alpha \in \mathbb{R}$. It is evident that the mappings $\boldsymbol{F}$ and $\boldsymbol{\Phi}$ have the same set of fixed points. For a small enough "damping factor" α, the iterative process

$$C_s(N,M)^{(k+1)} = \boldsymbol{\Phi}(C_s(N,M)^{(k)});\ k = 0, 1, \ldots \tag{5}$$

converges to one of the fixed points of the function $\boldsymbol{\Phi}$, and consequently of $\boldsymbol{F}$.

It was also shown [17] that, generally, for ν large enough, all fixed points correspond to spherical codes whose minimal distances are close enough to the minimal distance of corresponding BSCs. Consequently by finding any fixed point using (5) with ν large enough, the corresponding spherical code will be very close to the best one.

D. Obtaining BASC

The construction of BSC can be easily adapted for BASC, by considering additional antipodal points [1], leading to a new mapping and new forces acting on the particles respectively:

$$\left\{\underline{\boldsymbol{f}}_m = \underbrace{\sum_{l\neq m}\left[\frac{\underline{\boldsymbol{s}}_m - \underline{\boldsymbol{s}}_l}{|\underline{\boldsymbol{s}}_m - \underline{\boldsymbol{s}}_l|^{\nu}} + \frac{\underline{\boldsymbol{s}}_m + \underline{\boldsymbol{s}}_l}{|\underline{\boldsymbol{s}}_m + \underline{\boldsymbol{s}}_l|^{\nu}}\right]}_{}\right\}_{m=1}^{M}. \tag{6}$$

After the mapping (4) is applied, the antipodal points need to be updated. The resulting algorithm is given in Fig. 1.

III. Noise-Resilience Determined by Numerical Evaluations

The frequency of successful reconstruction[1] is evaluated over the sparsity of $\boldsymbol{x}$, where the non-zero values are drawn from a Gaussian distribution with zero mean and unit variance, and over the Signal-to-Noise-(power)-Ratio SNR, with

$$\mathsf{SNR}\ [\mathsf{dB}] = 10 \cdot \log_{10}\left(\frac{\sum\limits_{i=1}^{\mathcal{M}}|b_i|^2}{\sum\limits_{i=1}^{\mathcal{M}}|n_i|^2}\right),$$

where b_i and n_i are the components of the corresponding vectors $\boldsymbol{b}$ and $\boldsymbol{n}$. For the construction of BASC-based matrices, we used the initial values as given in the algorithm description presented in Fig. 1. For the stopping criterion of the OMP algorithm, we assume knowledge of the noise power: If the ℓ_2-norm of the residual is less than the ℓ_2-norm of the noise plus some small threshold $\left(10^{-6}\right)$, the OMP algorithm will stop. Our simulations indicate that OMP also performs well for an overestimation of the noise power, therefore, the assumption of a known noise power is not too restrictive. All simulations have been performed in MATLAB® [18].

Column-normalized random matrices with entries drawn from a Gaussian distribution have also been considered for comparisons. For the numerical evaluation, a version of each matrix type has been computed. The frequency of successful reconstruction has been determined over 7500 simulations. The corresponding result is shown in Fig. 2 and Fig. 3 for matrices of size 64×128. We repeated such simulations for multiple different realizations of the discussed matrices, however, the results did not show significant differences.

As it can be seen in Fig. 2 and Fig. 3, the signal must be strong enough in order to allow sparse recovery by the OMP. For high SNR levels, the sparsity is the dominating factor,

[1]The reconstruction is considered to be successful, if the condition $|\hat{\boldsymbol{x}} - \boldsymbol{x}| < 10^{-3}$ is fulfilled.

procedure BASC-BASED SENSING MATRIX($\mathcal{M}, \mathcal{N}$)
  $N \leftarrow \mathcal{M}$
  $M \leftarrow 2\mathcal{N}$
  $\alpha_{init} \leftarrow 0.9$
  $\nu \leftarrow 2$
  $\nu_{max} \leftarrow 2^{10}$
  $i_{max} \leftarrow 10^5$
  $\epsilon \leftarrow 10^{-10}$
  $C_s \leftarrow$ arbitrary ▷ Random spherical code
  $C_{as} \leftarrow [C_s \quad -C_s]$ ▷ Antipodal spherical code
  $\alpha \leftarrow \alpha_{init}$
  while $\nu < \nu_{max}$ **do**
    FixedPoint $\leftarrow$ **false**
    $i \leftarrow 0$
    while $i < i_{max}$ **AND** FixedPoint $=$ **false do**
      for $m = 1$ to $\frac{M}{2}$ **do**
        $\boldsymbol{f}_m \leftarrow \mathbf{0}$
        for $l = 1$ to M **do**
          if $l \neq m$ **AND** $l \neq m + \mathcal{N}$ **then**
            $\boldsymbol{f}_m \leftarrow \boldsymbol{f}_m + \frac{\underline{\boldsymbol{s}}_m - \underline{\boldsymbol{s}}_l}{\left|\underline{\boldsymbol{s}}_m - \underline{\boldsymbol{s}}_l\right|^\nu}$
          end if
        end for
      end for
      $\{\boldsymbol{s}_m\}_{m=1}^{\frac{M}{2}} \leftarrow \left\{\underline{\boldsymbol{s}}_m + \alpha \underline{\boldsymbol{f}}_m\right\}_{m=1}^{\frac{M}{2}}$
      $\{\boldsymbol{s}_m\}_{m=1+\frac{M}{2}}^{M} \leftarrow \{-\underline{\boldsymbol{s}}_m\}_{m=1}^{\frac{M}{2}}$
      if all $\left|\underline{\boldsymbol{f}}_m - \underline{\boldsymbol{s}}_m\right| < \epsilon$ **then**
        FixedPoint $\leftarrow$ `true`
      end if
      $i \leftarrow i + 1$
    end while
    $\nu \leftarrow 2\nu$
    $\alpha \leftarrow \frac{\alpha_{init}}{\nu - 1}$
  end while
  return $A \leftarrow \{\underline{\boldsymbol{s}}_m\}_{m=1}^{\frac{M}{2}}$
end procedure

Fig. 1. The construction algorithm for BASC-based sensing matrices.

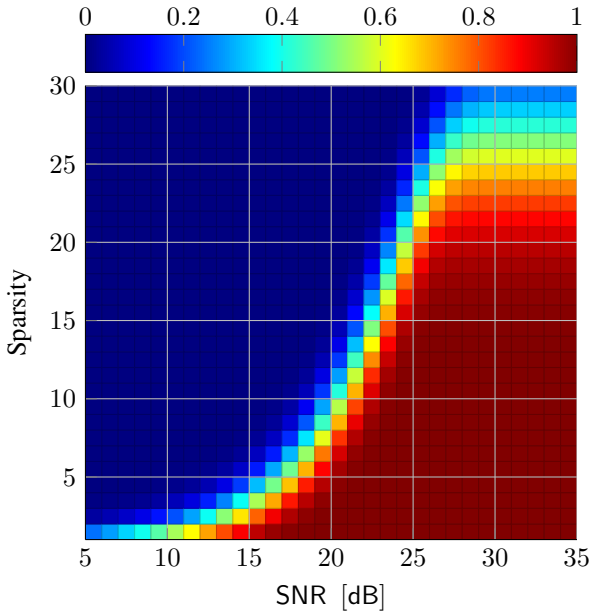


Fig. 2. Frequency of exact reconstruction for normalized Gaussian matrices with $\mathcal{M} = 64$ and $\mathcal{N} = 128$.

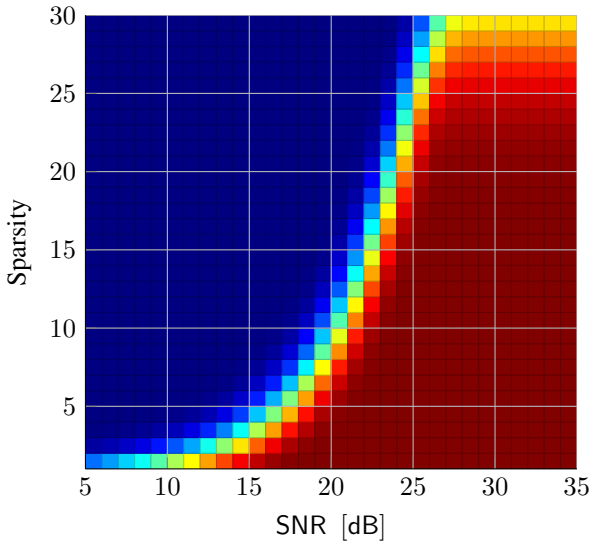


Fig. 3. Frequency of exact reconstruction for BASC-based matrices with $\mathcal{M} = 64$ and $\mathcal{N} = 128$.

and it can be seen that the OMP performs better for BASC-based matrices. Comparing the presented results with those of [1], where a noise-free setting was investigated with a BP algorithm, it is obvious that OMP gains more from the low coherence BASC-based matrices than the BP algorithm.

In Fig. 4, the difference of the reconstruction frequencies is shown in order to give a clearer comparison. Green areas indicate that the OMP algorithm was more often successful with the BASC-based matrix then with the normalized Gaussian matrix. Red areas would indicate better results in favor of the Gaussian matrices. Obviously, BASC-based matrices work on average better with the OMP algorithm. However, it should also be noted that this superiority is not always observable. For certain individual realizations of the noise $\boldsymbol{n}$ and the sparse vector $\boldsymbol{x}$, the Gaussian matrices performed slightly better for low SNR levels ($10 - 20$db). Taking more simulations into account, these differences average out cf. Fig. 4.

IV. CONCLUSIONS

The OMP algorithm clearly benefits more from the low coherence of BASC-based sensing matrices than the BP algorithm (cf. [1]).

For higher SNR levels, the lower coherence between the columns of the BASC-based sensing matrices can be exploited by the OMP, and therefore, a better performance can be achieved in such situations.

However, no significant gain in performance can be expected for lower SNR regions.

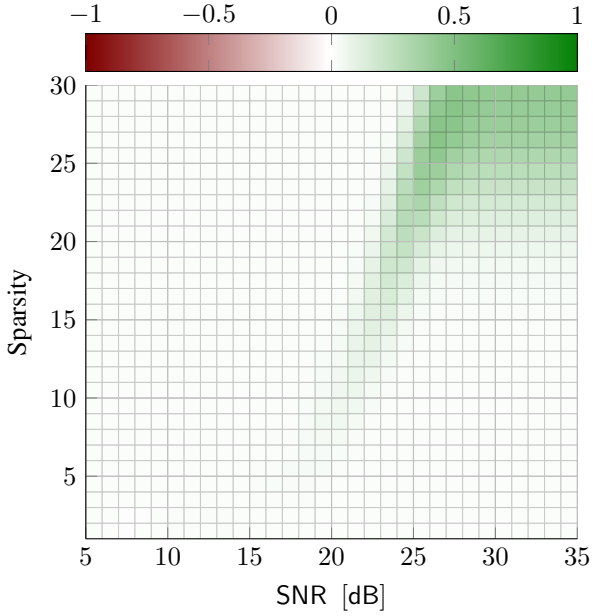


Fig. 4. Difference of the frequency of exact reconstruction between BASC-based matrices and normalized Gaussian matrices. Both are of size $\mathcal{M} = 64$ and $\mathcal{N} = 128$. Green areas indicate better performance of BASC-based matrices, whilst red areas show the same for Gaussian matrices.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their insightful and constructive comments, and their students Faisal Akram and Kotchapol Yoosooksawat for the assisting simulations.

This work was supported by the German research council Deutsche Forschungsgemeinschaft (DFG) under Grant Bo 867/27-1.

REFERENCES

[1] D. E. Lazich, H. Zörlein, and M. Bossert, "Low coherence sensing matrices based on best spherical codes," in *9th International ITG Conference on Systems, Communications and Coding 2013 (SCC'2013)*, Munich, Germany, Jan. 2013.

[2] T. Cai and L. Wang, "Orthogonal matching pursuit for sparse signal recovery with noise," *IEEE Transactions on Information Theory*, vol. 57, no. 7, pp. 4680 – 4688, Jul. 2011.

[3] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific Computing*, vol. 20, pp. 33 – 61, 1998.

[4] Y. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Proceedings of the 27 th Annual Asilomar Conference on Signals, Systems, and Computers*, 1993, pp. 40 – 44.

[5] D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 5, pp. 2197 – 2202, 2003.

[6] M. Elad and A. M. Bruckstein, "A generalized uncertainty principle and sparse representation in pairs of bases," *IEEE Transactions on Information Theory*, vol. 48, no. 9, pp. 2558 – 2567, Sep. 2002.

[7] R. Gribonval and M. Nielsen, "Sparse representations in unions of bases," *IEEE Transactions on Information Theory*, vol. 49, no. 12, pp. 3320 – 3325, Dec. 2003.

[8] W. Bajwa, R. Calderbank, and S. Jafarpour, "Model selection: Two fundamental measures of coherence and their algorithmic significance," in *Information Theory Proceedings (ISIT), 2010 IEEE International Symposium on*, Jun. 2010, pp. 1568 – 1572.

[9] W. U. Bajwa, R. Calderbank, and D. G. Mixon, "Two are better than one: Fundamental parameters of frame coherence," *Applied and Computational Harmonic Analysis*, vol. 33, no. 1, pp. 58 – 78, 2012.

[10] E. J. Candès and T. Tao, "Decoding by Linear Programming," *IEEE Transactions on Information Theory*, vol. 51, no. 12, pp. 4203 – 4215, Dec. 2005.

[11] J. Conway and N. Sloane, *Sphere Packings, Lattices, and Groups*, 3rd ed., ser. Grundlehren der Mathematischen Wissenschaften. Springer, 1999.

[12] T. Ericson and V. Zinoviev, *Codes On Euclidean Spheres*, ser. North-Holland Mathematical Library. Elsevier, 2001.

[13] N. J. A. Sloane, "Spherical codes: Nice arrangements of points on a sphere in various dimensions." [Online]. Available: http://www2.research.att.com/ njas/packings/

[14] J. Leech, "Equilibrium of sets of particles on a sphere," *The Mathematical Gazette*, vol. 41, no. 336, pp. 81 – 90, 1957.

[15] D. E. Lazic, "Class of block codes for the gaussian channel," *Electronics Letters*, vol. 16, no. 5, pp. 185 – 186, Feb. 1980.

[16] D. E. Lazic, T. Bece, and P. J. Krstajic, "On the construction of the best spherical code by computing the fixed point," in *IEEE International Symposium on Information Theory, Abstract of papers*, Ann Arbor, Michigan, USA, 1986, p. 74.

[17] D. E. Lazic, V. Senk, and R. Zamurovic, "An efficient numerical procedure for generating best spherical arrangements of points," in *Proceedings of the International AMSE'88, Istanbul Conference "Modelling and Simulation"*, vol. 1C, Istanbul, Turkey, 1988, pp. 267 – 278.

[18] MathWorks®, "MATLAB® Version R2012b," Natick, Massachusetts. [Online]. Available: http://www.mathworks.com/

Tight frames in spiral sampling

Enrico Au-Yeung
Pacific Institute for the Mathematical Sciences
4176-2207 Main Mall, Vancouver
British Columbia, V6T 1Z4, Canada
Email: enricoauy@math.ubc.ca

Somantika Datta
Department of Mathematics
University of Idaho
Moscow, ID 83844-1103, USA
Email: sdatta@uidaho.edu

***Abstract*—The paper deals with the construction of Parseval frames for $L^2(B(0,R))$, the space of square integrable functions whose domain is the ball of radius R. The focus is on Fourier frames on a spiral. Starting with a Fourier frame on a spiral, a Parseval frame that spans the same space can then be obtained by a symmetric approximation of the original Fourier frame.**

I. Introduction

Earlier work by Benedetto et al. [1], [2], [3], [4] gave the construction of a set of points on a given spiral such that these points give rise to a frame for $L^2(B(0,R))$, the space of all square integrable functions on the ball centered at the origin and of radius R. This means that given a spiral A_c, the authors in [1], [2], [3], [4] were able to construct a sequence of points Λ on this spiral and its interleaves such that every signal f belonging to $L^2(B(0,R))$ can be written as $\sum_{\lambda\in\Lambda} a_\lambda(f)e_\lambda$ where $e_\lambda(x) = e^{2\pi i x\cdot\lambda}$. The incentive of choosing points on a spiral comes from the applicability in MRI (Magnetic Resonance Imaging) where a signal is sampled in the Fourier domain along interleaving spirals, resulting in fast imaging methods. For practical purposes, the reconstruction of signals using such infinite frames entails inverting the frame operator and/or using only finitely many samples. Such numerical issues are mitigated if one can use a tight frame. The possibility of expanding a function as a non-harmonic Fourier series was discovered by Paley and Wiener. For a sequence Λ of real numbers, it is natural to ask whether every band-limited signal with spectrum E can be reconstructed in a stable way from its samples $\{F(\lambda), \lambda \subseteq \Lambda\}$. Landau [5] proved a necessary condition for $\{e^{2\pi i x\cdot\lambda}, \lambda \in \Lambda\}$ to be a frame for the space of band-limited functions with spectrum E by relating the lower density of Λ to the measure of E. There is an extensive literature on the stable reconstruction problem, (see, e.g., [6], [7], [8], [9], [10], [11]). Many of the contributions to this area focus on the theoretical aspect, while our emphasis is on explicit construction.

The main contribution of this article is to give an explicit procedure to convert a frame which is not a tight frame into a Parseval frame, with the requirement that each element in the resulting Parseval frame can be expressed as a linear combination of the elements in the original frame. To be precise, this requirement means that if $\{f_1, f_2, f_3\}$ is the original frame for the Hilbert space $\mathcal{H}$, and $\{g_1, g_2, g_3\}$ is the resulting Parseval frame, then each g_n is a linear combination of f_1, f_2, f_3. For any function $f \in \mathcal{H}$, one has $f = \sum_{n=1}^{3}\langle f, g_n\rangle g_n$. Since each g_n is a linear combination of f_1, f_2, and f_3, each number $\langle f, g_n\rangle$ can be calculated from the three numbers $\langle f, f_1\rangle, \langle f, f_3\rangle, \langle f, f_3\rangle$. Hence, from the numbers $\langle f, f_n\rangle$ for $n = 1, 2, 3$, one can recover f. In the reconstruction formula using the Parseval frame, only the measurements obtained from the original frame are needed. This feature is extremely important, especially in the aforementioned application to MRI, when the measurements from the original frame are the only available measurements. The procedure explained in this article applies to other frames, and not just to Fourier frames, but motivated by applications to medical imaging as in MRI, the focus here is only on spiral sampling with Fourier frames.

In [12], Frank, Paulsen, and Tiballi obtain a Parseval frame from a given frame that spans the same subspace as the original frame and is closest to it in some sense, which they call *symmetric approximation*. The approach used in [12] is to use the polar decomposition of the synthesis operator of the original frame. This idea inspires the method developed in the present work to obtain Parseval frames for the spiral sampling case. Presently, the work is only focused on finite frames. The symmetric approximation of infinite Fourier frames on spirals and the best N-term approximation of such frames constitute ongoing research.

A. Notation and preliminaries

Let $\mathbb{R}^d$ be the d-dimensional Euclidean space, and let $\widehat{\mathbb{R}}^d$ denote $\mathbb{R}^d$ when it is considered as the domain of the Fourier transforms of signals defined on $\mathbb{R}^d$. $L^2(\widehat{\mathbb{R}}^d)$ is the space of square integrable functions ϕ on $\widehat{\mathbb{R}}^d$, i.e.,

$$||\phi||_{L^2(\widehat{\mathbb{R}}^d)} = \left(\int_{\widehat{\mathbb{R}}^d} |\phi(\gamma)|^2 \mathrm{d}\gamma\right)^{1/2} < \infty,$$

$\phi^\vee$ is the inverse Fourier transform of ϕ defined as

$$\phi^\vee(x) = \int_{\widehat{\mathbb{R}}^d} \phi(\gamma) e^{2\pi i x\cdot\gamma}\mathrm{d}\gamma,$$

and $supp\ \phi^\vee$ denotes the support of $\phi^\vee$. Let $E \subseteq \widehat{\mathbb{R}}^d$ be closed. The *Paley-Wiener* space PW_E is

$$PW_E = \{\phi \in L^2(\widehat{\mathbb{R}}^d) : supp\ \phi^\vee \subseteq E\}.$$

Let $\mathcal{H}$ be a separable Hilbert space. A sequence $\{f_n : n \in \mathbb{Z}^d\} \subseteq \mathcal{H}$ is a *frame* for $\mathcal{H}$ if there exist constants $0 < A \leq$

$B < \infty$ such that

$$\forall y \in \mathcal{H}, \quad A||y||^2 \leq \sum_n |\langle y, f_n\rangle|^2 \leq B||y||^2.$$

The constants A and B are called the lower and upper frame bounds, respectively. If $A = B$, the frame is said to be *tight* and if $A = B = 1$, the frame is called a *Parseval* frame. Let $\{f_n\}$ be a frame for $\mathcal{H}$. The *synthesis operator* is the linear mapping $T : \ell_2 \to \mathcal{H}$ given by $T(\{c_i\}) = \sum_k c_k f_k$. The frame operator $S : \mathcal{H} \to \mathcal{H}$ is TT^* and is given by

$$\forall y \in \mathcal{H}, \quad S(y) = \sum_n \langle y, f_n\rangle f_n.$$

For every $y \in \mathcal{H}$,

$$y = \sum_n \langle y, S^{-1}f_n\rangle f_n = \sum_n \langle y, f_n\rangle S^{-1}f_n.$$

For more on frames one can look at [13] or [14].

Let $\Lambda \subseteq \widehat{\mathbb{R}}^d$ be a sequence and let $E \subset \mathbb{R}^d$ have finite Lebesgue measure. By the Parseval Formula, the following are equivalent ([3], [4]).

(i) $\{e_\lambda : \lambda \in \Lambda\}$ is a frame for $L^2(E)$.

(ii) There exist $0 < A \leqslant B < \infty$ such that

$$A||\phi||_2^2 \leqslant \sum_{\lambda\in\Lambda} |\phi(\lambda)|^2 \leqslant B||\phi||_2^2,$$

for all ϕ in PW_E. In this case, we say that Λ is a *Fourier frame* for PW_E.

A set Λ is *uniformly discrete* if there exists $r > 0$ such that

$$\forall \lambda, \gamma \in \Lambda, \quad |\lambda - \gamma| \geq r,$$

where $|\lambda - \gamma|$ is the Euclidean distance between λ and γ.

If for two frames $\{f_i\}_{i\in\mathbb{N}}$ and $\{g_i\}_{i\in\mathbb{N}}$ of two Hilbert subspaces $\mathcal{K}$ and $\mathcal{L}$ of $\mathcal{H}$, respectively, there exists an invertible bounded linear operator $T : \mathcal{K} \to \mathcal{L}$ such that $T(f_i) = g_i$ for every index i, then these two frames are said to be *weakly similar* [12]. A Parseval frame $\{\nu_i\}_{i=1}^n$ in a finite dimensional Hilbert subspace $\mathcal{L} \subseteq \mathcal{H}$ is said to be a *symmetric approximation* of a finite frame $\{f_i\}_{i=1}^n$ in a Hilbert subspace $\mathcal{K} \subseteq \mathcal{H}$ if the frames $\{f_i\}_{i=1}^n$ and $\{\nu_i\}_{i=1}^n$ are weakly similar and the inequality

$$\sum_{j=1}^n \|\mu_j - f_j\|^2 \geq \sum_{j=1}^n \|\nu_j - f_j\|^2$$

is valid for all Parseval frames $\{\mu_i\}_{i=1}^n$ in Hilbert subspaces of $\mathcal{H}$ that are weakly similar to $\{f_i\}_{i=1}^n$ [12]. If $\mathcal{K} = \mathcal{L}$, the frames are called *similar*.

When a 3 by 3 matrix W is acting on a sequence of elements $\{f_1, f_2, f_3\}$, this action is denoted by $\{e_1, e_2, e_3\} = W \cdot \{f_1, f_2, f_3\}$, or in matrix notation,

$$\begin{bmatrix} e_1 \\ e_2 \\ e_3 \end{bmatrix} = \begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \\ w_{31} & w_{23} & w_{33} \end{bmatrix} \begin{bmatrix} f_1 \\ f_2 \\ f_3 \end{bmatrix},$$

to denote

$$e_1 = \sum_{j=1}^3 w_{1j}f_j, \ e_2 = \sum_{j=1}^3 w_{2j}f_j, \ e_3 = \sum_{j=1}^3 w_{3j}f_j.$$

B. Background

The following theorem [1], [2], [3], [4] is based on a deep result of Beurling [15].

Theorem I.1 (Beurling Covering Theorem)**.** *Let $\Lambda \subseteq \widehat{\mathbb{R}}^d$ be uniformly discrete, and define $\rho = \sup_{\mu\in\widehat{\mathbb{R}}^d} \operatorname{dist}(\mu, \Lambda)$ where $\operatorname{dist}(\mu, \Lambda)$ is the Euclidean distance between the point μ and the set Λ. If $R\rho < 1/4$, then Λ is a Fourier frame for $\mathrm{PW}_{B(0,R)}$.*

In [1], [2], [3], [4] the authors have used the Beurling Covering Theorem to give an explicit construction of Fourier frames from points that lie on a spiral. In particular, the following result can be found in [2].

Example I.2. *Fix $c > 0$. In $\widehat{\mathbb{R}}^2$, consider the spiral*

$$A_c = \{c\theta\cos 2\pi\theta, c\theta \sin 2\pi\theta : \theta \geq 0\}.$$

For R and δ satisfying $Rc < 1/2$ and $(\frac{c}{2} + \delta)R < 1/4$, one chooses a uniformly discrete set of points Λ such that the curve distance between any two consecutive points is less than 2δ, and beginning within 2δ of the origin. Then Λ satisfies the Beurling Covering Theorem and hence gives rise to a Fourier frame for $\mathrm{PW}_{B(0,R)}$.

The synthesis operator T defined in Section I-A is bounded and has a natural polar decomposition $T = W|T|$, where W is a partial isometry from ℓ_2 into $\mathcal{H}$. To obtain a symmetric approximation of a given frame, the following has been shown in [12].

Theorem I.3. *Let $\{\mu_i\}_{i=1}^n$ be a Parseval frame in a Hilbert subspace $\mathcal{L} \subseteq \mathcal{H}$ and let $\{f_i\}_{i=1}^n$ be a frame in a Hilbert subspace $\mathcal{K} \subseteq \mathcal{H}$ such that both these frames are weakly similar. Letting the standard orthonormal basis for $\mathbb{C}^n$ be denoted by $\{e_i\}_{i=1}^n$, the following inequality*

$$\sum_{j=1}^n \|\mu_j - f_j\|^2 \geq \sum_{j=1}^n \|W(e_j) - f_j\|^2$$

holds. Equality appears if and only if $\mu_j = W(e_j)$ for $j = 1, \ldots, n$. (Consequently, the symmetric approximation of a frame $\{f_i\}_{i=1}^n$ in a finite dimensional Hilbert space $\mathcal{K} \subseteq \mathcal{H}$ is a Parseval frame spanning the same Hilbert subspace $\mathcal{L} \equiv \mathcal{K}$ of $\mathcal{H}$ and being similar to $\{f_i\}_{i=1}^n$.)

Similar results for infinite frames in separable Hilbert spaces have also been established in [12] but for now the focus is on the finite dimensional case.

II. Parseval frames from a finite Fourier frame on a spiral

Three examples are discussed below. In the first two examples, the frame under consideration is on $\widehat{\mathbb{R}}$. The third example is for a Fourier frame on a spiral in $\widehat{\mathbb{R}}^2$.

In the first two examples, the procedure suggested by Theorem I.3 is modified so that in the final step, matrix multiplication is replaced by a matrix acting on a sequence of elements in a Hilbert space.

Example II.1. *Let* $\{f_1 = e^{2\pi i\lambda_1 x}, f_2 = e^{2\pi i\lambda_2 x}, f_3 = e^{2\pi i\lambda_3 x}\}$ *be a frame that spans a subspace of* $L^2([-1/2, 1/2])$. *Choose* $\lambda_1 = 3 + \frac{1}{3}, \lambda_2 = 4 + \frac{1}{4}, \lambda_3 = 5 + \frac{1}{5}$.

This frame is used to construct a Parseval frame that spans the same subspace. Let $\mathcal{H}$ *be the span of* $\{f_1, f_2, f_3\}$ *and let* $\{e_1, e_2, e_3\}$ *be an orthonormal basis of* $\mathcal{H}$. *One can construct an orthonormal basis* $\{e_1, e_2, e_3\}$ *by applying the Gram-Schmidt orthogonalization process to* $\{f_1, f_2, f_3\}$. *The resulting orthonormal basis can be written as*

$$\begin{bmatrix} e_1 \\ e_2 \\ e_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ -c_{21} & 1 & 0 \\ c_{21}\theta - c_{31} & -\theta & 1 \end{bmatrix} \begin{bmatrix} f_1 \\ f_2 \\ f_3 \end{bmatrix},$$

where

$$c_{21} = sinc(\lambda_2 - \lambda_1), c_{32} = sinc(\lambda_3 - \lambda_2), c_{31} = sinc(\lambda_3 - \lambda_1),$$

and

$$sinc(x) \equiv \frac{\sin(\pi x)}{\pi x},\ \theta = \frac{c_{32} - c_{21}c_{31}}{1 - c_{21}^2}.$$

Then

$$\begin{aligned} f_1 &= e_1, \\ f_2 &= c_{21}e_1 + e_2, \\ f_3 &= c_{31}e_1 + \theta e_2 + e_3, \end{aligned}$$

and the synthesis operator T *of the frame* $\{f_1, f_2, f_3\}$ *can be written in matrix form as*

$$\begin{bmatrix} 1 & c_{21} & c_{31} \\ 0 & 1 & \theta \\ 0 & 0 & 1 \end{bmatrix}.$$

Next the polar decomposition of the matrix of T *is computed, so that* $T = W|T|$, *where* W *is a partial isometry and* $|T| = (T^*T)^{1/2}$. *In this case, since* T *is invertible,* W *is in fact a unitary matrix. Finally, let* $\{g_1, g_2, g_3\} = W^* \cdot \{e_1, e_2, e_3\}$. *Then* $\{g_1, g_2, g_3\}$ *forms a Parseval frame for* $\mathcal{H}$.

Remark: (1). In this example, since the original frame is linearly independent and therefore a basis for $\mathcal{H}$, *what is obtained as a Parseval frame is in fact an orthonormal basis for* $\mathcal{H}$. *(2). Since each* g_n *can be written as a linear combination of* f_1, f_2, *and* f_3, *the Parseval frame constructed indeed spans the same subspace as the original frame.*

Example II.2. *Let* $\lambda_1 = 3 + \frac{1}{3}, \lambda_2 = 4 + \frac{1}{4}, \lambda_3 = 5 + \frac{1}{5}$ *and let* $f_1 = e^{2\pi i\lambda_1 x}, f_2 = e^{2\pi i\lambda_2 x}, f_3 = e^{2\pi i\lambda_3 x}, f_4 = f_1 + f_2, f_5 = f_1 + f_3$, *and* $f_6 = f_2 + f_3$. *Consider the frame* $\{f_1, f_2, f_3, f_4, f_5, f_6\}$ *of a subspace of* $L^2([-1/2, 1/2])$. *Denote this subspace by* $\mathcal{H}$. *Starting from the linearly independent set* $\{f_1, f_2, f_3\}$ *that spans* $\mathcal{H}$, *one can construct an orthonormal basis* $\{e_1, e_2, e_3\}$ *for* $\mathcal{H}$ *as done in Example II.1. From Example II.1,*

$$\begin{aligned} f_1 &= e_1, \\ f_2 &= c_{21}e_1 + e_2, \\ f_3 &= c_{31}e_1 + \theta e_2 + e_3, \\ f_4 &= f_1 + f_2 = (1 + c_{21})e_1 + e_2, \\ f_5 &= f_1 + f_3 = (1 + c_{31})e_1 + \theta e_2 + e_3, \\ f_6 &= f_2 + f_3 = (c_{21} + c_{31})e_1 + (1 + \theta)e_2 + e_3, \end{aligned}$$

where c_{21}, c_{31}, *and* θ *are as defined in Example II.1. The synthesis operator* T *has the matrix representation*

$$\begin{bmatrix} 1 & c_{21} & c_{31} & 1 + c_{21} & 1 + c_{31} & c_{21} + c_{31} \\ 0 & 1 & \theta & 1 & \theta & 1 + \theta \\ 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix}.$$

Let the polar decomposition of T *be given by* $T = W|T|$. *Let* $\{g_1, g_2, g_3, g_4, g_5, g_6\} = W^* \cdot \{e_1, e_2, e_3\}$. *Note that* W^* *is a 6 by 3 matrix. Then it can be shown that* $\{g_k \colon 1 \le k \le 6\}$ *forms a Parseval frame for* $\mathcal{H}$.

Example II.3. *A Fourier frame of three elements is first constructed using Example I.2. Let* $c = 1,\ R = 1/4$, *and* $\delta = 1/4$. *Three points on the spiral* $A_{c=1} = \{\theta \cos 2\pi\theta, \theta \sin 2\pi\theta\}$ *that have arc-length between them less than* 2δ, *starting with* 2δ *from the origin, can be obtained by taking three values of* θ *to be* $\theta_1 = 1/16$, $\theta_2 = 1/8$, *and* $\theta_3 = 1/4$. *This choice gives the following three points on the spiral*

$$\lambda_1 = (\frac{1}{16}\cos\frac{\pi}{8}, \frac{1}{16}\sin\frac{\pi}{8}) = (0.06, 0.02),$$
$$\lambda_2 = (\frac{1}{8}\cos\frac{\pi}{4}, \frac{1}{8}\sin\frac{\pi}{4}) = (0.09, 0.09),$$

and

$$\lambda_3 = (\frac{1}{4}\cos\frac{\pi}{2}, \frac{1}{4}\sin\frac{\pi}{2}) = (0, 1/4).$$

Thus $X = \{e_{\lambda_1}, e_{\lambda_2}, e_{\lambda_3}\}$ *is a Fourier frame for* $\mathrm{span}\{e_{\lambda_1}, e_{\lambda_2}, e_{\lambda_3}\}$.

For implementation purposes, to get the symmetric approximation, one can think of discretizing the ball $B(0, 1/4)$ *by changing into polar coordinates and looking at the rectangle* $\{(r, \theta) : 0 \le r \le 1/4, 0 \le \theta \le 2\pi\}$. *One can then divide each side of the rectangle into* N *subintervals partitioning it into* N^2 *rectangles. The exponential functions from the set* X *are then evaluated at* N^2 *grid-points, taking one point from each small rectangle and thus obtaining a vector* v_i *of length* N^2 *for each* $e_{\lambda_i},\ i = 1, 2, 3$. *Looking at the synthesis operator* F *of* X *as the matrix* $[F]$ *whose columns are* v_i; *such a matrix will be of size* N^2 *by* 3. *After carrying out the polar decomposition of* $[F]$ *using Matlab, one can get the discretized Parseval frame* $\{u_i\}_{i=1}^3$ *that will be considered as the symmetric approximation of the above Fourier frame.*

Suppose one is interested in reconstructing a function f *in* $\mathrm{span}\{e_{\lambda_1}, e_{\lambda_2}, e_{\lambda_3}\}$. *First* f *is converted into a vector* $[f]$ *of size* N^2 *by evaluating it at the* N^2 *points on the rectangular grid above. Then* f *is reconstructed at the* N^2 *points as*

$$[\tilde{f}] = \sum_{j=1}^{3} \langle [f], u_i \rangle u_i.$$

The results are shown in Figures 1 and 2 for the reconstruction of $f = e_{\lambda_1}$ and $f = e_{\lambda_1} - 2e_{\lambda_2} + e_{\lambda_3}$, respectively. Only the real part of the original and the reconstructed functions are plotted. Also, for clarity of reading the figures, only a certain number of points are plotted instead of all the N^2 points.

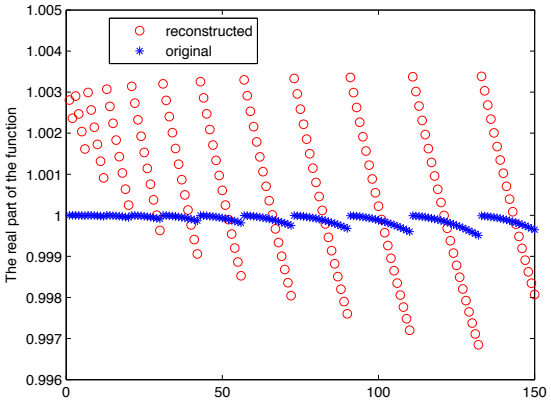


Fig. 1. Reconstruction of the function $f = e_{\lambda_1}$ using $N = 50$.

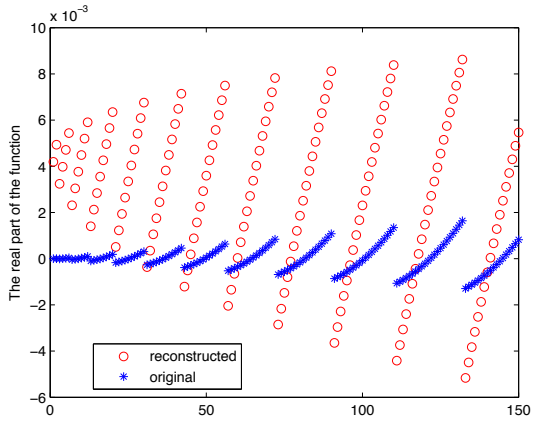


Fig. 2. Reconstruction of the function $f = e_{\lambda_1} - 2e_{\lambda_2} + e_{\lambda_3}$ using $N = 50$.

III. Conclusion

In this paper, the construction of a Parseval frame that is a symmetric approximation of a Fourier frame on a spiral has been considered. Presently, the focus is only on finite frames. This is done by means of the polar decomposition of the matrix corresponding to the synthesis operator of the Fourier frame. The reconstruction of functions lying in the span of such Fourier frames on spirals has been studied. By using a Parseval frame that spans the same space as the original Fourier frame, the reconstruction avoids the need to compute the inverse of the frame operator of the original frame. Besides, the Parseval frame that is obtained by considering the symmetric approximation enables one to reconstruct a function by only using the measurements obtained from the original Fourier frame.

Finding a Parseval frame for some general separable Hilbert space that is a symmetric approximation of a given frame involves finding the polar decomposition of the synthesis operator. This constitutes ongoing research. For practical purposes, even after finding a Parseval frame, it is not possible to use an infinite frame and one should think of finding the best N-term approximation. This will be a part of future research.

Acknowledgment

The authors are immensely grateful to John Benedetto for being a constant source of inspiration and a mathematical role model. The authors would also like to thank the anonymous reviewers for their valuable comments and suggestions. The first named author is supported by a postdoctoral fellowship from the Pacific Institute for the Mathematical Sciences. The second named author was partially supported by AFOSR Grant No. FA9550-10-1-0441.

References

[1] J. J. Benedetto, A. Powell, and H. C. Wu, "MRI signal resonstruction by Fourier frames on interleaving spirals," in *Proceedings of IEEE International Symposium on Biomedical Imaging*, 2002, pp. 717 – 720.

[2] J. J. Benedetto and H. C. Wu, "A Beurling covering theorem and multidimensional irregular sampling," in *SampTA*, Loen, 1999.

[3] J. J. Benedetto and H. Wu, "A multidimensional irregular sampling algorithm and applications," in *ICASSP*, 1999.

[4] J. J. Benedetto and H. C. Wu, "Nonuniform sampling and spiral MRI reconstruction," in *Proc. SPIE, Wavelet Applications in Signal and Image Processing*, vol. 4119, 2000, pp. 130–141.

[5] H. J. Landau, "Necessary density conditions for sampling and interpolation of certain entire functions," *Acta Math. J.*, vol. 117, pp. 37–52, 1967.

[6] R. M. Redheffer and R. M. Young, "Completeness and basis properties of complex exponentials," *Trans. Amer. Math. Soc.*, vol. 277, no. 1, pp. 93–111, 1983.

[7] A. Olevskii and A. Ulanovskii, "Universal sampling and interpolation of band-limited signals," *Geom. Funct. Anal.*, vol. 18, no. 3, pp. 1029–1052, 2008.

[8] ——, "Universal sampling of band-limited signals," *C. R. Math. Acad. Sci. Paris*, vol. 342, no. 12, pp. 927–931, 2006.

[9] G. Kozma and N. Lev, "Exponential Riesz bases, discrepancy of irrational rotations and BMO," *J. Fourier Anal. Appl.*, vol. 17, no. 5, pp. 879–898, 2011.

[10] H. G. Feichtinger and K. Grochenig, "Irregular sampling theorems and series expansions of band-limited functions," *J. Math Anal. Appl.*, vol. 167, no. 2, pp. 530–556, 1992.

[11] B. Matei and Y. Meyer, "Simple quasicrystals are sets of stable sampling," *Complex Var. Elliptic Equ.*, vol. 55, no. 8-10, pp. 947–964, 2010.

[12] M. Frank, V. I. Paulsen, and T. R. Tiballi, "Symmetric approximation of frames and bases in Hilbert spaces," *Trans. Amer. Math. Soc.*, vol. 354, no. 2, pp. 777–793, 2002. [Online]. Available: http://dx.doi.org/10.1090/S0002-9947-01-02838-0

[13] O. Christensen, *An Introduction to Frames and Riesz Bases*. Birkhäuser, 2003.

[14] I. Daubechies, *Ten Lectures on Wavelets*. SIAM, 1992.

[15] A. Beurling, "Local harmonic analysis with some applications to differential operators," in *Some Recent Advances in the Basic Sciences, Vol. 1 (Proc. Annual Sci. Conf., Belfer Grad. School Sci., Yeshiva Univ., New York, 1962–1964)*. Belfer Graduate School of Science, Yeshiva Univ., New York, 1966, pp. 109–125.

Measure-based diffusion kernel methods

Amit Bermanis
School of Computer Science,
Tel Aviv University,
Tel Aviv 69978, Israel
Email: amitberm@post.tau.ac.il

Guy Wolf
School of Computer Science,
Tel Aviv University,
Tel Aviv 69978, Israel
Email: guy.wolf@cs.tau.ac.il

Amir Averbuch
School of Computer Science,
Tel Aviv University,
Tel Aviv 69978, Israel
Email: amir@math.tau.ac.il

***Abstract*—A commonly used approach for analyzing massive high dimensional datasets is to utilize diffusion-based kernel methods. The kernel in these methods is based on a Markovian diffusion process, whose transition probabilities are determined by local similarities between data points. When the data lies on a low dimensional manifold, the diffusion distances according to this kernel encompass the geometry of the manifold. In this paper, we present a generalized approach for defining diffusion-based kernels by incorporating measure-based information, which represents the density or distribution of the data, together with its local distances. The generalized construction does not require an underlying manifold to provide a meaningful kernel interpretation but assumes a more relaxed assumption that the measure and its support are related to a locally low dimensional nature of the analyzed phenomena.**

I. Introduction

The diffusion maps (DM) method [3] is a popular kernel method that utilizes a stochastic diffusion process to analyze the data. It defines diffusion affinities via symmetric conjugation of a transition probability operator. These probabilities are based on local distances between the data points. The Euclidean distances in the embedded space represent the diffusion distances in the original space. When the data is sampled from a low dimensional manifold, the diffusion paths follow the manifold and the diffusion distances capture its geometry.

In this paper, we enhance the DM method by incorporating information about the distribution of the data, in addition to local distances on which DM is based. This distribution is expressed in term of a measure over the observable space. The measure (and its support) replace the manifold assumption. We assume that the measure quantifies the likelihood for the presence of data over the geometry of the space. This assumption is significantly less restrictive than the need to have a manifold present. In practice this measure can either be provided as an input (e.g., by a-priori knowledge or a statistical model), or deduced from a given training set (e.g., by a density estimator). The manifold assumption can be expressed in terms of the measure assumption by setting the measure to be concentrated around an underlying manifold or (in the extremely restrictive case), to be supported by the manifold. Therefore, the measure assumption is not only less restrictive than the manifold assumption but it also generalizes it.

In the suggested construction, the used measure, which can represent densities, is separated from the distances and from the analyzed dataset. Therefore, when dealing with discrete data, this construction can utilize two different sets of samples: the analyzed dataset and the measure-related set with attached empirical measure values. Furthermore, from theoretical point of view, this construction combines continuous measures with either discrete or continuous datasets.

II. Problem setup

Let $\Omega \subseteq \mathbb{R}^n$, for some natural n, be a metric space with the Euclidean distance metric $\|\cdot\|$. The integration notation $\int \cdot dy$ in this paper will refer to the Lebesgue integral $\int_\Omega \cdot dy$ over the subspace Ω, instead of the whole space $\mathbb{R}^n$. Let μ be a probability measure defined on Ω and let $q(x)$ be the distribution function of μ, i.e., $d\mu(x) = q(x)dx$. This measure represents the distribution of data in Ω. We aim to combine the distance metric of Ω and the measure μ to define a kernel function $k(x, y)$, $x, y \in \Omega$, which represents the affinities between data points in Ω. Then, these affinities can be used to construct a diffusion map, as described in Section II-A, and utilize it to embed the data into a low-dimensional representation that considers both proximities and distributions of the data points.

A. Diffusion maps

The diffusion maps (DM) framework utilizes a set of affinities to define a Markovian (random-walk) diffusion process over the analyzed data [3]. The spectral properties of this process are then used to obtain a representation of the data, where diffusion distances are expressed as Euclidean distances. The achieved representation reveals the underlying patterns of the data such as clusters and differences between normal and abnormal regions.

Technically, DM is based on an affinity kernel k and the associated integral operator that is defined as $Kf(x) = \int k(x, y)f(y)dy$, $x \in \Omega$, for any function $f \in L^2(\Omega)$. The affinity kernel k is normalized by a set of degrees $\nu(x) \triangleq \int k(x, y)dy$, $x \in \Omega$, to obtain the transition probabilities $p(x, y) \triangleq k(x, y)/\nu(x)$, from $x \in \Omega$ to $y \in \Omega$, of the Markovian diffusion process. Under mild conditions on the kernel k, the resulting transition probability operator has a discrete decaying spectrum of eigenvalues $1 = \lambda_0 \geq |\lambda_1| \geq |\lambda_2| \geq \ldots$, which are used together with their corresponding eigenvectors $\vec{1} = \phi_0, \phi_1, \phi_2, \ldots$ to achieve the diffusion map of the data.

Each data point $x \in \Omega$ is embedded by this diffusion map to the diffusion coordinates $(\lambda_1\phi_1(x), \ldots, \lambda_\delta(x)\phi_\delta(x))$, where the exact value of δ depends on the spectrum of the transition probabilities operator P, whose kernel is $p(x,y)$. The relation between the diffusion distance metric $\|p(x,\cdot) - p(y,\cdot)\|$ and the Euclidean distances in the embedded space, is a result of the spectral theorem [3], [5]. When the data in Ω lies on a low dimensional manifold, its tangent spaces can be utilized to express the infinitesimal generator of the associated diffusion process in terms of the Laplacian operators on the manifold.

III. Measure-based diffusion and affinity kernels

In this section, we define and analyze an affinity kernel that is based on the distances in Ω and on the measure μ. We use this kernel together with the DM method, which was briefly described in Section II-A, to obtain a measure-based diffusion affinity kernel and its resulting diffusion map. In Section III-A, we show the relations between the infinitesimal generator of the resulting diffusion operator and the Laplacian operator on the space Ω and the measure μ.

In order to define the desired kernel, we first define the function

$$g_\varepsilon(t) \triangleq \begin{cases} e^{-t^2/\varepsilon} & t \le \rho\sqrt{\varepsilon} \\ 0 & \text{otherwise} \end{cases}, \qquad \text{(III.1)}$$

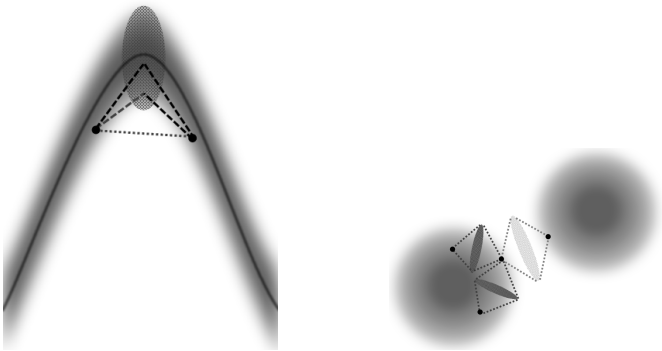
for any $\varepsilon > 0$ and some constant $\rho \gg 1$. Notice that for a sufficiently large ρ, the Gaussian kernel, which is usually used in the DM method, can be defined as $k_\varepsilon(x,y) \triangleq g_{2\varepsilon}(\|x-y\|)$, and this definition will be used in the rest of the paper. Definition III.1 uses the function g_ε to define an alternative kernel that incorporates both local distance information, as the Gaussian kernel does, and measure information, which the Gaussian kernel lacks.

Definition III.1 (Measure-based Gaussian Correlation kernel)**.** *The Measure-based Gaussian Correlation (MGC) affinity function* $\tilde{k}_\varepsilon : \Omega \times \Omega \to \mathbb{R}$ *is defined as* $\tilde{k}_\varepsilon(x,y) \triangleq \int g_\varepsilon(\|x-r\|) \cdot g_\varepsilon(\|y-r\|) d\mu(r)$. *The MGC integral operator is defined by this function as* $\tilde{K}_\varepsilon f(x) = \int \tilde{k}_\varepsilon(x,y) f(y) dy$ *for every function* $f \in L^2(\Omega)$ *and data point* $x \in \Omega$.

The MGC affinity from Definition III.1, is in fact the inner product in $L^2(\Omega, \mu)$ (correlation) between two Gaussians of width ε that are centered at x and y, respectively. This affinity takes into consideration the measure μ, between the described Gaussians around at the examined data points. The numerically significant positions of r in this correlation must be close enough to x and to y (based on their Gaussians of radius ε), but they must also be in an area with a high enough concentration of the measure μ. Notice that the measure information is considered and incorporated in the affinity definitions. From the identity $\|x-r\|^2 + \|y-r\|^2 = \frac{1}{2}\|x-y\|^2 + 2\left\|\frac{x+y}{2} - r\right\|^2$, the MGC affinity function becomes

$$\tilde{k}_\varepsilon(x,y) = k_\varepsilon(x,y) \cdot \int g_{\varepsilon/2}\left(\left\|\frac{x+y}{2} - r\right\|\right) d\mu(r). \qquad \text{(III.2)}$$

Equation III.2 shows the relation between the MGC kernel and the Gaussian kernel $k_\varepsilon(x,y)$. While the Gaussian affinity only considers the distances between the examined data points, the MGC affinity also considers the region in which this distance is measured by using a Gaussian around the midpoint between them. This midpoint represents the direct path that determines the distance between the two data points. For a given distance between two data points, the MGC affinity increases when its path lies in an area with a high concentration of the measure μ, and decreases when it lies in an area with a low concentration of μ. If the measure μ is uniform over Ω, then the MGC kernel becomes the same as the Gaussian kernel up to a constant.



(a) When the data lies around a curve, the MGC affinities consider paths that follow the curve.

(b) When the data lies in two separate clusters, the affinities between data points within a cluster are higher than data points from a different cluster.

Fig. III.1. An illustration of the MGC affinities in two common data analysis scenarios. For every pair of compared data points, the significant values of the integration variable r, from Definition III.1 or the equivalent representation from Eq. III.2, are marked.

The dual representation of the MGC kernel in Definition III.1 and Eq. III.2 can be used to detect and consider several common patterns in data analysis directly from the initial construction of the kernel. Figure III.1(a) uses the formulation in Definition III.1 to illustrate a case when the data is concentrated in areas around a curve with significant curvatures. In this case, the affinity will be more affected by the distances over the path that follows the "noisy" curve and not by the directions that follow sparse areas and bypass the curve. Figure III.1(b) uses the formulation in Eq. III.2 to illustrate the affinities when the data is concentrated in two distinct clusters. In this case, we can see that the affinity between data points from different clusters is significantly reduced due to the measure even if they are relatively close.

As proved in [1], the presented MGC affinity kernel satisfies the spectral properties that are required (and assumed) in [3], [5] for its utilization with the DM framework. These properties enable us to define a diffusion process that is based on the MGC affinities. Then, the resulting diffusion map is used to embed the data in a way that considers the distances and the measure distribution.

A. Infinitesimal generator

The DM framework is based on Markovian diffusion process, which is defined and represented by a transition probability operator denoted by P_ε. The infinitesimal generator of this operator encompasses the nature of the diffusion process. In [3], [5], it was shown that when the data is sampled from a low dimensional underlying manifold, the infinitesimal generator of P_ε has the form of *Laplacian+Potential*. In this section, we show a similar result, when using the MGC-based diffusion without requiring the underlying manifold assumption to hold.

The MGC affinity function $\tilde{k}_\varepsilon$ is symmetric and positive, i.e., $\tilde{k}_\varepsilon(x,y) > 0$ for any pair of data points $x, y \in \Omega$. To convert it to be a transition kernel of a Markov chain on Ω, we normalize it to be $\tilde{p}_\varepsilon(x,y) \triangleq \frac{\tilde{k}_\varepsilon(x,y)}{\nu_\varepsilon(x)}$. We define the corresponding stochastic operator $\tilde{P}_\varepsilon f(x) \triangleq \int \tilde{p}_\varepsilon(x,y) f(y) dy$.

The infinitesimal generator of the diffusion transition operator $\tilde{P}_\varepsilon$ is defined as $\mathcal{L} \triangleq \lim_{\varepsilon \to 0} (I - \tilde{P}_\varepsilon)/\varepsilon$. Theorem III.1, whose proof appears in [1], shows that the operator $\mathcal{L}$ takes the form *Laplacian+potential*, which is similar to the result shown in [5, Corollary 2]. The expression, which Theorem III.1 provides for $\mathcal{L}$, characterizes the differential equation for diffusion processes [2], [4].

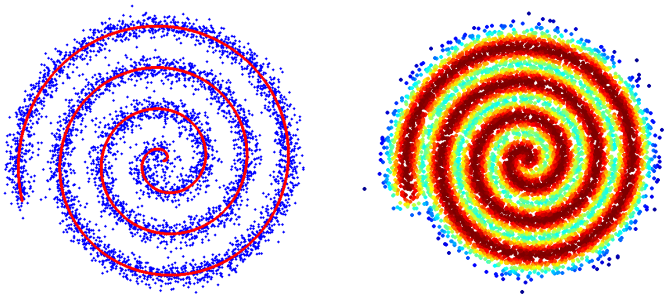
Theorem III.1. *If the density function q is in $C^4(\Omega)$, then the infinitesimal generator $\mathcal{L}$ of the MGC-based diffusion operator is*

$$\mathcal{L}f = -\frac{m_2}{m_0}\left(\Delta f + \left\langle \frac{\nabla q}{q}, \nabla f \right\rangle\right), \qquad f \in C^4(\Omega),$$

where, $m_0 = \int g_1(\|x\|)dx$ and $m_2 = \int g_1(\|x\|)(x^{(j)})^2 dx$.

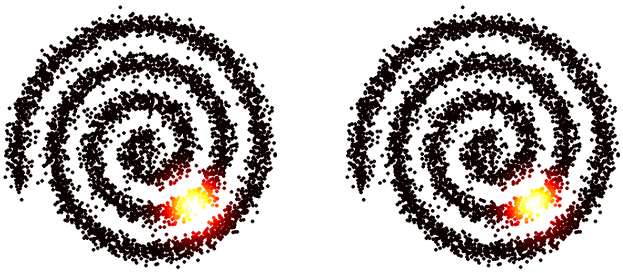
IV. Geometric example

In this section, we demonstrate the MGC kernel and the resulting diffusion map. A noisy data that is spread around a spiral curve is analyzed, and the results are compared with the "classic" DM [3]. This example also demonstrates the separation between the analyzed data and the data distribution, which is a unique feature of the presented method.



(a) Noisy data around the curve (b) An exponentially-decaying measure around the curve

Fig. IV.1. A spiral curve with 5000 noisy data points concentrated around it, and 10^4 points that represent an exponentially-decaying measure around the curve. Red color indicates large measure weights and blue color indicates small measure weights.



(a) K neighborhood (b) $\tilde{K}_v$ neighborhood

Fig. IV.2. A neighborhoods from the Gaussian kernel and the MGC kernels on the spiral curve. Close points are colored by white, and far points are colored by black.

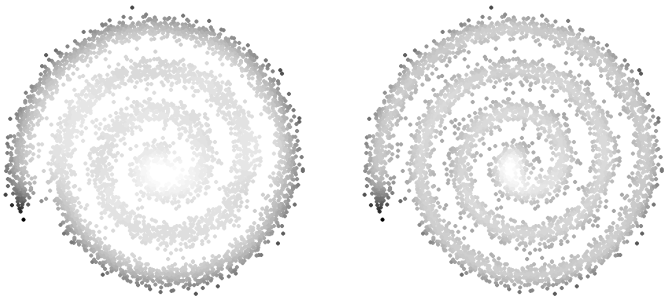
We use a noisy spiral curve (see Fig. IV.1(a)) for the comparison between MGC-based DM and the classical DM. The dataset was produced by sampling 500 equally spaced points from the curve and then sampling 10 normally distributed data points around each of these curve points. The resulting data has 5000 data points that lie in areas around the curve, as shown in Fig. IV.1(a), where the curve is marked in red and the noisy data points are marked in blue. We used the same scale meta-parameter ε to the compared DM applications. This meta-parameter was set to be sufficiently high to overcome the noise and to detect the high affinity between data points that originated from the same position (out of the 500 curve points) on the curve.

The MGC kernel from Definition III.1 requires to define a measure over the area where the data lies. Notice that the measure of the actual data points is not required. We can define a completely different set of points r from Definition III.1 and then define their weights, which represent their measure values. The measure we used is based on 10^4 points, distributed normally around a spiral curve. The weights of the point decay exponentially in relation to their distance from the curve. The resulting measure is denoted by μ_v and it is presented in Fig. IV.1(b).

We use the notation $\tilde{K}_v$ to denote the matrix that results from Definition III.1, with the measure μ_v. Notice that even though the measure is based on 10^4 positions of the integration variable r (from Definition III.1), the kernel and its normalized versions are of size 5000×5000, since the data has only 5000 data points.

Figure IV.2 compares the neighborhoods that are represented by the kernels K and $\tilde{K}_v$. While the Gaussian kernel captures inter-level affinities (i.e., it links different levels of the spiral), the MGC kernel only capture relations in the same level of the spiral, thus, it is able to separate between these levels. In addition, the shape of the neighborhoods of the MGC kernel form ellipses whose major axes clearly follow the significant tangential directions of the curve. The Gaussian kernel, however, captures circular neighborhoods that do not express any information about the significant directions of the data.

The embedding, which is achieved by DM, is based on



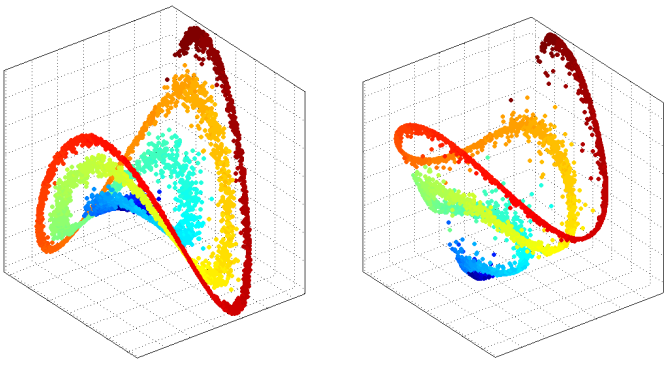
(a) Gaussian-based stationary distribution

(b) MGC-based stationary distribution

Fig. IV.3. The stationary distributions of: (a) the Gaussian-based diffusion process, and (b) the MGC-based diffusion process (low densities are represented by dark gray levels, and vise-versa.)

a diffusion process that has a stationary distribution when the time is taken to infinity. This distribution reveals the concentrations and the underlying potential of the diffusion process. It is represented by the first left eigenvector of the diffusion transition operator. Figure IV.3 compares the stationary distributions of the Gaussian-based diffusion with the MGC-based diffusion. This comparison shows that the Gaussian-based diffusion considers the entire spiral as one pit of potential. At infinity, the diffusion is distributed over the entire region of the curve. The MGC-based diffusion, on the other hand, separates different levels of the spiral. At infinity, this diffusion is concentrated on the curve levels themselves and not on the areas between them.

Finally, we compare between the embedded spaces of the Gaussian-based DM and the MGC-based DM. Figure IV.4 presents these spaces based on the first three diffusion coordinates. The comparison in Fig. IV.4 clearly shows that the MGC-based embedding results in a better separation between the spiral levels. Figure IV.4 further establishes this observation by showing that, in fact, the Gaussian-based diffusion considers the whole noisy spiral as a two-dimensional disk. The MGC-based embedding, on the other hand, separates the levels of the spiral by "stretching" it apart in the three-dimensional embedded space.



(a) Gaussian-based DM

(b) MGC-based DM

Fig. IV.4. The first three diffusion coordinates of the Gaussian-based and MGC-based DM embeddings.

The superior results (e.g., separation between the spiral levels) of the MGC-based DM demonstrate its robustness to noise. The reason for this robustness is because the noise is part of the model on which the MGC construction is based. The Gaussian-based DM assumes that the data lies on (or it is sampled from) an underlying manifold, and any significant noise outside this manifold may violate this assumption. The MGC-based DM, on the other hand, already assumes variable concentrations and distributions of the data, which are represented by the measure and incorporated into the affinities. Therefore, this setting is more natural when dealing with data that is concentrated *around* an underlying manifold structure but does not necessarily lie on the manifold.

V. Conclusion

We presented a generalized version of DM, which is based on the MGC kernel instead of the Gaussian kernel. We replaced the commonly-used manifold assumption in DM with a measure assumption. Namely, we assume access to a measure that represents the locally low dimensional nature of the analyzed data, its distributions and its densities. The MGC kernel was presented and formulated in two equivalent forms that incorporate the measure-based information together with local distances between data points. The infinitesimal generator of the MGC-based diffusion process is similar to the diffusion process in [3], and its spectral properties enable its utilization for dimensionality reduction.

We demonstrated the robustness of the MGC-based DM to noise, which is due to the noise being considered as part of the measure assumption while it violates the manifold assumption. Since the MGC-based construction considers the measure and the data points separately, it is able to analyze a given measure distribution by using a separated grid, as we will show in future work. This application cannot be achieved by the classic DM [3], which is based solely on local distances and does not consider a separately-provided measure.

Acknowledgments

This research was supported by the Israel Science Foundation (Grant No. 1041/10) and the Eshkol Fellowship from the Israeli Ministry of Science & Technology.

References

[1] A. Bermanis, G. Wolf, and A. Averbuch. Diffusion-based kernel methods on Euclidean metric measure spaces. *Submitted*, 2012.

[2] R.R. Coifman, I.G. Kevrekidis, S. Lafon, M. Maggioni, and B. Nadler. Diffusion maps, reduction coordinates, and low dimensional representation of stochastic systems. *Multiscale Modeling and Simulation*, pages 842–864, 2008.

[3] R.R. Coifman and S. Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, 2006.

[4] R.R. Coifman, S. Lafon, A.B. Lee, M. Maggioni, B. Nadler, F. Warner, and S.W. Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proceedings of the National Academy of Sciences of the United States of America*, 102(21):7426–7431, 2005.

[5] S. Lafon. *Diffusion Maps and Geometric Harmonics*. PhD thesis, Yale University, May 2004.

Spectral properties of dual frames

Felix Krahmer
Georg-August-Universität Göttingen
Institut für Numerische und Angewandte Mathematik
37083 Göttingen
Germany
Email: f.krahmer@math.uni-goettingen.de

Gitta Kutyniok
Technische Universität Berlin
Institut für Mathematik
10623 Berlin
Germany
Email: kutyniok@math.tu-berlin.de

Jakob Lemvig
Technical University of Denmark
Department of Mathematics
2800 Kgs. Lyngby
Denmark
Email: J.Lemvig@mat.dtu.dk

***Abstract*—We study spectral properties of dual frames of a given finite frame. We give a complete characterization for which spectral patterns of dual frames are possible for a fixed frame. For many cases, we provide simple explicit constructions for dual frames with a given spectrum, in particular, if the constraint on the dual is that it be tight.**

I. Introduction

In signal processing, one of the primary objectives is to obtain suitable representations of the signals of interest. Finite frames are redundant systems in a finite-dimensional Hilbert space, which give redundant representations of finite-dimensional signals. The representation process can be split into two steps: the decomposition and the reconstruction. For each frame decomposition method, there is one canonical reconstruction using a least-squares approach. However, due to the redundancy of frames, there are many alternative reconstruction methods. Each of these alternative reconstruction methods is associated to a so-called dual frame.

It is therefore natural to ask which dual frame for the reconstruction step is the best to choose in case the decomposition frame is given by the application at hand, e.g., by the way of measuring the data. The precise answer to this question is, of course, dependent on the application, but universal desirable properties of the dual can, nonetheless, be recognized. Among such desirable properties are fast and stable reconstruction. It turns out that the computational properties of the dual frames such as the *stability* of the reconstructions are directly linked to spectral properties of the frame. In particular, the Frobenius norm and the spectral norm of the so-called dual frame matrix play an important role in this context. In Subsection I-B below, we will illustrate the importance of these matrix norms in a situation, where we want to minimize the effect of noise from a noisy decomposition. Before we embark on this, we will need some basic definition from frame theory.

A. Setup and basic observations

Let us recall some basic definitions and facts from frame theory. For an extensive exposition on frames and their applications, we refer the reader to the books [1], [2]. We let $\mathbb{K}$ denote either $\mathbb{C}$ or $\mathbb{R}$ and define frames in $\mathbb{K}^n$ as follows.

Definition I.1. A collection of vectors $\Phi = (\phi_i)_{i=1}^m \subset \mathbb{K}^n$ is called a *frame* for $\mathbb{K}^n$ if there are two constants $0 < A \leq B$ such that

$$A \|x\|_2^2 \leq \sum_{i=1}^m |\langle x, \phi_i \rangle|^2 \leq B \|x\|_2^2, \quad \text{for all } x \in \mathbb{K}^n.$$

If the frame bounds A and B are equal, the frame $(\phi_i)_{i=1}^m$ is called a *tight frame* for $\mathbb{K}^n$.

In this paper, we are interested in the case $m > n$, where the frame $(\phi_i)_{i=1}^m$ is *redundant*, i.e., it consists of more vectors than necessary for the spanning property. For these frames there exist infinitely many dual frames. The precise definition of dual frames is the following:

Definition I.2. Given a frame Φ, another frame $\Psi = (\psi_i)_{i=1}^m \subset \mathbb{K}^n$ is said to be a *dual frame* of Φ if the following reproducing formula holds:

$$x = \sum_{i=1}^m \langle x, \phi_i \rangle \psi_i \quad \text{for all } x \in \mathbb{K}^n.$$

In matrix notation this definition reads

$$\Psi \Phi^* = I_n, \tag{1}$$

where the maps induced by Φ^* and Ψ correspond to the decomposition and reconstruction procedure, respectively, and where I_n is the $n \times n$ identity matrix. Hence, the set of all duals of Φ is the set of all left-inverses Ψ to Φ^*. The particular choice of Ψ as the Moore-Penrose pseudoinverse of Φ^* is the *canonical dual frame* of Φ.

From (1) it is immediate that the set of all duals Ψ to a frame Φ is an $n(m-n)$-dimensional affine subspace of $\mathrm{Mat}(\mathbb{K}, n \times m)$. A natural parametrization of this space is obtained using the singular value decomposition. Let $\Phi = U \Sigma_\Phi V^*$ be a full SVD of Φ, i.e., $U \in \mathbb{K}^{n \times n}$ and $V \in \mathbb{K}^{m \times m}$ are unitary and $\Sigma_\Phi \in \mathbb{R}^{n \times m}$ is a diagonal matrix whose entries, namely $\sqrt{B} = \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n = \sqrt{A} > 0$, are non-negative and arranged in a non-increasing order. We will sometimes write the ith singular value of Φ as σ_i^Φ. Let Ψ be a frame and define $M_\Psi := U^* \Psi V \in \mathbb{K}^{n \times m}$, where U and V are the right and left singular vectors of Φ. Then Ψ factors as $\Psi = U M_\Psi V^*$. By $\Phi \Psi^* = I_n$, we then see that

$$I_n = U^* I_n U = U^* \Phi \Psi^* U = \Sigma_\Phi M_\Psi^*.$$

Therefore, Ψ is a dual frame of Φ precisely when

$$\Sigma_\Phi M_\Psi^* = I_n, \tag{2}$$

where $\Psi = UM_\Psi V^*$. The solutions to (2) are given by

$$M_\Psi = \begin{bmatrix} \frac{1}{\sigma_1} & 0 & \cdots & 0 & s_{1,1} & s_{1,2} & \cdots & s_{1,r} \\ 0 & \frac{1}{\sigma_2} & & 0 & s_{2,1} & s_{2,2} & \cdots & s_{2,r} \\ & & \ddots & & \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \frac{1}{\sigma_n} & s_{n,1} & s_{n,2} & \cdots & s_{n,r} \end{bmatrix}, \tag{3}$$

where $s_{i,k} \in \mathbb{K}$ for $i = 1, \dots, n$ and $k = 1, \dots, r = m - n$. Note that the canonical dual frame is obtained by taking $s_{i,k} = 0$ for all $i = 1, \dots, n$ and $k = 1, \dots, m-n$. More importantly, since U and V are unitaries, the possible spectrum of duals Ψ is completely described by the matrices M_Ψ in (3).

B. Measures of the goodness of dual frames

In this subsection we consider the important scenario when the frame coefficients $c = \Phi^* x$ of the signal $x \in \mathbb{K}^n$ are corrupted by noise e. We will assume that the noise components e_i corresponding to the different frame coefficients are centered, uncorrelated, and of the same variance. This is a standard setup used e.g., in [3] for unit-norm frames, and in [7] for the case of Gaussian white noise. We will here follow the above general setup from [5]. We remark that it is possible to study an alternative scenario of corruptions through erasures, see [8]–[10].

The reconstruction error is given by

$$\|\Psi\tilde{c} - x\|_2 = \|\Psi(\Phi^* x + e) - x\|_2 = \|\Psi e\|_2,$$

where the corrupted frame coefficients are $\tilde{c} = c + e$. Hence, we see that different duals Ψ yield different reconstruction accuracy. It can be shown, see e.g., [5], that the expected error is controlled by the Frobenius norm of the matrix Ψ. To be precise, one has the expected reconstruction accuracy

$$\mathbb{E}\|\Psi\tilde{c} - x\|_2 \leq \sqrt{\frac{\delta B}{m}} \|\Psi\|_F,$$

where the variance satisfies $\sigma^2 \leq \frac{\delta B}{m}$ with B being the upper frame bound of Φ and $\delta < 1$. This shows that the Frobenius norm of the dual frame matrix Ψ is crucial in the average case scenario.

For the worst case scenario the spectral norm of Ψ is the correct measure. This is seen as follows. Recall that the condition number of an $n \times n$ invertible matrix T is given by $\operatorname{cond}(T) = \max\left(\frac{\text{relative output error}}{\text{relative input error}}\right) = \sigma_1^T / \sigma_n^T$. For a pair of dual frames similar considerations give

$$\begin{aligned} \operatorname{cond}(\Phi, \Psi) &:= \max\left(\frac{\|\Psi e\|_2 / \|x\|_2}{\|e\|_2 / \|\Phi^* x\|_2} \right) \\ &= \max\left(\frac{\|\Psi e\|_2}{\|e\|_2} \frac{\|\Phi^* x\|_2}{\|x\|_2} \right) \\ &= \|\Phi\|_{2\to 2} \|\Psi\|_{2\to 2} = \sigma_1^\Phi \, \sigma_1^\Psi. \end{aligned}$$

Note that if Φ is an invertible matrix, we recover the usual definition: $\operatorname{cond}(\Phi, \Psi) = \sigma_1^\Phi / \sigma_n^\Phi$. We see that only the largest singular value of Ψ plays a role in the measure of goodness of dual frames for the worst case scenario.

In this subsection we have set up two important measures for the goodness of a dual frame. Since both of these measures are determined by the singular values of the dual frame, we are interested in understanding the possible spectra in the set of all duals of a given frame. This is the theme of the second part of this paper, Section II, where we characterize the possible spectral patterns of dual frames.

II. Spectral properties of duals

In this section we characterize the possible spectra in the set of all dual frames of a given frame. However, we begin with the special case of characterizing frames that admit tight duals, which is exactly the situation when the spectrum of the dual frame is a one point spectrum. The characterization was obtained in [11] and extended in [6].

It turns out that a frame always has a tight dual if the redundancy is two or larger. If the redundancy is less than two, it will only be possible under certain assumptions on the singular values of Φ.

Theorem II.1 ([6], [11]). *Let $n, m \in \mathbb{N}$. Suppose Φ is a frame for $\mathbb{K}^n$ with m frame vectors and lower frame bound A. Then the following assertions hold:*

- *(i) If $m \geq 2n$, then for every $c \geq \frac{1}{A}$, there exists a tight dual frame Ψ with frame bound c.*
- *(ii) If $m = 2n - 1$, then there exists a tight dual frame Ψ; the only possible frame bound is $\frac{1}{A}$.*
- *(iii) Suppose $m < 2n - 1$. Then there exists a tight dual frame Ψ if and only if the smallest $2n - m \in \{2, \dots, n\}$ singular values of Φ are equal. In the positive case, the only possible frame bound is $\frac{1}{A}$.*

Before we turn to a proof of Theorem II.1, let us give a simple dimension counting argument to explain why $m = 2n-1$ is the borderline case. Any dual frame Ψ will be row bi-orthogonal to Φ. Hence, for each $j_0 = 1, \dots, n$, the j_0th row vector ψ^{j_0} of Ψ needs to be orthogonal to the jth row vector ϕ^j of Φ for $j \neq j_0$. For the dual frame Ψ to be tight, the matrix Ψ furthermore needs to be row orthogonal, hence ψ^{j_0} needs to be orthogonal to ψ^j for each $j \neq j_0$. In total, the vector $\psi^{j_0} \in \mathbb{K}^m$ needs to be orthogonal to $2(n-1)$ other vectors. If $m \geq 2n - 1$, it is possible to find $2(n-1) + 1 = 2n - 1$ orthogonal vectors in $\mathbb{K}^m$, which shows that we can find n orthogonal vectors $(\psi^j)_{j=1}^n$ being bi-orthogonal to $(\phi^j)_{j=1}^n$. As a final step to make Ψ a tight dual, we need to scale the vectors ψ^j, $j = 1, \dots, n$, to have equal norm.

The above argument is almost a proof of Theorem II.1. However, we include the following proper proof adapted from [6] since it provides an *explicit* construction procedure for the tight duals.

Proof of Theorem II.1: Let $\Phi = U\Sigma_\Phi V^*$ be a full SVD of Φ, and let Ψ be an arbitrary dual frame. Following Section I-A, we factor the dual frame as $\Psi = UM_\Psi V^*$, where M_Ψ is given as in (3) with $s_{i,k} \in \mathbb{K}$ for $i = 1, \dots, n$ and

$k = 1, \dots, r = m - n$. For Ψ to be tight, we need to choose $s_{i,k}$ such that the rows of M_Ψ are orthogonal and have equal norm. This follows from the fact that Ψ is row orthogonal if and only if M_Ψ is row orthogonal.

As the diagonal block of M_Ψ is well-understood, the duality and tightness constraints translate to conditions for the inner products of the $s_i = (s_{i,1}, \dots, s_{i,r}) \in \mathbb{K}^r$, $i = 1, \dots, n$. Indeed, Ψ is a tight dual frame with frame bound c if and only if, for all $1 \leq i \leq n$, one has

$$c = \frac{1}{\sigma_i^2} + \|s_i\|_2^2 , \tag{4}$$

and, for all $i \neq j = 1, \dots, n$, one has $\langle s_i, s_j \rangle = 0$.

Now assume that $\sigma_n = \sigma_{n-1} = \cdots = \sigma_{p+1} < \sigma_p$ for some $p < n$. As $\sigma_{p+1} < \sigma_i$ for all $1 \leq i \leq p$, (4) implies that all s_i for $i = 1, \dots, p$ must be nonzero vectors even if $s_{p+1}, \dots, s_n$ are all zero. Furthermore, by the value of $\|s_{p+1}\| = \cdots = \|s_n\|$, (4) also determines the norms of $s_1, \dots, s_p$. If $\|s_{p+1}\| = \cdots = \|s_n\| \neq 0$, the sequence $(s_i)_{i=1}^n$ is orthogonal, else the sequence $(s_i)_{i=1}^p$.

If $r \geq n$, that is, if $m \geq 2n$, then any choice of s_n allows for an orthogonal system with compatible norms, so tight dual frames with any frame bound above $\frac{1}{\sigma_n}$ exist and can be efficiently constructed. If $r < n$, then no n vectors can form an orthogonal system, one needs to have $s_n = 0$ and hence also $s_j = 0$ for all $j > p$. So no frame bound other than $\frac{1}{\sigma_n}$ is possible. The remaining vectors $\{s_j\}_{j=1}^p$ are all non-zero, so they must form an orthogonal system. For $r \geq n-p+1$, this is possible, and again a solution satisfying the norm constraints can be efficiently constructed. For $r \leq n - p$, no such system exists, hence there cannot be a tight dual. ∎

We will now derive general conditions on which spectral patterns (now possibly consisting of more than one point) can be achieved by a dual frame of a given frame. The reason that, in the general framework, such an analysis is harder than in the context of tight duals is that in the tight case, the frame operator is a multiple of the identity, hence diagonal in any basis. This no longer holds true if we drop the tightness assumption, so when the orthogonality argument of Theorem II.1 fails, one cannot conclude that there is no dual with a given spectral pattern. However, the orthogonality approach allows us to choose a subset of the singular values of the dual frame freely. In particular, if the redundancy of the frame Φ is larger than 2, it follows that for all spectral patterns satisfying a set of lower bounds, which we will later show to be necessary (see Theorem II.4), a dual with that spectrum can be found using a constructive procedure analogous to the proof of Theorem II.1.

Theorem II.2 ([6]). *Let $n, m \in \mathbb{N}$, and let Φ be a frame for $\mathbb{K}^n$ with m frame vectors and singular values $(\sigma_i)_{i=1}^n$. Suppose that $r \leq m-n$ and that $I \subset \{1, \dots, n\}$ with $|I| = r$. Then, for any sequence $(q_i)_{i \in I}$ satisfying $q_i \geq 1/\sigma_i$ for all $i \in I$, there exists a dual frame Ψ of Φ such that $\{q_i\}_{i \in I}$ is contained in the spectrum of Ψ. Furthermore, it can be found constructively using a sequence of orthogonalization procedures.*

Proof: The proof is just a slight modification of the proof of Theorem II.1. Again, we choose $(s_i)_{i \in I}$ to be orthogonal and the remaining s_i's to be the zero vector. The non-zero s_i vectors are scaled to satisfy

$$q_i^2 = \frac{1}{\sigma_i^2} + \|s_i\|_2^2 ,$$

where $i \in I$. Hence, by this procedure we obtain a dual frame with spectrum $\{q_i\}_{i \in I} \cup \{\sigma_i^{-1}\}_{i \notin I}$. ∎

As a corollary we obtain that using the same simple constructive procedure, one can find dual frames with any frame bound that is possible.

Corollary II.3 ([6]). *Let Φ be a redundant frame for $\mathbb{K}^n$ with singular values $(\sigma_i)_{i=1}^n$. Fix an upper frame bound satisfying $B^\Psi \geq \frac{1}{\sigma_n^2}$ and a lower frame bound $\frac{1}{\sigma_{m-n+1}^2} \geq A^\Psi \geq \frac{1}{\sigma_1^2}$, where we use the convention $\frac{1}{\sigma_{m-n+1}} = \infty$ if $m \geq 2n$. Then a dual frame Ψ of Φ with these frame bounds can be found constructively using a sequence of orthogonalization procedures.*

We are now ready to state the complete characterization of the possible spectra of dual frames.

Theorem II.4 ([6]). *Let $n, m \in \mathbb{N}$, and set $r = m - n$. Let Φ be a frame for $\mathbb{K}^n$ with singular values $(\sigma_i)_{i=1}^n$. Suppose Ψ is any dual frame with singular values $(\sigma_i^\Psi)_{i=1}^n$ (also arranged in a non-increasing order). Then the following inequalities hold:*

$$\frac{1}{\sigma_{n-i+1}} \leq \sigma_i^\Psi \qquad \text{for } i = 1, \dots, r, \tag{5}$$

$$\frac{1}{\sigma_{n-i+1}} \leq \sigma_i^\Psi \leq \frac{1}{\sigma_{n-i+r+1}} \qquad \text{for } i = r+1, \dots, n. \tag{6}$$

Furthermore, for every sequence $(\sigma_i^\Psi)_{i=1}^n$ which satisfies (5) *and* (6)*, there is a dual Ψ of Φ with singular values $(\sigma_i^\Psi)_{i=1}^n$.*

The necessity of the conditions in Theorem II.4 follows by r applications of [4, Theorem 7.3.9] on the matrix M_Ψ defined in (3) or from the well-known interlacing inequalities for Hermitian matrices by Weyl. For the existence part, we refer to the proof in [6].

The inequalities (5) and (6), written in terms of the singular values $(\sigma_i^{\widetilde{\Psi}})_{i=1}^n$ of the canonical dual frame $\widetilde{\Psi} := S^{-1}\Phi$, have the following simple form:

$$\sigma_i^{\widetilde{\Psi}} \leq \sigma_i^\Psi \qquad \text{for } i = 1, \dots, r,$$

$$\sigma_i^{\widetilde{\Psi}} \leq \sigma_i^\Psi \leq \sigma_{i-r}^{\widetilde{\Psi}} \qquad \text{for } i = r+1, \dots, n.$$

In terms of eigenvalues of frame operators, Theorem II.4 states that the spectra in the set of all duals exhaust the set $\Lambda \subset \mathbb{R}^n$ defined by

$$\Lambda = \left\{ (\lambda_i) \in \mathbb{R}^n : \lambda_i^{\widetilde{\Psi}} \leq \lambda_i \leq \lambda_{i-r}^{\widetilde{\Psi}} \text{ for all } i = 1, \dots, n \right\},$$

where $\lambda_i^{\widetilde{\Psi}} = 1/\lambda_{n-i+1}^\Phi$ is the ith eigenvalue of the canonical dual frame operator; we again use the convention that $\lambda_i^{\widetilde{\Psi}} = \infty$ for $i \leq 0$. By considering the trace of $M_\Psi M_\Psi^*$, we see that the canonical dual frame is the unique dual frame that

minimizes the inequalities in Λ. Therefore, the canonical dual is a minimizer among all duals for any matrix norm related to the spectrum of an operator. In general, it is only a unique minimizer if the matrix norm involves all singular values. Moreover, any other spectrum in Λ will not be associated with a unique dual frame, in particular, if $s_i = (s_{i,1}, \dots, s_{i,r}) \neq 0$ in M_Ψ for some $i = 1, \dots, n$, then replacing s_i by zs_i for any $|z| = 1$ will yield a dual frame with unchanged spectrum.

For a better understanding of the more general framework where Theorem II.2 does not yield a complete characterization of the possible spectral patterns, we will continue by a discussion of the example of a frame of three vectors in $\mathbb{R}^2$.

Example 1. Suppose Φ is a frame in $\mathbb{R}^2$ with 3 frame vectors and frame bounds $0 < A^\Phi \leq B^\Phi$, and let $\Phi = U\Sigma_\Phi V^*$ be the SVD of Φ. Then all dual frames are given as $\Psi = UM_\Psi V^*$, where

$$M_\Psi = \begin{bmatrix} 1/\sigma_1 & 0 & s_1 \\ 0 & 1/\sigma_2 & s_2 \end{bmatrix}$$

for $s_1, s_2 \in \mathbb{R}$. Since the frame operator of the dual frame is given by $S_\Psi = \Psi\Psi^* = UM_\Psi M_\Psi^* U^*$, we can find the eigenvalues of S_Ψ by considering eigenvalues of

$$S := M_\Psi M_\Psi^* = \begin{bmatrix} 1/\sigma_1^2 + s_1^2 & s_1 s_2 \\ s_1 s_2 & 1/\sigma_2^2 + s_2^2 \end{bmatrix}.$$

These are given by

$$\lambda_{1,2} = \frac{1}{2} \operatorname{tr} S \pm \frac{1}{2} R, \quad \text{where} \quad R = \sqrt{(\operatorname{tr} S)^2 - 4 \det S}.$$

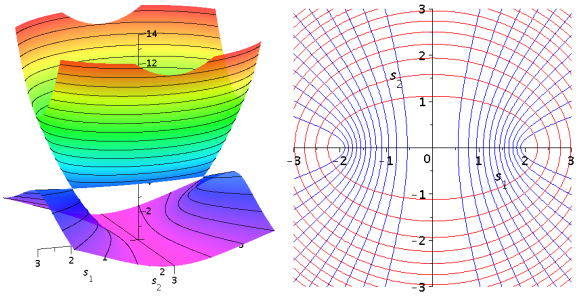
One easily sees that $\operatorname{tr} S$ monotonically grows as a function of $s_1^2 + s_2^2$, whereas for fixed $\operatorname{tr} S$, the term R grows as a function of $s_1^2 - s_2^2$. This exactly yields the two degrees of freedom predicted by the existence part of Theorem II.4. A straightforward calculation shows that $R + (s_1^2 + s_2^2) \geq \frac{1}{\sigma_2^2} - \frac{1}{\sigma_1^2} \geq 0$, hence we see that

$$\lambda_1 \geq \tfrac{1}{\sigma_2^2} \quad \text{and} \quad \tfrac{1}{\sigma_2^2} \geq \lambda_2 \geq \tfrac{1}{\sigma_1^2},$$

which is also the conclusion of the necessity part of Theorem II.4. We remark that the two eigenvalues depend only on quadratic terms of the form s_1^2 and s_2^2. Therefore, if s_1 and s_2 are non-zero, then the choices $(\pm s_1, \pm s_2)$ yield four different dual frames having the same eigenvalues. In this case the level sets of λ_1 as a function of (s_1, s_2) are origin-centered ellipses with major and minor axes in the s_1 and s_2 direction, respectively. Moreover, the semi-major axis is always greater than $s_0 := (\sigma_2^{-2} - \sigma_1^{-2})^{1/2}$. The level sets of λ_2 are origin-centered, East-West opening hyperbolas with semi-major axes greater than s_0. In Figure 1 the possible eigenvalues of the dual frame operator of the frame Φ defined by

$$\Phi = \frac{1}{50} \begin{bmatrix} 90 & -12 & -16 \\ 120 & 9 & 6 \end{bmatrix} \tag{7}$$

are shown as a function of the two parameters s_1 and s_2; Figure 1b shows the level sets and the four intersection points $(\pm s_1, \pm s_2)$ for each allowed spectrum in the interior of Λ. Note that the singular values are $\sigma_1 = 3$ and $\sigma_2 = 1/2$, hence $B^\Phi = 9$ and $A^\Phi = 1/4$.



(a) Graphs of λ_1 and λ_2 (b) Level curves of λ_1 and λ_2

Figure 1: The lower and upper frame bounds of dual frames Ψ (to Φ defined in (7)) as a function of s_1 and s_2. The two graphs in (a) meet at $s_1 = \pm\sqrt{35}/3$ and $s_2 = 0$ which correspond to tight dual frames.

When the difference between the singular values of Ψ goes to zero, the ellipses degenerate to a line segment (or even to a point if $\sigma_1 = \sigma_2$). The limiting case corresponds to tight dual frames so Theorem II.1(ii) applies, and we are forced to set $s_2 = 0$ to achieve row orthogonality of M_Ψ. We then need to pick s_1 such that the two row norms of M_Ψ are equal, thus

$$|s_1| = \sqrt{\frac{1}{\sigma_2^2} - \frac{1}{\sigma_1^2}} = \sqrt{\frac{1}{A^\Phi} - \frac{1}{B^\Phi}} = s_0,$$

which shows that the above lower bound for the semi-major axis is sharp.

References

[1] P. Casazza, G. Kutyniok, eds., *Finite Frames: Theory and Applications*, Birkhäuser, Boston, 2012.

[2] O. Christensen, Frames and Bases: An Introductory Course. Birkhäuser, Boston, 2008.

[3] V. Goyal, J. Kovačević, *Quantized Frame Expansions with Erasures* Appl. Comput. Harmon. Anal. **10** (2001), 203–233.

[4] R. A. Horn, C. R. Johnson, *Matrix analysis*, Cambridge University Press, Cambridge, 1985.

[5] F. Krahmer, G. Kutyniok, J. Lemvig, *Quantitative comparison of dual frames from the computational viewpoint*, in preparation.

[6] F. Krahmer, G. Kutyniok, J. Lemvig, *Sparsity and spectral properties of dual frames*, Linear Algebra Appl. in press, http://dx.doi.org/10.1016/j.laa.2012.10.016.

[7] M. Lammers, A. M. Powell, Ö. Yılmaz, *Alternative dual frames for digital-to-analog conversion in sigma-delta quantization*, Adv. Comput. Math. **32** (2010), 73–102.

[8] J. Leng and D. Han, *Optimal dual frames for erasures II*, Linear Algebra Appl. **435** (2011), 1464–1472.

[9] J. Leng, D. Han, T. Huang, *Optimal Dual Frames for Communication Coding With Probabilistic Erasures*, IEEE Trans. Signal Proc. **59**, 2011, 5380–5389.

[10] J. Lopez, D. Han, *Optimal dual frames for erasures*, Linear Algebra Appl. **432** (2010), 471–482.

[11] P. G. Massey, M. A. Ruiz, and D. Stojanoff, *Optimal dual frames and frame completions for majorization*, Appl. Comput. Harmon. Anal., to appear.

Local coherence sampling in compressed sensing

Felix Krahmer*, Holger Rauhut†, Rachel Ward‡,
* Institute for Numerical and Applied Mathematics, University of Göttingen, Germany
† Hausdorff Center for Mathematics, Bonn, Germany
‡ Mathematics Department, University of Texas at Austin, TX, USA

***Abstract*—Sparse recovery guarantees in compressive sensing and related optimization problems often assume incoherence between the 'sensing' and 'sparsity' domains. In practice, incoherence is rarely satisfied due to physical constraints and limitations. Here we discuss the notion of local coherence, and show that by matching the sampling strategy to the local coherence at hand, sparse recovery guarantees extend to a rich new class of sensing problems beyond incoherent systems. We discuss particular applications to compressive MRI imaging and polynomial interpolation.**

I. Introduction

One of the main results in the theory of compressed sensing is that signals which allow for an approximately sparse representation in a suitable basis or dictionary can be recovered from relatively few linear measurements via convex optimization, provided these measurements are sufficiently *incoherent* with the basis in which the signal is sparse.

In practice, incoherence is rarely satisfied due to physical constraints limiting the freedom of the sensing basis. Here we recall the notion of *local coherence*, as introduced in [6] and somewhat implicitly in [5], and summarize coherence-guided sampling strategies and reconstruction guarantees that extend beyond incoherent sampling. In short, local coherence sampling implies that, if Φ is an orthonormal basis from which we can subsample to construct a sensing matrix, and if our signal class is assumed sparse in an alternative orthonormal basis Ψ, then one should sample rows from Φ proportionately to their maximal correlation to any row from Ψ.

We illustrate the power of coherence-based sampling through two examples: compressed sensing imaging and and polynomial interpolation. In compressed sensing imaging, coherence-based sampling provides a theoretical justification for empirical studies [2], [3] pointing to variable-density sampling strategies for improved MRI compressive imaging. In polynomial interpolation, coherence-based sampling implies that sampling points drawn from the Chebyshev distribution are better suited for the recovery of polynomials and smooth functions than uniformly distributed sampling points, aligning with classical results on Lagrange interpolation [4].

II. Notation

Before continuing, let us fix some notation. We will refer to the set of natural numbers $\{1, 2, \dots, N\}$ using the shorthand notation $[N]$. For a vector $x = (x_j) \in \mathbb{C}^N$, the usual ℓ_p vector norm is $\|x\|_p$, and by an abuse of notation, the ℓ_0-"norm" is defined as $\|x\|_0 = \#\{x_j : x_j \neq 0\}$. A vector $x \in \mathbb{C}^N$ is called s-sparse if $\|x\|_0 \leq s$, and the best s-term approximation of a vector $x \in \mathbb{C}^N$ is the s-sparse vector $x_s \in \mathbb{C}^N$ satisfying $x_s = \inf_{u:\|u\|_0 \leq s} \|x - u\|_p$. Clearly, $x_s = x$ if x is s-sparse. Informally, x is called compressible if $\|x - x_s\|$ decays quickly as s increases. Finally, for two nonnegative functions $f(n)$ and $g(n)$ on the natural numbers, we write $f \gtrsim g$ (or $f \lesssim g$) if there exists a constant $C > 0$ such that $f(n) \geq Cg(n)$ (or $f(n) \leq Cg(n)$, respectively) for all $n \in \mathbb{N}$.

III. Incoherent sampling

Here we recall sparse recovery results for structured random sampling schemes corresponding to *bounded orthonormal systems*, of which the partial discrete Fourier transform is a special case. We refer the reader to [7] for an expository article including many references.

Definition 1 (Bounded orthonormal system (BOS)): Let $\mathcal{D}$ be a measurable subset of $\mathbb{R}^d$.

- A set of functions $\{\psi_j : \mathcal{D} \to \mathbb{C}, \quad j \in [N]\}$ is called an *orthonormal system* with respect to the probability measure ν if $\int_{\mathcal{D}} \bar{\psi}_j(u)\psi_k(u)d\nu(u) = \delta_{jk}$, where δ_{jk} denotes the Kronecker delta.
- Let μ be a probability measure on $\mathcal{D}$. A *random sample* of the orthonormal system $\{\psi_j\}$ is the random vector $(\psi_1(U), \dots, \psi_N(U))$ that results from drawing a sampling point U from the measure μ.
- An orthonormal system is said to be *bounded* with bound K if $\sup_{j\in[N]} \|\psi_j\|_\infty \leq K$.

Suppose now that we have an orthonormal system $\{\psi_j\}_{j\in[N]}$ and m random sampling points $U_1, U_2, \dots, U_m$ drawn independently from some probability measure μ. Here and throughout, we assume that the number of sampling points $m \ll N$. As shown in [7], if the system $\{\psi_j\}$ is *bounded*, and if the probability measure μ from which we sample points is the orthogonalization measure ν associated to the system, then the (underdetermined) structured random matrix $A : \mathbb{C}^N \to \mathbb{C}^m$ whose rows are the independent random samples will be well-conditioned, satisfying the so-called *restricted isometry property* [1] with nearly order-optimal restricted isometry constants with high probability. Consequently, matrices associated to random samples of bounded orthonormal systems have nice sparse recovery properties.

Proposition 2 (Sparse recovery through BOS): Consider the matrix $A \in \mathbb{C}^{m\times N}$ whose rows are independent random samples of an orthonormal system $\{\psi_j, j \in [N]\}$ with bound $\sup_{j\in[N]} \|\psi_j\|_\infty \leq K$, drawn from the orthogonalization measure ν associated to the system. If the number of random

samples satisfies

$$m \gtrsim K^2 s \log^3(s) \log(N), \tag{III.1}$$

for some $s \gtrsim \log(N)$, then the following holds with probability exceeding $1 - N^{-C\log^3(s)}$.

For each $x \in \mathbb{C}^N$, given noisy measurements $y = Ax + \sqrt{m}\eta$ with $\|\eta\|_2 \leq \varepsilon$, the approximation

$$x^\# = \arg\min_{z \in \mathbb{C}^N} \|z\|_1 \text{ subject to } \|Az - y\|_2 \leq \sqrt{m}\epsilon$$

satisfies the error guarantee

$$\|x - x^\#\|_2 \lesssim \frac{1}{\sqrt{s}}\|x - x_s\|_1 + \varepsilon.$$

An important special case of such a matrix construction is the *subsampled discrete Fourier matrix*, constructed by sampling $m \ll N$ rows uniformly at random from the unitary discrete Fourier matrix $\Psi \in \mathbb{C}^{N\times N}$ with entries $\psi_{j,k} = \frac{1}{\sqrt{N}} e^{i2\pi(j-1)(k-1)}$. Indeed, the system of complex exponentials $\psi_j(u) = e^{i2\pi(j-1)u}$, $j \in [N]$, is orthonormal with respect to the uniform measure over the discrete set $\mathcal{D} = \{0, \frac{1}{N}, \ldots, \frac{N-1}{N}\}$, and is bounded with optimally small constant $K = 1$. In the discrete setting, we may speak of a more general procedure for forming matrix constructions adhering to the conditions of Proposition 2: given any two unitary matrices Φ and Ψ, the composite matrix $\Phi^*\Psi$ is also a unitary matrix, and this composite matrix will have uniformly bounded entries if the orthonormal bases (ϕ_j) and (ψ_k), corresponding to the rows of Φ and Ψ respectively, are *mutually incoherent:*

$$\mu(\Phi, \Psi) \quad := \sqrt{N}\sup_{1\leq j,k\leq N} |\langle \phi_j, \psi_k \rangle| \leq K \tag{III.2}$$

Indeed, if Φ and Ψ are mutually incoherent, then the rows of $B = \sqrt{N}\Psi^*\Phi$ constitute a bounded orthonormal system with respect to the uniform measure on $\mathcal{D} = \{0, \frac{1}{N}, \ldots, \frac{N-1}{N}\}$. Proposition 2 then implies a sampling strategy for reconstructing signals $x \in \mathbb{C}^N$ with assumed sparse representation in the basis Ψ, that is $x = \Psi b$ and $b \approx b_s$, from a few linear measurements: form a sensing matrix $A \in \mathbb{C}^{m\times N}$ by sampling rows i.i.d. uniformly from an incoherent basis Φ, collect measurements $y = Ax + \eta$, $\|\eta\|_2 \leq \epsilon$, and solve the ℓ_1 minimization program,

$$x^\# = \arg\min_{z \in \mathbb{C}^N} \|\Psi^* z\|_1 \text{ subject to } \|Az - y\|_2 \leq \sqrt{m}\epsilon$$

This scenario is referred to as *incoherent sampling*.

IV. Local coherence sampling

Consider more generally the setting where we aim to compressively sense signals $x \in \mathbb{C}^N$ with assumed sparse representation in the orthonormal basis $\Psi \in \mathbb{C}^{N\times N}$, but our sensing matrix $A \in \mathbb{C}^{m\times N}$ can only consist of rows from some fixed orthonormal basis $\Phi \in \mathbb{C}^{N\times N}$ that is not necessarily incoherent with Ψ. In this setting, we ask: *Given a fixed sensing basis Ψ and sparsity basis Φ, how should we sample rows of Ψ in order to make the resulting system as incoherent as possible?* We will answer this question by introducing the concept of *local coherence* between two bases as described in [5], [6], whereby in the discrete setting the coherences of individual elements of the sensing basis are calculated and used to derive the sampling strategy.

The following result says that regions of the sensing basis that are *more* coherent with the sparsity basis should be sampled with higher density. The following is essentially a generalization of Theorem 2.1 in [5], but for completeness, we include a short self-contained proof.

Theorem 3 (Sparse recovery via local coherence sampling): Consider a measurable set $\mathcal{D}$ and a system $\{\psi_j,\ j \in [N]\}$ that is orthonormal with respect to a measure ν on $\mathcal{D}$ which has square-integrable local coherence,

$$\sup_{j\in[N]} |\psi_j(u)| \leq \kappa(u), \qquad \int_{u\in\mathcal{D}} |\kappa(u)|^2 \nu(u) du = B. \tag{IV.1}$$

We can define the probability measure $\mu(u) = \frac{1}{B}\kappa^2(u)\nu(u)$ on $\mathcal{D}$. Draw m sampling points $u_1, u_2, \ldots, u_m$ independently from the measure μ, and consider the the matrix $A \in \mathbb{C}^{m\times N}$ whose rows are the random samples $\psi_j(u_k), j \in [N]$. Consider also the diagonal preconditioning matrix $\mathcal{P} \in \mathbb{C}^{m\times m}$ with entries $p_{k,k} = 1/\mu(u_k)$. If the number of sampling points

$$m \gtrsim B^2 s \log^3(s) \log(N), \tag{IV.2}$$

for some $s \gtrsim \log(N)$, then the following holds with probability exceeding $1 - N^{-C\log^3(s)}$.

For each $x \in \mathbb{C}^N$, given noisy measurements $y = Ax + \sqrt{m}\eta$ with $\|\mathcal{P}\eta\|_2 \leq \sqrt{m}\varepsilon$, the approximation

$$x^\# = \arg\min_{z \in \mathbb{C}^N} \|z\|_1 \text{ subject to } \|\mathcal{P}Az - \mathcal{P}y\|_2 \leq \sqrt{m}\epsilon$$

satisfies the error guarantee

$$\|x - x^\#\|_2 \lesssim \frac{1}{\sqrt{s}}\|x - x_s\|_1 + \varepsilon$$

Proof: Consider the functions $Q_j(u) = \frac{\sqrt{B}}{\kappa(u)}\psi_j(u)$. The system $\{Q_j\}$ is bounded with $\sup_{j\in[N]} \|Q_j\|_\infty \leq \sqrt{B}$, and this system is orthonormal on $\mathcal{D}$ with respect to the sampling measure μ:

$$\begin{aligned}
&\int_{u\in\mathcal{D}} \bar{Q}_j(u) Q_k(u) \mu(u) du \\
&= \int_{u\in\mathcal{D}} (\frac{1}{\kappa(u)}\bar{\psi}_j(u))(\frac{1}{\kappa(u)}\psi_k(u))(\kappa^2(u)\nu(u)) du \\
&= \int_{u\in\mathcal{D}} \bar{\psi}_j(u)\psi_k(u)\nu(u) du = \delta_{jk}
\end{aligned} \tag{IV.3}$$

Thus we may apply Proposition 2 to the system $\{Q_j\}$, noting that the matrix of random samples of the system $\{Q_j\}$ may be written as $\mathcal{P}A$. ∎

In the discrete setting where $\{\psi_j\}_{j\in[N]}$ and $\{\phi_k\}$ are rows of unitary matrices Ψ and Φ, and ν is the uniform measure over the set $\mathcal{D} = \{0, \frac{1}{N}, \ldots, \frac{N-1}{N}\}$, the integral in condition IV.1 reduces to a sum,

$$\sup_{k\in[N]} \sqrt{N}|\langle \psi_j, \phi_k \rangle| \leq \kappa_j, \quad \frac{1}{N}\sum_{j=1}^{N} \kappa_j^2 = B. \tag{IV.4}$$

This motivates the introduction of the local coherence of an orthonormal basis $\{\phi_j\}_{j=1}^N$ of $\mathbb{C}^N$ with respect to the orthonormal basis $\{\psi_k\}_{k=1}^N$ of $\mathbb{C}^N$ is the function $\mu^{loc} = (\mu_j) \in \mathbb{R}^N$ defined coordinate-wise by

$$\mu_j = \sup_{1\le k\le N} \sqrt{N}|\langle \varphi_j, \psi_k\rangle|.$$

We have the following corollary of Theorem 3.

Corollary 4: Consider a pair of orthonormal basis (Φ, Ψ) with local coherences bounded by $\mu_j \le \kappa_j$. Let $s \ge 1$, and suppose that

$$m \gtrsim s(\frac{1}{N}\sum_{j=1}^N \kappa_j^2)\log^4(N).$$

Select m (possibly not distinct) rows of Φ^* independent and identically distributed from the multinomial distribution on $\{1, 2, \dots, N\}$ with weights $c\kappa_j^2$ to from the sensing matrix $A : \mathbb{C}^N \to \mathbb{C}^m$. Consider also the diagonal preconditioning matrix $\mathcal{P} \in \mathbb{C}^{m\times m}$ with entries $p_{k,k} = \frac{1}{\sqrt{c}\kappa_j}$.

Then the following holds with probability exceeding $1 - N^{-C\log^3(s)}$.

For each $x \in \mathbb{C}^N$, given measurements $y = Ax + \eta$, with $\|\mathcal{P}\eta\|_2 \le \sqrt{m}\varepsilon$, the approximation

$$x^\# = \arg\min_{u\in\mathbb{C}^N} \|\Psi^* u\|_1 \text{ subject to } \|y - \mathcal{P}Au\|_2 \le \sqrt{m}\varepsilon$$

satisfies the error guarantee

$$\|x - x^\#\|_2 \lesssim \frac{1}{\sqrt{s}}\|\Psi^* x - (\Psi^* x)_s\|_1 + \varepsilon.$$

Remark 5: Note that the local coherence not only influences the embedding dimension m, it also influences the sampling measure. Hence a priori, one cannot guarantee the optimal embedding dimension if one only has suboptimal bounds for the local coherence. That is why the sampling measure in Theorem 3 is defined via the (known) upper bounds κ and $\|\kappa\|_2$ rather than the (usually unknown) exact values μ_{loc} and $\|\mu_{loc}\|_2$, showing that local coherence sampling is *robust with respect to the sampling measure:* suboptimal bounds still lead to meaningful bounds on the embedding dimension.

We now present two applications where incoherent sampling fails, but local coherence sampling provides a sampling scheme with sparse recovery guarantees.

V. Applications

A. Variable-density sampling for compressed sensing MRI

In Magnetic Resonance Imaging, after proper discretization, the unknown image (x_{j_1,j_2}) is a two-dimensional array in $\mathbb{R}^{n\times n}$, and allowable sensing measurements are two-dimensional Fourier transform measurements:

$$\phi_{k_1,k_2} = \frac{1}{n}\sum_{j_1,j_2} x_{j_1,j_2} e^{2\pi i(k_1 j_1 + k_2 j_2)/n}, \quad -n/2{+}1 \le k_1, k_2 \le n/2$$

Natural sparsity domains for images, such as discrete spatial differences, are not incoherent to the Fourier basis.

A number of empirical studies, including the very first papers on compressed sensing MRI , observed that image reconstructions from compressive frequency measurements could be significantly improved by variable-density sampling.

Note that lower frequencies are more coherent with wavelets and step functions than higher frequencies. In [6], the local coherence between the two-dimensional Fourier basis and bivariate Haar wavelet basis was calculated:

Proposition 6: The local coherence between frequency ϕ_{k_1,k_2} and the bivariate Haar wavelet basis $\Psi = (\psi_I)$ can be bounded by

$$\mu(\phi_{k_1,k_2}, \Psi) \lesssim \frac{\sqrt{N}}{(|k_1+1|^2 + |k_2+1|^2)^{1/2}}$$

Note that this local coherence is *almost square integrable independent of discretization size* n^2, as

$$\frac{1}{N}\sum_{j=1}^N \mu_j^2 \lesssim \log(n).$$

Applying Corollary 4 to compressive MRI imaging, we then have

Corollary 7: Let $n \in \mathbb{N}$. Let Ψ be the bivariate Haar wavelet basis and let $\Phi = (\phi_{k_1,k_2})$ be the two-dimensional discrete Fourier transform. Let $s \ge 1$, and suppose that $m \gtrsim s(\frac{1}{N}\log^5(N)$. Select m (possibly not distinct) frequencies (ϕ_{k_1,k_2}) independent and identically distributed from the multinomial distribution on $\{1, 2, \dots, N\}$ with weights proportional to the inverse squared Euclidean distance to the origin, $\frac{1}{(|k_1+1|^2+|k_2+1|^2)}$, and form the sensing matrix $A : \mathbb{C}^N \to \mathbb{C}^m$.

Then the following holds with probability exceeding $1 - N^{-C\log^3(s)}$.

For each image $x \in \mathbb{C}^{n\times n}$, given measurements $y = Ax$, the approximation

$$x^\# = \arg\min_{u\in\mathbb{C}^{n\times n}} \|\Psi^* u\|_1 \text{ subject to } \|\mathcal{D}y - Au\|_2 \le \epsilon$$

satisfies the error guarantee

$$\|x - x^\#\|_2 \lesssim \frac{1}{\sqrt{s}}\|\Psi^* x - (\Psi^* x)_s\|_1 + \varepsilon.$$

Numerical results such as those detailed in [**?**] and illustrated below in Figure 1 confirm that variable-density sampling strategies significantly outperform uniform sampling strategies as well as deterministic sampling strategies, and Corollary 7 provides theoretical justification for such observations. Below we provide a numerical comparison of various sampling strategies, including the sampling distribution given in Corollary 7. The following images were made from total variation minimization rather than Haar wavelet minimization, but the theory for Fourier-Wavelet sampling is extended to the total variation minimization setting in [6].

B. Sparse Legendre expansions for smooth function interpolation

Here we consider the problem of recovering polynomials g from m sample values $g(x_1), g(x_2) \dots, g(x_m)$, with sampling points $x_\ell \in [-1, 1]$ for $\ell = 1, \dots, m$. If the number of

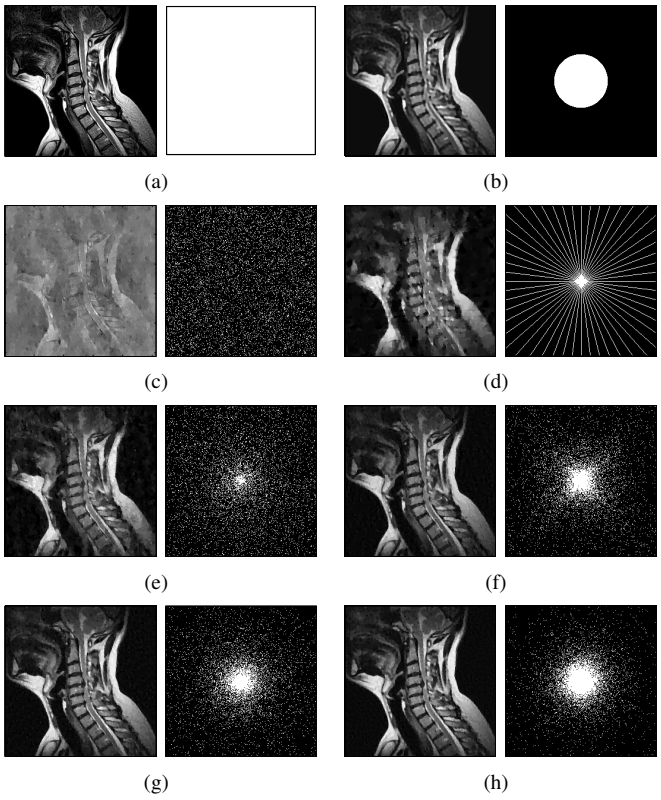


Fig. 1. Various reconstructions of an MRI image $x \in \mathbb{R}^{256\times 256}$ with total variation minimization from $m = 6400$ noiseless partial DFT measurements sampled from various distributions. Beside each reconstruction is a plot of frequency space $\{(k_1, k_2) : -N/2 + 1 \le k_1, k_2 \le N/2\}$ and the frequencies used in its reconstruction. (a) Original image. (b) Reconstruction using only lowest frequencies: $\Omega = \{(k_1, k_2) : k_1^2 + k_2^2 \le 80\}$. (c) Prob$\big((k_1, k_2) \in \Omega\big) \sim 1$ (Uniform subsampling) (d) Ω comprised of frequencies in equispaced radial lines. (e) Prob$\big((k_1, k_2) \in \Omega\big) \propto (k_1^2+k_2^2+1)^{-1/2}$(f) Prob$\big((k_1, k_2) \in \Omega\big) \propto \big(\max(|k_1|, |k_2|)+1\big)^{-1}$ (g) Prob$\big((k_1, k_2) \in \Omega\big) \propto (k_1^2+k_2^2+1)^{-1}$. (h) Prob$\big((k_1, k_2) \in \Omega\big) \propto (k_1^2 + k_2^2 + 1)^{-3/2}$. The relative reconstruction error $\|f - f^{\#}_{TV}\|_2/\|f\|_2$ corresponding to each reconstruction is (b) .2932, (c) .8229, (d) .4074, (e) .3192, (f) .2603, (g) .2537, and (h) .2463.

sampling points is less or equal to the degree of g, then in general such reconstruction is impossible due to dimension reasons. However, the situation becomes tractable if we make a sparsity assumption. In order to introduce a suitable notion of sparsity, we consider the orthonormal basis of Legendre polynomials.

Definition 8: The (orthonormal) Legendre polynomials

$$P_0, P_1, \ldots, P_n, \ldots$$

are uniquely determined by the following conditions:

- $P_n(x)$ is a polynomial of precise degree n in which the coefficient of x^n is positive,
- the system $\{P_n\}_{n=0}^{\infty}$ is orthonormal with respect to the normalized Lesbegue measure on $[-1, 1]$:

$$\frac{1}{2}\int_{-1}^{1} P_n(x)P_m(x)dx = \delta_{n,m}, \quad n, m = 0, 1, 2, \ldots$$

Since the interval $[-1, 1]$ is symmetric, the Legendre polynomials satisfy $P_n(x) = (-1)^n P_n(-x)$. For more information see [Szego].

An arbitrary real-valued polynomial g of degree $N-1$ can be expanded in terms of Legendre polynomials,

$$g(x) = \sum_{j=0}^{N-1} c_j P_j(x), \quad x \in [-1, 1]$$

with coefficient vector $c \in \mathbb{R}^N$. The vector is s-sparse if $\|c\|_0 \le s$. Given a set of m sampling points $(x_1, x_2, \ldots, x_m)$, the samples $y_k = g(x_k)$, $k = 1, \ldots, m$, may be expressed concisely in terms of the coefficient vector according to

$$y = \Phi c,$$

where $\phi_{k,j} = P_j(x_k)$. If the sampling points $x_1, \ldots, x_m$ are random variables, then the matrix $\Phi \in \mathbb{R}^{m\times N}$ is exactly the sampling matrix corresponding to random samples from the Legendre system $\{P_j\}_{j=1}^{N}$. This is not a bounded orthonormal system, however, as the Legendre polynomials grow like

$$|P_n(x)| \le (n + 1/2)^{1/2}, \quad -1 \le x \le 1.$$

Nevertheless the Legendre system does have bounded local coherence. A classic result [szego] follows.

Proposition 9: For all $n > 0$ and for all $x \in [-1, 1]$,

$$|P_n(x)| < \kappa(x) = 2\pi^{-1/2}(1 - x^2)^{-1/4}.$$

here, the constant is 2 $\pi^{-1/2}$ cannot be replaced by a smaller one.

Indeed, $\kappa(x)$ is a square integrable function proportional to the Chebyshev measure $\pi^{-1}(1 - x^2)^{-1/2}$. We arrive at the following result for Legendre polynomial interpolation as a corollary of Theorem 3.

Corollary 10: Let $x_1, \ldots, x_m$ be chosen independently at random on $[-1, 1]$ according to the Chebyshev measure $\pi^{-1}(1 - x^2)^{-1/2}dx$. Let Ψ be the matrix with entries $\Psi_{k,j} = \sqrt{\pi/2}(1 - x_k^2)^{1/4}P_n(x_k)$. Suppose that

$$m \gtrsim s \log^3$$

Consider the matrix $A \in \mathbb{C}^{m\times N}$ whose rows are independent random vectors$(\psi_j(X_k))$ drawn from the measure μ. If

$$m \gtrsim B^2 s \log^3(s) \log(N), \tag{V.1}$$

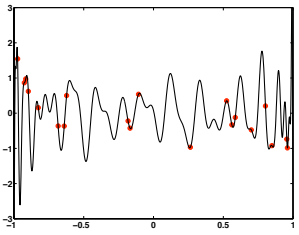
for some $s \gtrsim \log(N)$, then the following holds with probability exceeding $1 - N^{-C\log^3(s)}$. Let $\mathcal{D} \in \mathbb{C}^{m\times m}$ be the diagonal matrix with entries $d_{k,k} = \frac{1}{\mu(X_k)}$. For each $x \in \mathbb{C}^N$, given noisy measurements $y = Ax + \sqrt{m}\eta$ with $\|\mathcal{D}\eta\|_2 \le \sqrt{m}\varepsilon$, the approximation

$$x^{\#} = \arg\min_{u\in\mathbb{C}^N} \|u\|_1 \text{ subject to } \|\mathcal{D}Au - \mathcal{D}y\|_2 \le \sqrt{m}\epsilon$$

satisfies the error guarantee

$$\|x - x^{\#}\|_2 \lesssim \frac{1}{\sqrt{s}}\|x - x_s\|_1 + \varepsilon$$

where x_s is the best s-term approximation to x.



We illustrate exact recovery of a Legendre sparse polynomial from randomly sampled points from the Chebyshev measure.

In fact, more general theorems exist: the Chebyshev measure is a universal sampling strategy for interpolation with any set of orthogonal polynomials [5].

An extension to the setting of interpolation with spherical harmonics can be found in [5], [**?**].

VI. Conclusion

Here we summarize local coherence sampling, and demonstrate its power for generalized sparse recovery results in compressed sensing in two seemingly disparate settings - MRI compressive imaging and Legendre polynomial interpolation. Unlike incoherence-based results, local coherence sampling gives a sampling strategy for fixed sparsity basis and fixed sensing basis from which one can subsample; if the local coherence function is square integrable and this integral depends only mildly on the ambient dimension of the signal, then stable and robust sparse recovery results for incoherent sampling generalize to this setting. Several questions remain, such as the optimality of the local coherence sampling, extensions to frames rather than orthonormal dictionaries, and connections to designing sensing matrices via minimizing the local coherence [**?**].

Acknowledgment

The authors would like to thank Wotao Yin, Mark Tygert for helpful discussions related to this paper.

References

[1] E. J. Candes, J. Romberg, and T. Tao. *Stable signal recovery from incomplete and inaccurate measurements.* Comm. Pure Appl. Math., 59(8):1207–1223, 2006.

[2] M. Lustig, D. Donoho, and J.M. Pauly, *Sparse MRI: the application of compressed sensing for rapid MRI imaging*, Magnetic Resonance in Medicine, 58(6):1182-1195, 2007.

[3] M. Lustig, D.L. Donoho, J.M. Santos, and J.M. Pauly. *Compressed sensing MRI*. IEEE Sig. Proc. Mag., 25(2):72-82, 2008.

[4] L. Brutman. *Lebesgue functions for polynomial interpolatio - a survey.* Ann. Numer. Math., 4 (1-4): 111– 127, 1997.

[5] H. Rauhut and R. Ward, *Sparse Legendre expansions via ℓ_1-minimization*, Journal of Approximation Theory, 164:517–533, 2012.

[6] F. Krahmer and R. Ward, *Beyond incoherence: Stable and robust sampling strategies for compressive imaging*, Preprint, 2012.

[7] H. Rauhut, *Compressive sensing and structured random matrices*, In M. Fornasier, editor, *Theoretical Foundations and Numerical Methods for Sparse Recovery*, Volume 9 of *Radon Series Comp. Appl. Math.*, pages 1–92. deGruyter, 2010. 0.5em minus 0.4emHarlow, England: Addison-Wesley, 1999.

Structured-signal recovery from single-bit measurements

Yaniv Plan
Department of Mathematics
University of Michigan, Ann Arbor, Michigan 48104
Email: yplan@umich.edu

***Abstract*—1-bit compressed sensing was introduced by Boufounos and Baraniuk in 2008 as a model of extreme quantization; only the sign of each measurement is retained. Recent theoretical and algorithmic advances, combined with the ease of hardware implementation, show that it is an effective method of signal acquisition. Surprisingly, in the high-noise regime there is almost no information loss from 1-bit quantization. We review and revise recent results, and compare to closely related statistical problems: sparse binary regression and binary matrix completion.**

I. Introduction

Discrete measurements arise both in signal processing and statistical inference, but for different reasons. In some cases, they are inherent to the data—consider a statistical experiment in which the response is a binary variable indicating the presence or absence of a certain disease. In other cases the level of discretization is chosen—consider quantization in analog-to-digital conversion. We focus on the extreme case in which all measurements are binary. For further signal-processing motivation, see [1].

It turns out that the abstract statistical models and signal-processing models nearly match, but with subtle differences that have strong influence on the methods of signal reconstruction and the theoretical challenges. We point out these differences and how the ideas from 1-bit compressed sensing allow new methods and results in binary regression.

In Section II, we describe recent results in *1-bit compressed sensing* and give connections to standard *compressed sensing*. These methods allow for a new semi-parametric approach to *sparse binary regression*, described in Section III. In Section IV we describe modern theoretical results in binary PCA with missing entries, or *binary matrix completion*.

II. 1-bit compressed sensing

Unquantized compressed sensing [7] concerns the reconstruction of sparse signals from linear measurements. Let $\|\boldsymbol{x}\|_0$ give the number of nonzero entries of $\boldsymbol{x}$. We assume that $\|\boldsymbol{x}\|_0 \leq s$ i.e., $\boldsymbol{x}$ is sparse. One observes data of the form

$$y_i = \langle \boldsymbol{a}_i, \boldsymbol{x} \rangle \qquad i = 1, \dots, m$$

and would like to reconstruct $\boldsymbol{x} \in \mathbb{R}^n$ from $\{y_i, \boldsymbol{a}_i\}$.

Supported by NSF Postdoctoral Research Fellowship under award No. 1103909.

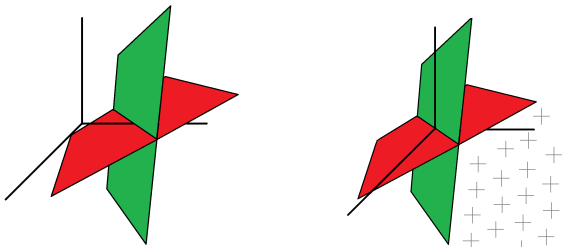


Fig. 1: On left: Linear measurements $y_1 = \langle \boldsymbol{a}_1, \boldsymbol{x} \rangle$ and $y_2 = \langle \boldsymbol{a}_2, \boldsymbol{x} \rangle$ determine that $\boldsymbol{x}$ must lie in the intersection of the two hyperplanes. On right: Single bit measurements $y_1 = \text{sign}(\langle \boldsymbol{a}_1, \boldsymbol{x} \rangle)$ and $y_2 = \text{sign}(\langle \boldsymbol{a}_2, \boldsymbol{x} \rangle)$ determine that $\boldsymbol{x}$ must lie in the region denoted by $+$ signs.

In 1-bit compressed sensing [4], only the sign of each measurement is retained:

$$y_i = \text{sign}(\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle) \qquad i = 1, \dots, m.$$

Above $\text{sign}(t) = 1$ if $t \geq 0$ and $\text{sign}(t) = -1$ it $t < 0$. In matrix form,

$$\boldsymbol{y} = \text{sign}(\boldsymbol{A}\boldsymbol{x})$$

where $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ is a matrix whose i-th row is equal to $\boldsymbol{a}_i$, and we allow the sign function to act on a vector by acting on each individual entry.

There is a stark geometric difference between these two observation models. In unquantized compressed sensing, each measurement determines a hyperplane in which $\boldsymbol{x}$ must reside. In 1-bit compressed sensing, each measurement determines a hyperplane, but now we are only told which side of the hyperplane $\boldsymbol{x}$ resides on (see Figure 1).

Do the 1-bit measurements contain sufficient information to reconstruct $\boldsymbol{x}$? Clearly, exact reconstruction is impossible because the measurements only give a finite number of bits of information and the signal lies in an infinite set. Furthermore, the measurements retrieve *no information* about the norm of $\boldsymbol{x}$. Thus, we may only hope to approximate the direction of $\boldsymbol{x}$. Equivalently, we assume that $\boldsymbol{x} \in S^{n-1}$ and endeavor to approximate $\boldsymbol{x}$ itself.

A natural method to reconstruct $\boldsymbol{x}$ is to find a vector that

matches the data and has the required structure:

$$\text{Find } \boldsymbol{x}' \quad \text{such that} \quad \|\boldsymbol{x}'\|_0 \leq s, \|\boldsymbol{x}'\|_2 = 1 \quad \text{and} \quad \operatorname{sign}(\boldsymbol{A}\boldsymbol{x}') = y. \tag{1}$$

This program has recently been shown to give nearly optimal accuracy. If $\boldsymbol{A}$ is a Gaussian matrix, Jacques et al. [10] show that $O(\delta^{-1}s\log(n/\delta))$ measurements are sufficient to reconstruct $\boldsymbol{x}$ with ℓ_2 error at most δ. Aside from the logarithmic factor, Theorem 1 in [10] shows that this error bound in nearly minimax. It is further shown that a variation on this program provides stability to adversarial noise. Yet there still remain important challenges because the above program contains two nonconvex constraints: $\|\boldsymbol{x}\|_0 \leq s$ and $\|\boldsymbol{x}\|_2 = 1$. Thus, there is no known algorithm that is guaranteed to return the solution to the above program in polynomial time.

In order to give a polynomial-time solver, Plan and Vershynin [17] propose a convex programming approach:

$$\min_{\boldsymbol{x}'} \|\boldsymbol{x}'\|_1 \quad \text{such that} \quad \|\boldsymbol{A}\boldsymbol{x}'\|_1 = 1 \quad \text{and} \quad \operatorname{sign}(\boldsymbol{A}\boldsymbol{x}') = y. \tag{2}$$

Above, $\|\boldsymbol{A}\boldsymbol{x}\|_1 = \sum_i |\langle \boldsymbol{a}_i, \boldsymbol{x}\rangle| = \sum_i y_i \langle \boldsymbol{a}_i, \boldsymbol{x}\rangle$ is a linear constraint; in fact, the program can be recast as a linear program. Let $\hat{\boldsymbol{x}}$ be the solution to the above program. Theorem 1.1 in [17] shows that

$$\left\| \frac{\hat{\boldsymbol{x}}}{\|\hat{\boldsymbol{x}}\|_2} - \boldsymbol{x} \right\|_2 \leq \delta$$

with high probability provided that $m \geq O(\delta^{-5}s\log^2(n/s))$. We leverage recent results on discrete embeddings [16] to give a slight refinement of this result.

Theorem 1. *Let $s \leq m \leq n$. Let $\boldsymbol{A}$ have i.i.d. standard normal entries. Suppose that*

$$m \geq C\delta^{-4}s\log^2(n/s).$$

Then, with probability at least $1 - C_1\exp(-c\delta m)$ the following holds uniformly over all signals $\boldsymbol{x} \in \mathbb{R}^n$ satisfying $\|\boldsymbol{x}\|_1 \leq \sqrt{s}, \|\boldsymbol{x}\|_2 = 1$. Let $\boldsymbol{y} = \operatorname{sign}(\boldsymbol{A}\boldsymbol{x})$. Then the solution $\hat{\boldsymbol{x}}$ to the linear program (2) *satisfies*

$$\left\| \frac{\hat{\boldsymbol{x}}}{\|\hat{\boldsymbol{x}}\|_2} - \boldsymbol{x} \right\|_2 \leq \delta.$$

Above, and in what follows, C and c are absolute numeric constants.

Proof: Proceed as in the proof of Theorem 1.1 in [17], but replace Theorem 2.1 in [17] with Theorem 3.1 in [16]. ■

Remark 1 (Soft sparsity). The assumption that $\|\boldsymbol{x}\|_1 \leq \sqrt{s}$ is a relaxation of the exact sparsity constraint $\|\boldsymbol{x}\|_0 \leq s$. Indeed, suppose that $\|\boldsymbol{x}\|_0 \leq s$. Then by the Cauchy-Schwarz inequality,

$$\|\boldsymbol{x}\|_1 \leq \sqrt{\|\boldsymbol{x}\|_0} \cdot \|\boldsymbol{x}\|_2 = \sqrt{\|\boldsymbol{x}\|_0} \leq \sqrt{s}.$$

However, the constraint $\|\boldsymbol{x}\|_1 \leq \sqrt{s}$ allows for $\boldsymbol{x}$ to be *compressible* instead of exactly sparse—it only requires a fast decay rate of the entries of $\boldsymbol{x}$.

Remark 2 (Optimality and δ dependence). Up to the power of 2 on the logarithm, the number of measurements required for a *fixed* level of accuracy matches what is needed for *unquantized* compressed sensing, and also matches the error bound achieved by the non-convex program (1). Let us also consider the dependence of m on δ and compare to the solution to the non-convex program. If $\hat{\boldsymbol{x}}$ is the solution to (1) the number of measurements required is essentially proportional to δ^{-1}. If $\hat{\boldsymbol{x}}$ is the solution to the convex program (2), Theorem 1 requires m to be proportional to δ^{-4}. On one hand, the former requires exact sparsity while the latter softens this requirement. Further, as shown in [15], in the noisy problem and with *soft sparsity* the δ^{-4} dependence is sometimes optimal. Nevertheless, in the noiseless problem it is an open problem whether the δ dependence can be improved for an efficient solver.

For adaptive approaches to 1-bit compressed sensing with impressive reconstruction guarantees, see [8], [9].

III. Noisy 1-bit compressed sensing

In noisy 1-bit compressed sensing, the data takes the form

$$\boldsymbol{y} = \operatorname{sign}(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{z}) \tag{3}$$

where $\boldsymbol{z}$ is a noise term with i.i.d. entries.

In order to reconstruct $\boldsymbol{x}$, one would like to soften the constraint, $\operatorname{sign}(\boldsymbol{A}\boldsymbol{x}) = y$ used in the noiseless problem. A natural way to do this would be to bound the Hamming distance between $\operatorname{sign}(\boldsymbol{A}\boldsymbol{x})$ and $\boldsymbol{y}$. Unfortunately, this would give a non-convex constraint. Thus, Plan and Vershynin [15] suggest a different convex program to estimate $\boldsymbol{x}$:

$$\max_{\boldsymbol{x}'} \sum_i y_i \langle \boldsymbol{a}_i, \boldsymbol{x}'\rangle \quad \text{such that} \quad \|\boldsymbol{x}'\|_2 \leq 1, \|\boldsymbol{x}'\|_1 \leq \sqrt{s}. \tag{4}$$

The solution enjoys a high level of accuracy.

Theorem 2 ([15], Corollary 3.1)**.** *Fix $\boldsymbol{x} \in S^{n-1}$ satisfying $\|\boldsymbol{x}\|_1 \leq \sqrt{s}$. Let $\boldsymbol{A}$ have independent standard normal entries. Let $\boldsymbol{y} = \operatorname{sign}(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{z})$ and suppose that $\boldsymbol{z}$ is a Gaussian noise vector with independent $N(0, \sigma^2)$ entries. Let $\delta > 0$ and suppose that*

$$m \geq C\delta^{-4}(\sigma^2 + 1)s\log(2n/s).$$

Then, with probability at least $1 - 8\exp(-c\delta^4 m)$, the solution $\hat{\boldsymbol{x}}$ to the convex program (4) *satisfies*

$$\|\hat{\boldsymbol{x}} - \boldsymbol{x}\|_2 \leq \delta.$$

In contrast to Theorem 1, this is a non-uniform result—it holds for one fixed $\boldsymbol{x}$ with a random draw of $\boldsymbol{A}$. See Theorem 1.3 in [15] for a uniform result which also considers adversarial noise and for the treatment of much more general signal structures outside of sparsity. We also note that when $\sigma > 1$ this error bound nearly matches the minimax error bound achievable by any estimator from *unquantized measurements*. See [18, Theorem 1] and [15, Section 3]. This has the following implication: *When the signal-to-noise ratio is low, 1-bit measurements contain almost as much information*

as unquantized measurements. The preceding theoretical conclusion is backed up with numerical evidence in [12].

A. Sparse binary regression

Sparse binary regression, and in particular sparse logistic regression, are often used to explain statistical data in which the response variable is binary. It is common to assume that the data is generated according to the *generalized linear model:* $y_i \in \{+1, -1\}$ is a Bernoulli random variable satisfying

$$\mathbb{E}\, y_i = \theta(\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle) \tag{5}$$

for some function $\theta : \mathbb{R} \to [0, 1]$. Note that this model implies that

$$P(y_i = 1) = \frac{\theta(\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle) + 1}{2} =: f(\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle).$$

Thus, the noisy 1-bit compressed sensing model 3 can always be recast using the generalized linear model by taking $f(t) := P(z_i \geq -t)$. The two are equivalent as long as $1 - f$ is a distribution function.

There are a number of theoretical results in sparse binary regression, focusing on sparse logistic regression [3], [5], [11], [13], [14], [19], [20]. A main message is that $O(\delta^{-2} s \log(n))$ measurements are sufficient to reconstruct $\boldsymbol{x}$ up to error δ by using ℓ_1-penalized maximum likelihood estimation [14]. Interestingly, these results allow for the reconstruction of both the direction, and norm of $\boldsymbol{x}$. However, there are two limitations to this maximum-likelihood-based approach: 1) knowledge of the function θ defining the generalized linear model is imperative, and 2) as the norm of $\boldsymbol{x}$ increases, the negative log-likelihood loses the strong convexity needed in the theoretical treatment.

The ideas from 1-bit compressed sensing allow us to overcome these two limitations. Indeed, the solution to (4) remains accurate for nearly any generalized linear model, but knowledge of the function θ is unnecessary in the reconstruction of $\boldsymbol{x}$. To make this precise, define

$$\lambda := \mathbb{E}\, g\, \theta(g)$$

where $g \sim N(0, 1)$. λ gives a notion of how correlated the response y is with the linear functionals $\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle$. Higher correlation improves reconstruction. For example, if f is the logistic function, then $\lambda \approx 0.41$.

Theorem 3 ([15], Corollary 3.1)**.** *Fix $\boldsymbol{x} \in S^{n-1}$ satisfying $\|\boldsymbol{x}\|_1 \leq \sqrt{s}$. Let $\boldsymbol{A}$ have independent standard normal entries and suppose that $\boldsymbol{y}$ follows the generalized linear model* (5)*. Let $\delta > 0$ and suppose that*

$$m \geq C\delta^{-4}\lambda^{-2} s \log(2n/s).$$

Then, with probability at least $1 - 8\exp(-c\delta^4 m)$, the solution $\hat{\boldsymbol{x}}$ to the convex program (4) *satisfies*

$$\|\hat{\boldsymbol{x}} - \boldsymbol{x}\|_2 \leq \delta.$$

Remark 3. The assumption that $\|\boldsymbol{x}\|_2 = 1$ in the theorem is really no assumption at all, since the norm of $\boldsymbol{x}$ may be absorbed into the definition of θ simply by rescaling the function. Further, suppose the following two mild assumptions on the model: 1) θ is monotonically increasing and 2) $\theta(0) = 0$. Then for a positive scalar t and standard normal random variable g, $\mathbb{E}\,\theta(tg)g$ is an increasing function of t. The implication is that rescaling θ, to absorb the norm of $\boldsymbol{x}$ causes λ to *increase* as long as $\|\boldsymbol{x}\|_2 \geq 1$. Thus, the reconstruction of the direction of $\boldsymbol{x}$ only improves as the magnitude of $\boldsymbol{x}$ increases. This contrasts with the maximum-likelihood approach discussed above.

B. sub-gaussian measurements

Up until now, we have considered Gaussian measurement vectors. One may ask whether 1-bit compressed sensing is possible with other random measurement schemes.

Let us consider Bernoulli measurement vectors in which each entry of $\boldsymbol{a}_i$ is an independent Bernoulli random variable, so that $\boldsymbol{a}_i \in \{+1, -1\}^n$. It is well known [7] that measurements of this form lead to near-optimal results in unquantized compressed sensing. Does the same hold true for 1-bit compressed sensing?

Consider two candidate signals $\boldsymbol{x} = (1, 0, 0, \ldots, 0)$ and $\bar{\boldsymbol{x}} = (1, 0.9, 0, 0, \ldots, 0)$. Then one has $\operatorname{sign}(\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle) = \operatorname{sign}(\langle \boldsymbol{a}_i, \bar{\boldsymbol{x}} \rangle)$ *deterministically*. Thus, when the measurement vectors are Bernoulli, $\boldsymbol{x}$ and $\bar{\boldsymbol{x}}$ are indistinguishable, and reconstruction of either is ill-posed.

However, it turns out that the above negative example is atypical. Signal reconstruction with Bernoulli measurements is possible—as long as the signal is not *too sparse*. This will be quantified by a bound on the maximum entry of $\boldsymbol{x}$ in Theorem 4 below.

Recall that a random variable η is called sub-gaussian if it has a sub-gaussian tail: $\mathbb{P}\left(\eta > t\right) \leq Ce^{-ct^2}$. Recall also that Bernoulli random variables are sub-gaussian.

Theorem 4 ([2], Theorem 1.3)**.** *Fix $\boldsymbol{x} \in S^{n-1}$ satisfying $\|\boldsymbol{x}\|_1 \leq \sqrt{s}$. Let a be a symmetric, sub-gaussian random variable with unit variance. Let $\boldsymbol{A}$ be generated with coordinates that are independent copies of a. Assume that $\boldsymbol{y}$ follows the generalized linear model* (5) *and the first three derivatives of θ are bounded. Suppose*

$$m \geq C\delta^{-4}\lambda^{-2} s \log(n/s)$$

and let $\hat{\boldsymbol{x}}$ be the solution to the convex program (4)*. Then with probability at least $1 - 4\exp(-c\delta^4 m)$*

$$\|\hat{\boldsymbol{x}} - \boldsymbol{x}\|_2 \leq \delta + C\left(\frac{\|\boldsymbol{x}\|_\infty}{\lambda^3}\right)^{1/4}.$$

For a more precise treatment which also allows for θ to be the discontinuous sign function, see [2]. We note that this theorem allows no correlation between the entries of $\boldsymbol{A}$. For a treatment of the case when each row of $\boldsymbol{A}$ is Gaussian with correlations between entries see Section 3.4 in [15].

C. General signal structures

While sparse signal structures are intrinsic in the sparse binary regression model and in some compressed sensing

problems, it is often of interest to consider more general signal structures. As a common example, $\boldsymbol{x}$ may not be sparse itself, but it may be sparse in a known dictionary, so that $\boldsymbol{x} = \boldsymbol{D}\boldsymbol{v}$ for a sparse vector v. Alternatively, $\boldsymbol{x}$ itself could be a matrix with low rank.

In general, the signal structure of $\boldsymbol{x}$ is defined by knowledge of a set K to which $\boldsymbol{x}$ belongs. Since we assume that $\|\boldsymbol{x}\|_2 = 1$, we may also assume that $K \subset B_2$ where B_2 is the unit ball. In this case, Vershynin and Plan [15] suggest to take the estimate of $\boldsymbol{x}$ to be the solution to the following program.

$$\max_{\boldsymbol{x}'} \sum_{i=1}^{m} y_i \langle \boldsymbol{a}_i, \boldsymbol{x}' \rangle \quad \text{such that} \quad \boldsymbol{x}' \in K. \tag{6}$$

For example, we may take $K = B_2 \cap \sqrt{s} B_1^n$ where B_1^n is the ℓ_1 ball. This recovers the convex program (4).

In this general case, reconstruction of $\boldsymbol{x}$ to accuracy δ requires $O(\delta^{-4} w(K)^2)$ binary measurements, where $w(K)$ is the *Gaussian mean width* of K. See [15] for details.

IV. Binary matrix completion

In a complementary line of research, Davenport et al. [6] analyze the following problem. Suppose that you see a subset of entries of a binary matrix, i.e., a matrix filled with ± 1 entries. From the observed entries, what information can be determined about unobserved entries? Problems of this nature arise in various applications. Consider for example the voting history of US senators on a number of bills, but with missing votes when a senator is out of town; or consider binary recommendation systems such as Pandora, in which one wishes to recommend unrated songs based on observed user ratings. For more applications see [6]. Davenport et al. assume that the data follows from the generalized linear model, but with three large differences from the considerations of the previous sections: 1) the measurements only give information about single entries of the matrix, 2) a low-rank structure is assumed in place of a sparse structure, and 3) θ (from Equation (5)) is assumed to be known. Under these assumptions, the authors show that nuclear-norm constrained maximum likelihood estimation gives minimax optimal reconstruction of the probability distribution of the unseen entries.

V. Conclusion

Binary data is intrinsic to many naturally arising inverse problems, and also arises in extreme quantization. But the signal or model that is to be reconstructed often comes from an infinite, albeit low-dimensional, set. This blend of continuous and discrete leads to interesting challenges in developing and analyzing accurate methods of signal reconstruction. We reviewed a number of recent results, which show that 1-bit measurements can give comparable information to unquantized measurements. Further, the methods of 1-bit compressed sensing allow for a semi-parametric treatment of sparse binary regression. One question that arises naturally from the above work is whether we can give a semi-parametric treatment of binary matrix completion which does not assume knowledge of θ.

Acknowledgment

The author would like to thank Roman Vershynin, Mark Davenport, Ewout Vandenberg, Mary Wootters, Albert Ai, and Alex Lapanowski for beneficial discussions. Most of the work reviewed in this paper is based on joint work with the above authors.

References

[1] One-bit compressive sampling webpage. http://dsp.rice.edu/1bitCS/.

[2] A. Ai, A. Lapanowski, Y. Plan, and R. Vershynin. One-bit compressed sensing with non-gaussian measurements. Submitted. Available at http://arxiv.org/abs/1208.6279.

[3] F. Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4:384–414, 2010.

[4] P. Boufounos and R. Baraniuk. 1-Bit compressive sensing. In *42nd Annual Conference on Information Sciences and Systems (CISS)*, March 2008.

[5] F. Bunea. Honest variable selection in linear and logistic regression models via ℓ_1 and $\ell_1 + \ell_2$ penalization. *Electronic Journal of Statistics*, 2:1153–1194, 2008.

[6] M. Davenport, Y. Plan, E. van den Berg, and M. Wootters. One-bit matrix completion. Submitted. Available at http://arxiv.org/abs/1209.3672.

[7] Y. Eldar and G. Kutyniok, editors. *Compressed Sensing*. Cambridge University Press, 2012.

[8] C. Gunturk, M. Lammers, A. Powell, R. Saab, and O. Ylmaz. Sigma delta quantization for compressed sensing. In *44th Annual Conference on Information Sciences and Systems (CISS)*. IEEE, 2010.

[9] A. Gupta, R. Nowak, and B. Recht. Sample complexity for 1-bit compressed sensing and sparse classification. In *International Symposium on Information Theory (ISIT)*. IEEE, 2010.

[10] L. Jacques, J. Laska, P. Boufounos, and R. Baraniuk. Robust 1-bit compressive sensing via binary stable embeddings of sparse vectors. 2011. Preprint. Available at http://arxiv.org/abs/1104.3160.

[11] S.M. Kakade, O. Shamir, K. Sridharan, and A. Tewari. Learning exponential families in high-dimensions: Strong convexity and sparsity. 2009. Preprint. Available at http://arxiv.org/abs/0911.0054.

[12] J. Laska and R. Baraniuk. Regime change: Bit-depth versus measurement-rate in compressive sensing. 2011. Preprint. Available at http://arxiv.org/abs/1110.3450.

[13] L. Meier, S. Van De Geer, and P. Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71, 2008.

[14] S. Negahban, P. Ravikumar, M.J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. 2010. Preprint. Available at http://arxiv.org/abs/1010.2731.

[15] Y. Plan and R. Vershynin. Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *IEEE Transactions on Information Theory*. Accepted for publication, 2012.

[16] Y. Plan and R. Vershynin. Dimension reduction by random hyperplane tessellations. 2011. Submitted. Available at http://arxiv.org/abs/1111.4452.

[17] Y. Plan and R. Vershynin. One-bit compressed sensing by linear programming. *Communications on Pure and Applied Math*, 2011. Accepted for publication. Available at http://arxiv.org/abs/1109.4299.

[18] G. Raskutti, M.J. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *Information Theory, IEEE Transactions on*, 57(10):6976–6994, 2011.

[19] P. Ravikumar, M.J. Wainwright, and J.D. Lafferty. High-dimensional ising model selection using 1-regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319, 2010.

[20] S.A. Van De Geer. High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36(2):614–645, 2008.

Dictionary Identification Results for K-SVD with Sparsity Parameter 1

Karin Schnass
Computer Vision Laboratory
University of Sassari
Porto Conte Ricerche, 07041 Alghero, Italy
kschnass@uniss.it

***Abstract*—In this paper we summarise part of the results from our recent work [1]. We give theoretical insights into the performance of K-SVD, a dictionary learning algorithm that has gained significant popularity in practical applications, by answering the question when a dictionary Φ can be recovered as local minimum of the minimisation criterion underlying K-SVD from a set of training signals $y_n = \Phi x_n$. Assuming the training signals are generated from a tight frame with coefficients drawn from a random symmetric distribution, then in expectation the generating dictionary can be recovered as a local minimum of the K-SVD criterion if the coefficient distribution exhibits sufficient decay. This decay can be characterised by the coherence of the dictionary and the ℓ_1-norm of the coefficients. Further it is demonstrated that given a finite number of training samples N with probability $O(\exp(-N^{1-4q}))$ there is a local minimum of the K-SVD criterion within a radius $O(N^{-q})$ of the generating dictionary.**

***Index Terms*—dictionary learning, sparse coding, finite samples, K-SVD, sampling complexity, dictionary identification, minimisation criterion, sparse representation**

I. Introduction

Research in the last decade has proven that sparsity provides an efficient way of dealing with high-dimensional data, since sparse signals are easily compressed, are robust to corruption and can therefore easily be restored from incomplete information. Triggered by this success an increasingly important research direction is how to learn dictionaries providing sparse representations for the data at hand, known as dictionary learning or sparse coding. The problem under investigation is usually formulated as follows. Given N signals $y_n \in \mathbb{R}^d$, stored as columns in a matrix $Y = (y_1, \dots y_N)$ find a decomposition,

$$Y \approx \Phi X,$$

into a $d \times K$ dictionary matrix Φ with unit norm columns and a $K \times N$ coefficient matrix with sparse columns.

So far research has provided several dictionary learning algorithms, which are efficient in practice and therefore popular in applications, but there exists only a handful of dictionary learning schemes, for which theoretical results available, [3], [4], [5], [6], [7]. Unfortunately, however, these then tend to be rather cumbersome in practice. In this talk we start bridging the gap between practically efficient and provably efficient dictionary learning schemes, by shedding some light on the theoretical performance of K-SVD, one of the most widely applied dictionary algorithms.

This work was supported by the Austrian Science Fund (FWF) under Grant no. Y432 an J3335.

K-SVD was introduced by Aharon, Elad and Bruckstein in [8] as an algorithm to solve the following minimisation problem. Given some signals $Y = (y_1, \dots, y_N)$, $y_n \in \mathbb{R}^d$, find

$$\min_{\Phi \in \mathcal{D}, X \in \mathcal{X}_S} \|Y - \Phi X\|_F^2, \tag{1}$$

for $\mathcal{D} := \{\Phi = (\phi_1, \dots, \phi_K), \phi_i \in \mathbb{R}^d, \|\phi_i\|_2 = 1\}$ and $\mathcal{X}_S := \{X = (x_1, \dots, x_N), x_n \in \mathbb{R}^K, \|x_n\|_0 \leq S\}$, where $\|x\|_0$ counts the number of non-zero entries of x, and $\|\cdot\|_F$ denotes the Frobenius norm. In short we are looking for the dictionary Φ that provides on average the best S-term approximation to the signals in Y.

Since for a signal y_n the best S-term approximation using Φ is given by the largest projection onto a set of S atoms $\Phi_I = (\phi_{i_1} \dots \phi_{i_S})$, ie. $P_I(\Phi) = \Phi_I \Phi_I^\dagger$ where $\Phi_I^\dagger$ denotes the Moore-Penrose pseudo inverse of Φ_I, instead of (1) we can equivalently consider the following maximisation problem,

$$\max_{\Phi \in \mathcal{D}} \sum_i \max_{|I| \leq S} \|P_I(\Phi) y_n\|_2^2. \tag{2}$$

Let us assume that the training signals are all created from an admissible generating dictionary $\bar{\Phi} \in \mathcal{D}$, and coefficients drawn at random from a distribution ν of sparse or rapidly decaying coefficient, ie.

$$y_n = \bar{\Phi} \bar{x_n}. \tag{3}$$

The goal of dictionary identification is to give conditions under which an algorithm can locally identify the generating dictionary from the training signals. To achieve this we will first study when $\bar{\Phi}$ is exactly at a local maximum in the limiting case, ie. when we replace the sum in (2) with the expectation,

$$\max_{\Phi \in \mathcal{D}} \mathbb{E}_y \left(\max_{|I| \leq S} \|P_I(\Phi) y\|_2^2 \right). \tag{4}$$

In the next section we will provide identification results for the case when in (4) we have $S = 1$, ie. $\mathcal{X}_S = \mathcal{X}_1$, assuming first a simple (discrete, noise-free) signal model and then progressing

to a noisy, continuous signal model. In Section III we will extend these asymptotic results to the case of a finite number of samples and finally we will discuss the implications of our results for practical applications and compare them to related dictionary identification results.

II. Asymptotic identification results

A. The problem for $S = 1$

In case $S = 1$ the objective function in (4) can be radically simplified and the maximisation problem we want to analyse reduces to,

$$\max_{\Phi\in\mathcal{D}} \mathbb{E}_y \left(\|\Phi^\star y\|_\infty^2\right). \tag{5}$$

Clearly if the signals y are all 1-sparse in a dictionary $\bar{\Phi}$ then $\bar{\Phi}$ is a global maximiser of (5). However what happens if we do not have perfect sparsity? Let us start with a very simple negative example.

Example 2.1: Let U be an orthonormal basis and x be randomly 2-sparse with 'flat' coeffcients, ie. pick two indices i, j uniformly at random, choose $\sigma_{i/j} = \pm 1$ uniformly at random and set $x_k = \sigma_k$ for $k = i, j$ and zero else. Then U is not a local maximum of (5), since we can construct an ascent direction. Choose $U_\varepsilon = (u_1, \dots, u_{d-1}, (u_d + \varepsilon u_1)/\sqrt{1+\varepsilon^2})$, then we have

$$\begin{aligned}\mathbb{E}_y \left(\|U_\varepsilon^\star y\|_\infty^2\right) &= \mathbb{E}_x \left(\|(x_1, \dots, x_{d-1}, \tfrac{x_d+\varepsilon x_1}{\sqrt{1+\varepsilon^2}})\|_\infty^2\right) \\ &= 1 + \tfrac{1}{d(d-1)} \tfrac{\varepsilon}{1+\varepsilon^2} > 1 = \mathbb{E}_y \left(\|U^\star y\|_\infty^2\right).\end{aligned}$$

From the above example we see that in order to have a local maximum at the original dictionary we need a signal/coefficient model where the coefficients show some type of decay.

B. A simple model of decaying coefficients

We first consider a very simple coefficient model, constructed from a non-negative, non-increasing sequence $c \in \mathbb{R}^K$ with $\|c\|_2 = 1$, which we permute uniformly at random and provide with random $\pm$ signs. To be precise for a permutation $p : \{1, ..., K\} \to \{1, ..., K\}$ and a sign sequence σ, $\sigma_i = \pm 1$, we define the sequence $c_{p,\sigma}$ component-wise as $c_{p,\sigma}(i) := \sigma_i c_{p(i)}$, and set $y = \Phi x$ where $x = c_{p,\sigma}$ with probability $(2^K K!)^{-1}$.

The normalisation $\|c\|_2 = 1$ has the advantage that for dictionaries, which are an orthonormal basis, the resulting signals also have unit norm and for general dictionaries the signals have unit square norm in expectation, ie. $\mathbb{E}(\|y\|_2^2) = 1$. This reflects the situation in practical application, where we would normalise the signals in order to equally weight their importance.

Armed with this model we can now prove a first dictionary identification result for (5).

Theorem 2.1: Let Φ be a unit norm tight frame with frame constant $A = K/d$ and coherence μ. Let $x \in \mathbb{R}^d$ be a random permutation of a sequence c, where $c_1 \geq c_2 \geq c_3 \dots \geq c_K \geq 0$ and $\|c\|_2 = 1$, provided with random $\pm$ signs, i.e. $x = c_{p,\sigma}$ with probability $\mathbb{P}(p, \sigma) = (2^K K!)^{-1}$. If c satisfies $c_1 > c_2 + 2\mu\|c\|_1$, then there is a local maximum of (5) at Φ. Moreover we have the following quantitative estimate for the basin of attraction around Φ. For all perturbations $\Psi = (\psi_1 \dots \psi_K)$ of $\Phi = (\phi_1 \dots \phi_K)$ with $0 < \max_i \|\psi_i - \phi_i\|_2 \leq \varepsilon$ we have $\mathbb{E}_x\|\Psi^\star\Phi x\|_\infty^2 < \mathbb{E}_x\|\Phi^\star\Phi x\|_\infty^2$ as soon as $\varepsilon < 1/5$ and

$$\varepsilon \leq \frac{\left(1 - 2\frac{c_2+\mu\|c\|_1}{c_2+c_1}\right)^2}{2A\log\left(2AK/(c_1^2 - \frac{1-c_1^2}{K-1})\right)}. \tag{6}$$

Proof: We briefly sketch the proof. The condition $c_1 > c_2 + 2\mu\|c\|_1$ ensures that the maximal inner product $|\langle\phi_i, \Phi c_{p,\sigma}\rangle|$ is always attained by $i_p = p^{-1}(1)$, leading to

$$\mathbb{E}_x\|\Phi^\star\Phi x\|_\infty^2 = c_1^2 + \frac{(1-c_1^2)}{K-1}(A-1).$$

The main idea now is that for small perturbations and most sign patterns σ the maximal inner product is still attained by i_p. For an ε-perturbations Ψ of the original dictionary Φ where $\psi_i = (1-\varepsilon_i^2/2)\phi_i + (\varepsilon_i^2 - \varepsilon_i^4/4)^{\frac{1}{2}} z_i$, for some z_i with $\langle\phi_i, z_i\rangle = 0, \|z_i\|_2 = 1$ and $\varepsilon_1 \leq \varepsilon$, we have

$$\max_{i=1\dots K} |\langle\psi_i, \Phi c_{p,\sigma}\rangle| = |\langle\psi_{i_p}, \Phi c_{p,\sigma}\rangle|,$$

except with probability

$$\eta := 2\sum_{i|\varepsilon_i\neq 0} \exp\left(-\frac{\left(1 - \frac{\varepsilon^2}{2} - 2\frac{c_2+\mu\|c\|_1}{c_2+c_1}\right)^2}{2A\varepsilon_i^2}\right),$$

which leads to the following bound

$$\begin{aligned}\mathbb{E}_x\|\Psi^\star\Phi x\|_\infty^2 \leq 2A\eta &+ \frac{c_1^2}{K}\sum_{i=1}^K (1-\varepsilon_i^2/2)^2 \\ &+ \frac{1-c_1^2}{K-1}\left(A - \frac{1}{K}\sum_{i=1}^K (1-\varepsilon_i^2/2)^2\right).\end{aligned}$$

Since e^{-c/ε^2} and therefore η decays much faster than ε^2 as ε goes to zero we have $\mathbb{E}_x\|\Psi^\star\Phi x\|_\infty^2 < \mathbb{E}_x\|\Phi^\star\Phi x\|_\infty^2$, as soon as ε is small enough. ∎

Remark 2.2: (i) Note that in some sense Theorem 2.1 is sharp. Assume that Φ is an orthonormal basis (ONB) then $\mu = 0$ and the condition to be a local maximum reduces to $c_1 > c_2$. However from Example 2.1 we see that if $c_1 = c_2$ we can again construct an ascent direction and so Φ is not a local maximum.

(ii) Similarly the condition that Φ is a tight frame is almost necessary in the non-trivial case where $|c_1| < 1$, as otherwise the candidate local maximiser at the generating dictionary may be distorted towards the maximal eigenvector of the frame.

C. A continuous model of decaying coefficients

Next we would like to extend the result from the last subsection to a wider range of coefficient distributions, especially continuous ones. To characterise suitable distributions we make the following definitions.

Definition 2.1: A probability measure ν on the unit sphere $S^{d-1} \subset \mathbb{R}^d$ is called symmetric if for all measurable sets $\mathcal{X} \subseteq S^{d-1}$, for all sign sequences $\sigma \in \{-1,1\}^d$ and all permutations p we have

$$\nu(\sigma\mathcal{X}) = \nu(\mathcal{X}), \quad \sigma\mathcal{X} := \{(\sigma_1 x_1, \dots, \sigma_d x_d) : x \in \mathcal{X}\}$$
$$\nu(p(\mathcal{X})) = \nu(\mathcal{X}), \quad p(\mathcal{X}) := \{(x_{p(1)}, \dots, x_{p(d)}) : x \in \mathcal{X}\}$$

Definition 2.2: A probability distribution ν on the unit sphere $S^{K-1} \subset \mathbb{R}^K$ is called (β,μ)-decaying if there exists a $\beta < 1/2$ such that for $c_1(x) \geq c_2(x) \geq \ldots \geq c_d(x) \geq 0$ a non increasing rearrangement of the absolute values of the components of x we have,

$$\nu\left(\frac{c_2(x) + \mu\|c(x)\|_1}{c_2(x) + c_1(x)} \leq \beta\right) = 1 \tag{7}$$

For the case $\mu = 0$ it will also be useful to define the following notion. A probability distribution ν on the unit sphere $S^{d-1} \subset \mathbb{R}^d$ is called f-decaying if there exists a function f such that

$$\exp\left(-\frac{f(\varepsilon)^2}{8\varepsilon^2}\right) = o(\varepsilon^2)$$
$$\text{and} \qquad \nu\left(\frac{c_2(x)}{c_1(x)} \geq 1 - f(\varepsilon)\right) = o(\varepsilon^2).$$

Note that $(\beta,0)$-decaying is a special case of f-decaying, ie. $f(\varepsilon)$ can be chosen constant β. To illustrate both concepts we give simple examples for (β,μ)- and f-decaying distributions on S^1.

Example 2.3:
- Let ν be the symmetric distribution on S^1 defined by $c_2(x)$ being uniformly distributed on $[0, \frac{1}{\sqrt{2}} - \theta]$ for $\theta > 0$ (and accordingly $c_1(x) = \sqrt{1 - c_2^2(x)}$), then ν is (β,μ)-decaying for all $\mu < \frac{\theta}{\sqrt{2}}$.
- Let ν be the symmetric distribution on S^1 defined by $c_2(x)$ being distributed on $[0, \frac{1}{\sqrt{2}}]$ with density $20\sqrt{2}(\frac{1}{\sqrt{2}} - x)^4$, then ν is f-decaying for e.g. $f(\varepsilon) = \sqrt{\varepsilon}$.
- Let ν be the symmetric distribution on S^1 defined by $c_2(x)$ being distributed on $[0, \frac{1}{\sqrt{2}}]$ with density $4(\frac{1}{\sqrt{2}} - x)$, then ν is not f-decaying.

With these examples of suitable probability distributions in mind we can now give a continuous version of Theorem 2.1.

Theorem 2.2: (a) Let Φ be a unit norm tight frame with frame constant $A = K/d$ and coherence μ. If x is drawn from a symmetric (β,μ)-decaying probability distribution ν on the unit sphere S^{K-1}, then there is a local maximum of (5) at Φ.
(b) If Φ is an orthonormal basis, there is a local maximum of (5) at Φ whenever x is drawn from a symmetric f-decaying probability distribution ν on the unit sphere S^{d-1}.

D. Bounded white noise

With the tools used to prove the two noiseless identification results it is also possible to analyse the case of (very small) bounded white noise.

Theorem 2.3: Let Φ be a unit norm tight frame with frame constant $A = K/d$ and coherence μ. Assume that the signals y are generated from the following model

$$y = \Phi x + r, \tag{8}$$

where r is a bounded random white noise vector, ie. there exist two constants $\rho, \rho_{\max}$ such that $\|r\|_2 \leq \rho_{\max}$ almost surely, $\mathbb{E}(r) = 0$ and $\mathbb{E}(rr^\star) = \rho^2 I$. If x is drawn from a symmetric decaying probability distribution ν on the unit sphere S^{K-1} with $\mathbb{E}_x\|x\|_\infty^2 = \bar{c_1}^2$ and the maximal size of the noise is small compared to the size and decay of the coefficients c_1, c_2, meaning there exists $\beta < 1/2$, such that

$$\nu\left(\frac{c_2(x) + \mu\|c(x)\|_1 + \rho_{\max}}{c_1(x) - c_2(x)} \leq \beta\right) = 1 \tag{9}$$

then there is a local maximum of (5) at Φ.

III. Finite sample size

We are now ready to analyse the local maxima of the non-asymptotic maximisation problem for $S = 1$. For simplicity we choose a normalised version of (2).

$$\max_{\Phi\in\mathcal{D}} \frac{1}{N}\sum_{n=1}^{N} \|\Phi^\star y_n\|_\infty^2. \tag{10}$$

Theorem 3.1: Let Φ be a unit norm tight frame with frame constant $A = K/d$ and coherence μ. Assume that the signals y_n are generated as $y_n = \Phi x_n + r_n$, where r_n is a bounded random white noise vector, ie. there exist two constants $\rho, \rho_{\max}$ such that $\|r_n\|_2 \leq \rho_{\max}$ almost surely, $\mathbb{E}(r_n) = 0$ and $\mathbb{E}(r_n r_n^\star) = \rho^2 I$. Further let x_n be drawn from a symmetric decaying probability distribution ν on the unit sphere S^{K-1} with $\mathbb{E}_x\|x\|_\infty^2 = \bar{c_1}^2$ and the maximal size of the noise be small compared to the size and decay of the coefficients c_1, c_2, meaning there exists $\beta < 1/2$, such that

$$\nu\left(\frac{c_2(x) + \mu\|c(x)\|_1 + \rho_{\max}}{c_1(x) - c_2(x)} \leq \beta\right) = 1. \tag{11}$$

Abbreviate $\gamma := \bar{c_1}^2 - \frac{1-\bar{c_1}^2}{K-1}$ and $C_L = (\sqrt{A} + \rho_{\max})^2$. If for some $0 < q < 1/4$ the number of samples N satisfies

$$N^{-q} + N^{-2q}/K \leq \frac{(1-2\beta)^2}{4A\log(4AK/\gamma)} \tag{12}$$

then except with probability

$$\exp\left(\frac{-N^{1-4q}\gamma^2}{4K^2C_L^2} + Kd\log(NKC_L/\gamma)\right),$$

there is a local maximum of (10) resp. local minimum of (1) within distance at most $2N^{-q}$ to Φ, ie. for the local maximum $\tilde{\Psi}$ we have $\max_k \|\tilde{\psi}_k - \phi_k\|_2 \leq 2N^{-q}$.

Proof: We again give a brief sketch of the proof. From the last section we know that for any ε-perturbation we have

$$\mathbb{E}_y\|\Phi^\star y\|_\infty^2 - \mathbb{E}_y\|\Psi^\star y\|_\infty^2 \approx \varepsilon^2\gamma/K.$$

Hoeffding's inequality lets us estimate the probability that for a fixed perturbation the finite sample sum deviates from its expectation as

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_{n=1}^{N}\|\Psi^\star y_n\|_\infty^2 - \mathbb{E}_y\|\Psi^\star y\|_\infty^2\right| > t\right) \leq e^{-Nt^2/C_L^2}.$$

Using a union bound this leads to an estimate for the probability that the above holds for a δ-net $\mathcal{N}$ for the set of all ε-perturbations with $\varepsilon \leq \varepsilon_{\max}$. Since this set is the product of K $(d-1)$-dimensional balls with radius ε_{max} we have

$$\sharp\mathcal{N} \leq (3\varepsilon_{\max}/\delta)^{K(d-1)}.$$

Choosing δ and t to be $O(N^{-q})$ the final result then follows from a triangular inequality argument and the fact that

$$|\|\Psi^\star y_n\|_\infty^2 - \|\bar{\Psi}^\star y_n\|_\infty^2| \leq 3C_L \max_k \|\psi_k - \bar{\psi}_k\|_2.$$

■

IV. Discussion

We have shown that the K-SVD minimisation principle with sparsity parameter 1 can correctly identify a tight frame from signals generated from a wide class of decaying coefficients distributions. Since any simple greedy algorithm will always find the best 1-term approximation for any signal in any dictionary our results give conditions under which the K-SVD algorithm can identify the underlying dictionary given a good initialisation.

Before turning to a comparison of our results to other dictionary learning schemes we illustrate the limitations of the K-SVD principle for learning non-tight frames. We generated 1000 coefficients by drawing $c_2(x)$ uniformly at random from the interval $[0, 0.6]$, setting $c_1(x) = \sqrt{1 - c_2^2(x)}$, randomly permuting the resulting vector and providing it with random $\pm$ signs. We then generated two sets of signals, using an orthogonal and an oblique basis with the same coefficients, and for both sets of signals found the minimiser of the K-SVD criterion (1) with $S = 1$. Figure 1 shows the two signal sets, the generating bases and the recovered bases. As predicted by our theoretical results when the generating basis is orthogonal it is also the minimiser of the K-SVD criterion, while for the oblique generating basis the minimiser is distorted towards the maximal eigenvector of the basis.

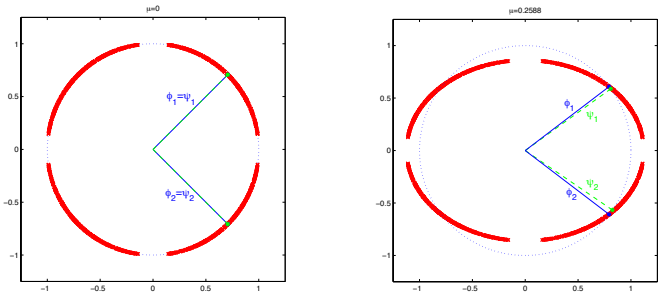


Fig. 1. Signals created from an orthogonal and an oblique basis $\Phi = (\phi_1, \phi_2)$ with decaying coefficients, together with the corresponding minimiser $\Psi = (\psi_1, \psi_2)$ of the K-SVD-criterion for $S = 1$.

Finally let us point out further research directions based on a comparison of our results for the K-SVD-minimisation principle to the identification results for the ℓ_1-minimisation principle,

$$\min_{\Phi\in\mathcal{D}, X:Y=\Phi X} \sum_{ij} |X_{ij}|, \tag{13}$$

derived in [5], [6]. At first glance it seems that the K-SVD-criterion requires a larger sample size than the ℓ_1-criterion, ie. $N^{1-4q}/\log N = O(K^3 d)$ as opposed to $O(d^2 \log d)$ reported in [5] for a basis and $O(K^3)$ reported in [6] for an overcomplete dictionary. Also it does not allow for exact identification with high probability but only guarantees stability. However this effect may be due to the more general signal model which assumes decay rather than exact sparsity. Indeed it is very interesting to compare our results to a recent result for a noisy version of the ℓ_1-minimisation principle, [7], which provides stability results under unbounded white noise and, omitting log factors, also derives a sampling complexity of $O(K^3 d)$.

Another difference, apparently intrinsic to the minimisation criteria is that the K-SVD criterion can only identify tight dictionary frames exactly, while the ℓ_1-criterion allows identification of arbitrary dictionaries. Thus to support the use of K-SVD for the learning of non-tight dictionaries also theoretically, we plan to study the stability of the K-SVD criterion under non-tightness by analysing the maximal distance between an original, non tight dictionary with condition number $\sqrt{B/A} > 1$ and the closest local maximum, cp. also Figure 1.

The last research direction we want to point out is how much decay of the coefficients is actually necessary. For the asymptotic results we used condition $c_1 > c_2 + 2\mu\|c\|_1$ to ensure that the maximal inner product is always attained at i_p. However typically we have $|\langle\phi_i, \Phi c_{p,\sigma}\rangle| \approx c_{p(i)} \pm \mu$. Therefore a condition such as $c_1 > c_2 + O(\mu)$, which allows for outliers, ie. signals for which the maximal inner product is not attained at i_p, might be sufficient to prove exact identifiability or - failing that - to again show stability. Together with the inspiring techniques from [7], we expect the tools developed in the course of such an analysis to allow us also to deal with unbounded white noise.

References

[1] K. Schnass, "On the identifiability of overcomplete dictionaries via the minimisation principle underlying K-SVD," *arXiv:1301.3375*, 2013.

[2] ——, "Sampling complexity of dictionary learning via the k-svd-minimisation criterion," *in preparation*, 2013.

[3] M. Aharon, M. Elad, and A. Bruckstein, "On the uniqueness of overcomplete dictionaries, and a practical way to retrieve them," *Journal of Linear Algebra and Applications*, vol. 416, pp. 48–67, July 2006.

[4] P. Georgiev, F. Theis, and A. Cichocki, "Sparse component analysis and blind source separation of underdetermined mixtures," *IEEE Transactions on Neural Networks*, vol. 16, no. 4, pp. 992–996, 2005.

[5] R. Gribonval and K. Schnass, "Dictionary identifiability - sparse matrix-factorisation via l_1-minimisation," *IEEE Transactions on Information Theory*, vol. 56, no. 7, pp. 3523–3539, July 2010.

[6] Q. Geng, H. Wang, and J. Wright, "On the local correctness of ℓ^1-minimization for dictionary learning," *arXiv:1101.5672v1 [cs.IT]*, 2011.

[7] R. Jenatton, F. Bach, and R. Gribonval, "Local stability and robustness of sparse dictionary learning in the presence of noise," *preprint*, 2012.

[8] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. on Signal Processing.*, vol. 54, no. 11, pp. 4311–4322, November 2006.

Sampling and Reconstruction of Bandlimited BMO-Functions

Holger Boche*
Technische Universität München
Lehrstuhl für Theoretische Informationstechnik
E-mail: boche@tum.de

Ullrich J. Mönich†
Massachusetts Institute of Technology
Research Laboratory of Electronics
E-mail: moenich@mit.edu

***Abstract*—Functions of bounded mean oscillation (BMO) play an important role in complex function theory and harmonic analysis. In this paper a sampling theorem for bandlimited BMO-functions is derived for sampling points that are the zero sequence of some sine-type function. The class of sine-type functions is large and, in particular, contains the sine function, which corresponds to the special case of equidistant sampling. It is shown that the sampling series is locally uniformly convergent if oversampling is used. Without oversampling, the local approximation error is bounded.**

I. Notation

Let $\hat{f}$ denote the Fourier transform of a function f. $L^p(\mathbb{R})$, $1 \leq p < \infty$, is the space of all pth-power Lebesgue integrable functions on $\mathbb{R}$, with the usual norm $\|\cdot\|_p$, and $L^\infty(\mathbb{R})$ is the space of all functions for which the essential supremum norm $\|\cdot\|_\infty$ is finite. For $0 < \sigma < \infty$ let $\mathcal{B}_\sigma$ be the set of all entire functions f with the property that for all $\epsilon > 0$ there exists a constant $C(\epsilon)$ with $|f(z)| \leq C(\epsilon)\exp((\sigma+\epsilon)|z|)$ for all $z \in \mathbb{C}$. The Bernstein space $\mathcal{B}_\sigma^p$, $1 \leq p \leq \infty$, consists of all functions in $\mathcal{B}_\sigma$, whose restriction to the real line is in $L^p(\mathbb{R})$. A function in $\mathcal{B}_\sigma^p$ is called bandlimited to σ.

II. Introduction and Motivation

A well-known result in sampling theory is Brown's theorem, which states that the Shannon sampling series

$$\sum_{k=-\infty}^{\infty} f(k)\frac{\sin(\pi(t-k))}{\pi(t-k)}$$

is locally uniformly convergent for all functions in the Paley–Wiener space $\mathcal{PW}_\pi^1$. $\mathcal{PW}_\pi^1$ is the space of all functions f with a representation $f(z) = 1/(2\pi)\int_{-\sigma}^{\sigma} g(\omega)\,\mathrm{e}^{iz\omega}\ \mathrm{d}\omega$, $z \in \mathbb{C}$, for some $g \in L^1[-\pi,\pi]$. This sampling theorem has been extended in various directions, for example, larger signal spaces and non-equidistant sampling patterns [1].

In this paper we consider the sampling series

$$\sum_{k=-\infty}^{\infty} f(t_k)\phi_k(t), \tag{1}$$

where the sampling points $\{t_k\}_{k\in\mathbb{Z}}$ are the zero sequence of some sine-type function and the functions $\{\phi_k\}_{k\in\mathbb{Z}}$ are certain reconstruction functions, and analyze its convergence behavior for bandlimited $\mathrm{BMO}(\mathbb{R})$-functions.

Definition 1. A function $f : \mathbb{R} \to \mathbb{C}$ is said to belong to $\mathrm{BMO}(\mathbb{R})$, provided that it is locally in $L^1(\mathbb{R})$ and $\frac{1}{|I|}\int_I |f(t) - m_I(f)|\ \mathrm{d}t \leq C_1$ for all bounded intervals I, where $m_I(f) := \frac{1}{|I|}\int_I f(t)\ \mathrm{d}t$ and the constant C_1 is independent of I. $|I|$ denotes the Lebesgue measure of the set I. We further define

$$\|f\|_{\mathrm{BMO}(\mathbb{R})} := \sup_I \frac{1}{|I|}\int_I |f(t) - m_I(f)|\ \mathrm{d}t,$$

where the supremum is over all bounded intervals I. By BMO_π we denote the space of all functions in $\mathcal{B}_\pi$ that are in $\mathrm{BMO}(\mathbb{R})$ when restricted to the real axis.

Note that $\|\cdot\|_{\mathrm{BMO}(\mathbb{R})}$ is actually a seminorm, because we have $\|C\|_{\mathrm{BMO}(\mathbb{R})} = 0$ for all constants $C \in \mathbb{C}$.

A consequence of the famous Fefferman–Stein theorem [2] is the fact that an arbitrary $\mathrm{BMO}(\mathbb{R})$-function can be decomposed into a $L^\infty(\mathbb{R})$-function and the Hilbert transform of a $L^\infty(\mathbb{R})$-function [3, p. 248].

Theorem A (Fefferman–Stein)**.** *There exists a constant $C_2 > 0$ such that for all $f \in BMO(\mathbb{R})$ there exist two functions $f_1, f_2 \in L^\infty(\mathbb{R})$ and a constant α such that $f = f_1 + \mathfrak{H}f_2 + \alpha$ and $\|f_1\|_\infty \leq C_2\|f\|_{BMO(\mathbb{R})}$, $\|f_2\|_\infty \leq C_2\|f\|_{BMO(\mathbb{R})}$.*

Theorem A is an important theoretical result [3], but it also has a high significance for applications where the Hilbert transform is used, for example the calculation of the analytic signal [4] and the analysis of signal properties [5], [6], [7]. Since the Hilbert transform of bounded functions is of particular interest, Theorem A is interesting because it essentially describes the range of the Hilbert transform for $L^\infty(\mathbb{R})$.

III. Sampling for BMO_π

Definition 2. An entire function f of exponential type π is said to be of sine type if the zeros of f are separated and simple, and there exist positive constants A, B, and H such that $A\,\mathrm{e}^{\pi|y|} \leq |f(x+iy)| \leq B\,\mathrm{e}^{\pi|y|}$ whenever x and y are real and $|y| \geq H$.

Without loss of generality, we assume that the sequence of sampling points $\{t_k\}_{k\in\mathbb{Z}}$ is ordered strictly increasingly and

*This work was partly supported by the German Research Foundation (DFG) under grant BO 1734/13-2.
†U. Mönich was supported by the German Research Foundation (DFG) under grant MO 2572/1-1.

that $t_0 = 0$. Then, it follows that the product

$$\phi(z) = z \lim_{N\to\infty} \prod_{\substack{|k|\le N \\ k\neq 0}} \left(1 - \frac{z}{t_k}\right) \tag{2}$$

converges uniformly on $|z| \le R$ for all $R < \infty$, and ϕ is an entire function of exponential type π [8]. It can bee seen from (2) that ϕ, which is often called generating function, has the zeros $\{t_k\}_{k\in\mathbb{Z}}$. Moreover, it follows that

$$\phi_k(t) = \frac{\phi(t)}{\phi'(t_k)(t - t_k)} \tag{3}$$

is the unique function in $\mathcal{B}_\pi^2$ that solves the interpolation problem $\phi_k(t_l) = \delta_{kl}$, where $\delta_{kl} = 1$ if $k = l$, and $\delta_{kl} = 0$ otherwise.

Sampling point sequences that are made of the zeros of functions of sine type are also complete interpolating sequences for $\mathcal{B}_\pi^2$ [9, p. 143]. This means that we restrict our analysis to a subclass of complete interpolating sequences. We conjecture that for arbitrary complete interpolating sequences a result like the one in this paper cannot be obtained even for smaller signal spaces. In particular, we conjecture that there exist complete interpolating sequences and functions in $\mathcal{PW}_\pi^1$ such that the sampling series is even locally divergent [10]. If this conjecture is true it shows the specialty of sine type function generated sampling patterns.

In [11] it was shown that for signals in $\mathcal{B}_{\beta\pi}^\infty$, $0 < \beta < 1$, the sampling series (1) is uniformly convergent on all compact subsets of $\mathbb{R}$. The proof in [11] makes use of certain essential properties of sine type functions.

Theorem C. *Let ϕ be a function of sine type, whose zeros $\{t_k\}_{k\in\mathbb{Z}}$ are all real and ordered increasingly. Furthermore, let ϕ_k be defined as in* (3) *and $0 < \beta < 1$. Then, for all $T > 0$ and all $f \in \mathcal{B}_{\beta\pi}^\infty$ we have*

$$\lim_{N\to\infty} \max_{t\in[-T,T]} \left| f(t) - \sum_{k=-N}^{N} f(t_k)\phi_k(t) \right| = 0.$$

In the next theorem we provide a sampling theorem for BMO_π-functions, and thus extend Theorem C to a larger space.

Theorem 1. *Let ϕ be a function of sine type, whose zeros $\{t_k\}_{k\in\mathbb{Z}}$ are all real and ordered increasingly. Furthermore, let ϕ_k be defined as in* (3) *and $T > 0$.*

1) *We have*

$$\sup_{N\in\mathbb{N}} \max_{t\in[-T,T]} \left| f(t) - \sum_{k=-N}^{N} f(t_k)\phi_k(t) \right| < \infty$$

for all $f \in$ BMO$_\pi$.

2) *Let $0 < \beta < 1$. For all $f \in$ BMO$_{\beta\pi}$ we have*

$$\lim_{N\to\infty} \max_{t\in[-T,T]} \left| f(t) - \sum_{k=-N}^{N} f(t_k)\phi_k(t) \right| = 0.$$

Theorem 1 shows that without oversampling the local peak value of the approximation error is bounded. With oversampling the sampling series is uniformly convergent on all compact subsets of $\mathbb{R}$.

IV. Proof of Theorem 1

In this section we prove Theorem 1. For the proof we need several auxiliary results.

A. Basic Properties of BMO$_\pi$-Functions

For functions f in BMO_π, i.e., $\mathrm{BMO}(\mathbb{R})$-functions that are additionally bandlimited, the Fefferman–Stein decomposition (Theorem A) is of course also possible because $\mathrm{BMO}_\pi \subset \mathrm{BMO}(\mathbb{R})$. The functions f_1 and f_2 in this decomposition are in $L^\infty(\mathbb{R})$. However, since the function f is additionally bandlimited, it is reasonable to ask whether the decomposition can be performed in a such a way that f_1 and f_2 are also bandlimited, i.e., in $\mathcal{B}_\pi^\infty$. The next theorem, which has been proved in [12], answers this question in the affirmative.

Theorem 2. *There exists a constant $C_3 > 0$ such that for all $f \in$ BMO$_\pi$ there exist two functions $f_1, f_2 \in \mathcal{B}_\pi^\infty$ and a constant α such that $f = f_1 + \mathfrak{H}f_2 + \alpha$ and $\|f_1\|_\infty \le C_3\|f\|_{BMO(\mathbb{R})}$, $\|f_2\|_\infty \le C_3\|f\|_{BMO(\mathbb{R})}$.*

Theorem 2 has been stated in a form to have maximum similarity to the Fefferman–Stein theorem. However, the bandwidth of the function f_2 does not have to be π; it can be arbitrarily reduced. Hence, we have the next theorem [12]. It should be noted that a decrease of the bandwidth of the function f_2 comes in general with an increase of the L^∞-norm of f_1 and f_2.

Theorem 3. *For all $0 < \hat\beta \le 1$ there exists a constant C_4 such that for all $f \in$ BMO$_\pi$ there exist two functions $f_3 \in \mathcal{B}_\pi^\infty$ and $f_4 \in$ BMO$_{\hat\beta\pi}$ and a constant α such that $f = f_3 + f_4 + \alpha$ and $\|f_3\|_\infty \le C_4(\hat\beta)\|f\|_{BMO(\mathbb{R})}$, $\|f_4\|_{BMO(\mathbb{R})} \le C_4(\hat\beta)\|f\|_{BMO(\mathbb{R})}$.*

Finally, we need a theorem about the growth behavior of bandlimited $\mathrm{BMO}(\mathbb{R})$-functions [12].

Theorem 4. *Let $f \in$ BMO$_\sigma$, $0 < \sigma < \infty$. Then, for all $\gamma > \sigma$, there exists a constant C_5 such that $|f(z)| \le C_5\, \mathrm{e}^{\gamma|\mathrm{Im}(z)|} \log(2 + |\mathrm{Re}(z)|)$ for all $z \in \mathbb{C}$.*

B. Basic Properties of Sine-Type Functions

Two important properties of sine-type functions, which will be used in the proof, are stated in Lemmas 1 and 2.

Lemma 1. *Let f be a function of sine type, whose zeros $\{\lambda_k\}_{k\in\mathbb{Z}}$ are ordered increasingly according to their real parts. Then we have*

$$\inf_{k\in\mathbb{Z}} |\lambda_{k+1} - \lambda_k| \ge \underline{\delta} > 0 \tag{4}$$

and

$$\sup_{k\in\mathbb{Z}} |\lambda_{k+1} - \lambda_k| \le \overline{\delta} < \infty \tag{5}$$

for some constants $\underline{\delta}$ and $\overline{\delta}$.

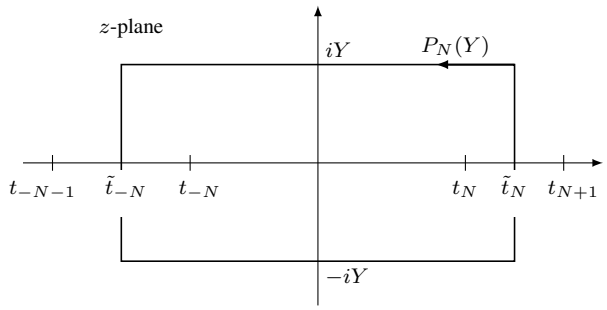


Fig. 1. Integration path $P_N(Y)$ in the complex plane.

Equation (4) follows directly from Definition 2 and the proof of (5) can be found in [8, p. 164].

Lemma 2. *Let f be a function of sine type. For each $\epsilon > 0$ there exists a number $C_6 > 0$ such that*

$$|f(x+iy)| \geq C_6\, \mathrm{e}^{\pi|y|}$$

outside the circles of radius ϵ centered at the zeros of f.

A proof of Lemma 2 can be found in [9, p. 144]. For further information about sine-type functions see for example [8], [9].

C. Proof of Theorem 1

We first prove the second assertion of the theorem. To this end, we extend the proof technique from [13] and [11], which was developed to obtain results, similar to those in this paper, for $\mathcal{B}_\pi^\infty$.

Let ϕ be an arbitrary but fixed sine-type function, whose zeros $\{t_k\}_{k\in\mathbb{Z}}$ are all real and ordered increasingly. Furthermore, let ϕ_k be defined as in (3), and let $0 < \beta < 1$, $f \in \mathrm{BMO}_{\beta\pi}$, and $T > 0$ be arbitrary but fixed. A key equation for the proof is the identity

$$f(t) - \sum_{k=-N}^{N} f(t_k)\phi_k(t) = \frac{1}{2\pi i}\oint_{P_N(Y)} \frac{\phi(t)}{(\zeta - t)}\frac{f(\zeta)}{\phi(\zeta)}\,\mathrm{d}\zeta, \quad (6)$$

which is valid for all $N \in \mathbb{N}$, $Y > 0$, and $t \in \mathbb{R}$ with $\tilde{t}_{-N} < t < \tilde{t}_N$, where

$$\tilde{t}_N = \begin{cases} (t_{N+1} + t_N)/2 & \text{for } N \geq 1 \\ (t_{N-1} + t_N)/2 & \text{for } N \leq -1. \end{cases} \quad (7)$$

The integration path $P_N(Y)$ is depicted in Figure 1. Equation (6) can be easily verified using the residue theorem.

Let

$$\underline{\delta} = \inf_{k\in\mathbb{Z}} |\lambda_{k+1} - \lambda_k|$$

and

$$\overline{\delta} = \sup_{k\in\mathbb{Z}} |\lambda_{k+1} - \lambda_k|.$$

According to Lemma 1, we have $\underline{\delta} > 0$ and $\overline{\delta} < \infty$. Further, let N_0 be the smallest natural number for which $N_0\underline{\delta} > T$. Since $\tilde{t}_N \geq N\underline{\delta}$ for all $N \in \mathbb{N}$, it follows that $\tilde{t}_{N_0} > T$. Furthermore, let $Y_N = N\overline{\delta}$. From (6) we see that

$$\begin{aligned}
&\left| f(t) - \sum_{k=-N}^{N} f(t_k)\phi_k(t) \right| \\
&\quad \leq \frac{1}{2\pi}\int_{-Y_N}^{Y_N} \left|\frac{f(\tilde{t}_N + iy)}{\phi(\tilde{t}_N + iy)}\right| \frac{|\phi(t)|}{|\tilde{t}_N + iy - t|}\,\mathrm{d}y \\
&\qquad + \frac{1}{2\pi}\int_{-Y_N}^{Y_N} \left|\frac{f(\tilde{t}_{-N} + iy)}{\phi(\tilde{t}_{-N} + iy)}\right| \frac{|\phi(t)|}{|\tilde{t}_{-N} + iy - t|}\,\mathrm{d}y \\
&\qquad + \frac{1}{2\pi}\int_{\tilde{t}_{-N}}^{\tilde{t}_N} \left|\frac{f(x + iY_N)}{\phi(x + iY_N)}\right| \frac{|\phi(t)|}{|x + iY_N - t|}\,\mathrm{d}x \\
&\qquad + \frac{1}{2\pi}\int_{\tilde{t}_{-N}}^{\tilde{t}_N} \left|\frac{f(x - iY_N)}{\phi(x - iY_N)}\right| \frac{|\phi(t)|}{|x - iY_N - t|}\,\mathrm{d}x \qquad (8)
\end{aligned}$$

for all $N \geq N_0$ and $t \in [-T, T]$. Next, we treat the integrals on the right hand side of (8) separately. Because of (4) and the definition of $\tilde{t}_N$, it follows that the distance between $\tilde{t}_N$ and the nearest zero of ϕ is at least $\underline{\delta}/2$. Hence, according to Lemma 2, there exists a constant $C_7 > 0$ such that $|\phi(\tilde{t}_N + iy)| \geq C_7\,\mathrm{e}^{\pi|y|}$ for all $y \in \mathbb{R}$. Further, let γ satisfy $\beta\pi < \gamma < \pi$. Then we have

$$|f(\tilde{t}_N + iy)| \leq C_5\,\mathrm{e}^{\gamma|y|}\log(2 + \tilde{t}_N)$$

for all $y \in \mathbb{R}$, according to Theorem 4. Therefore, for the first integral we obtain

$$\begin{aligned}
&\frac{1}{2\pi}\int_{-Y_N}^{Y_N} \left|\frac{f(\tilde{t}_N + iy)}{\phi(\tilde{t}_N + iy)}\right| \frac{|\phi(t)|}{|\tilde{t}_N + iy - t|}\,\mathrm{d}y \\
&\qquad \leq \frac{C_5\log(2 + \tilde{t}_N)\|\phi\|_\infty}{2\pi C_7}\int_{-Y_N}^{Y_N} \frac{\mathrm{e}^{-(\pi-\gamma)|y|}}{|\tilde{t}_N + iy - t|}\,\mathrm{d}y, \\
&\qquad \leq \frac{C_5\log(2 + \tilde{t}_N)\|\phi\|_\infty}{\pi C_7(N\underline{\delta} - T)}\frac{(1 - \mathrm{e}^{-(\pi-\gamma)Y_N})}{(\pi - \gamma)} \\
&\qquad \leq \frac{C_5\log(2 + (N+1)\overline{\delta})\|\phi\|_\infty}{\pi C_7(N\underline{\delta} - T)(\pi - \gamma)}
\end{aligned}$$

for all $N \geq N_0$ and $t \in [-T, T]$. It follows that

$$\lim_{N\to\infty}\max_{t\in[-T,T]} \frac{1}{2\pi}\int_{-Y_N}^{Y_N} \left|\frac{f(\tilde{t}_N + iy)}{\phi(\tilde{t}_N + iy)}\right| \frac{|\phi(t)|}{|\tilde{t}_N + iy - t|}\,\mathrm{d}y = 0. \quad (9)$$

For the second integral in (8) we obtain by the same considerations that

$$\begin{aligned}
&\frac{1}{2\pi}\int_{-Y_N}^{Y_N} \left|\frac{f(\tilde{t}_{-N} + iy)}{\phi(\tilde{t}_{-N} + iy)}\right| \frac{|\phi(t)|}{|\tilde{t}_{-N} + iy - t|}\,\mathrm{d}y \\
&\qquad \leq \frac{C_5\log(2 + (N+1)\overline{\delta})\|\phi\|_\infty}{\pi C_7(N\underline{\delta} - T)(\pi - \gamma)}
\end{aligned}$$

for all $N \geq N_0$ and $t \in [-T, T]$, and consequently

$$\lim_{N\to\infty}\max_{t\in[-T,T]} \frac{1}{2\pi}\int_{-Y_N}^{Y_N} \left|\frac{f(\tilde{t}_{-N} + iy)}{\phi(\tilde{t}_{-N} + iy)}\right| \frac{|\phi(t)|}{|\tilde{t}_{-N} + iy - t|}\,\mathrm{d}y = 0. \quad (10)$$

Next we treat the third integral in (8). Since all zeros of ϕ are real and $Y_N = N\overline{\delta} \geq \overline{\delta}$, it follows from Lemma 2 that there

exists a constant $C_8 > 0$ such that

$$|\phi(x + iY_N)| \geq C_8 \, \mathrm{e}^{\pi Y_N}$$

for all $x \in \mathbb{R}$. Further, we have

$$|f(x + iY_N)| \leq C_5 \, \mathrm{e}^{\gamma Y_N} \log(2 + |x|)$$

for all $x \in \mathbb{R}$, according to Theorem 4. Thus, we obtain

$$\begin{aligned}
&\frac{1}{2\pi} \int_{\tilde{t}_{-N}}^{\tilde{t}_N} \left| \frac{f(x + iY_N)}{\phi(x + iY_N)} \right| \frac{|\phi(t)|}{|x + iY_N - t|} \, \mathrm{d}x \\
&\quad \leq \frac{C_5 \, \mathrm{e}^{\gamma Y_N} \|\phi\|_\infty}{2\pi C_8 \, \mathrm{e}^{\pi Y_N}} \int_{\tilde{t}_{-N}}^{\tilde{t}_N} \frac{\log(2 + |x|)}{|x + iY_N - t|} \, \mathrm{d}x \\
&\quad \leq \frac{C_5 \, \mathrm{e}^{-(\pi-\gamma)Y_N} \|\phi\|_\infty (2N+1)\overline{\delta} \log(2 + (N+1)\overline{\delta})}{2\pi C_8 Y_N} \\
&\quad = \frac{C_5 \, \mathrm{e}^{-(\pi-\gamma)N\overline{\delta}} \|\phi\|_\infty (2N+1)\overline{\delta} \log(2 + (N+1)\overline{\delta})}{2\pi C_8 N\overline{\delta}} \\
&\quad \leq \frac{2C_5 \, \mathrm{e}^{-(\pi-\gamma)N\overline{\delta}} \|\phi\|_\infty \log(2 + (N+1)\overline{\delta})}{\pi C_8}
\end{aligned}$$

for all $N \geq N_0$ and $t \in [-T, T]$, and consequently

$$\lim_{N\to\infty} \max_{t\in[-T,T]} \frac{1}{2\pi} \int_{\tilde{t}_{-N}}^{\tilde{t}_N} \left| \frac{f(x + iY_N)}{\phi(x + iY_N)} \right| \frac{|\phi(t)|}{|x + iY_N - t|} \, \mathrm{d}x = 0. \tag{11}$$

By the same considerations we obtain for the fourth integral in (8) that

$$\begin{aligned}
&\frac{1}{2\pi} \int_{\tilde{t}_{-N}}^{\tilde{t}_N} \left| \frac{f(x - iY_N)}{\phi(x - iY_N)} \right| \frac{|\phi(t)|}{|x - iY_N - t|} \, \mathrm{d}x \\
&\qquad \leq \frac{2C_5 \, \mathrm{e}^{-(\pi-\gamma)N\overline{\delta}} \|\phi\|_\infty \log(2 + (N+1)\overline{\delta})}{\pi C_8}
\end{aligned}$$

for all $N \geq N_0$ and $t \in [-T, T]$, and

$$\lim_{N\to\infty} \max_{t\in[-T,T]} \frac{1}{2\pi} \int_{\tilde{t}_{-N}}^{\tilde{t}_N} \left| \frac{f(x - iY_N)}{\phi(x - iY_N)} \right| \frac{|\phi(t)|}{|x - iY_N - t|} \, \mathrm{d}x = 0. \tag{12}$$

Combining (8), (9), (10), (11), and (12) we see that

$$\lim_{N\to\infty} \max_{t\in[-T,T]} \left| f(t) - \sum_{k=-N}^{N} f(t_k)\phi_k(t) \right| = 0,$$

which proves the second assertion of the theorem.

Next, we prove the first assertion of the theorem. Let $T > 0$ and $f \in \mathrm{BMO}_\pi$ be arbitrary but fixed, and choose some $\hat{\beta}$ with $0 < \hat{\beta} < 1$. According to Theorem 3 there exist two functions $f_3 \in \mathcal{B}_\pi^\infty$ and $f_4 \in \mathrm{BMO}_{\hat{\beta}\pi}$ and a constant α such that $f = f_3 + f_4 + \alpha$ and $\|f_3\|_\infty \leq C_4(\hat{\beta})\|f\|_{\mathrm{BMO}(\mathbb{R})}$. It follows that

$$\begin{aligned}
\left| f(t) - \sum_{k=-N}^{N} f(t_k)\phi_k(t) \right| &\leq \left| f_3(t) - \sum_{k=-N}^{N} f_3(t_k)\phi_k(t) \right| \\
&\quad + \left| f_4(t) - \sum_{k=-N}^{N} f_4(t_k)\phi_k(t) \right| \\
&\quad + \left| \alpha - \sum_{k=-N}^{N} \alpha\phi_k(t) \right|. \qquad (13)
\end{aligned}$$

From Theorem 1 in [11] we know that there exists a constant C_9 such that

$$\sup_{N\in\mathbb{N}} \max_{t\in[-T,T]} \left| f_3(t) - \sum_{k=-N}^{N} f_3(t_k)\phi_k(t) \right| \leq C_9 \|f_3\|_\infty$$

and

$$\sup_{N\in\mathbb{N}} \max_{t\in[-T,T]} \left| \alpha - \sum_{k=-N}^{N} \alpha\phi_k(t) \right| \leq C_9\alpha.$$

For the second term on the right hand side of (13) we have

$$\lim_{N\to\infty} \max_{t\in[-T,T]} \left| f_4(t) - \sum_{k=-N}^{N} f_4(t_k)\phi_k(t) \right| = 0$$

according to the second assertion of the theorem. Thus, it follows that

$$\begin{aligned}
&\sup_{N\in\mathbb{N}} \max_{t\in[-T,T]} \left| f(t) - \sum_{k=-N}^{N} f(t_k)\phi_k(t) \right| \\
&\qquad \leq C_9(\|f_3\|_\infty + \alpha) + C_{10} \\
&\qquad < \infty,
\end{aligned}$$

which completes the proof of the first assertion.

References

[1] P. L. Butzer, W. Splettstößer, and R. L. Stens, "The sampling theorem and linear prediction in signal analysis," *Jahresbericht der Deutschen Mathematiker-Vereinigung*, vol. 90, no. 1, pp. 1–70, Jan. 1988.

[2] C. Fefferman and E. M. Stein, "H^p spaces of several variables," *Acta Mathematica*, vol. 129, no. 1, pp. 137–193, Dec. 1972.

[3] J. B. Garnett, *Bounded Analytic Functions*, S. Eilenberg and H. Bass, Eds. Academic Press, 1981.

[4] D. Gabor, "Theory of communication," *Journal of the Institute of Electrical Engineers*, vol. 93, no. 3, pp. 429–457, Nov. 1946.

[5] L. M. Fink, "Relations between the spectrum and instantaneous frequency of a signal," *Problemy peredachi informatsii*, vol. 2, no. 4, pp. 26–38, 1966 (in Russian), English translation: *Problems of Information Transmission*, vol. 2, no. 4, pp. 11–21.

[6] D. Y. Vakman, "On the definition of concepts of amplitude, phase and instantaneous frequency of a signal," *Radio Engineering and Electronic Physics*, vol. 17, no. 5, pp. 754–759, 1972, English translation from Russian.

[7] B. F. Logan, Jr., "Theory of analytic modulation systems," *Bell System Technical Journal*, vol. 57, no. 3, pp. 491–576, Mar. 1978.

[8] B. Y. Levin, *Lectures on Entire Functions*. AMS, 1996.

[9] R. M. Young, *An Introduction to Nonharmonic Fourier Series*. Academic Press, 2001.

[10] H. Boche and U. J. Mönich, "Global and local approximation behavior of reconstruction processes for Paley-Wiener functions," *Sampling Theory in Signal and Image Processing*, vol. 8, no. 1, pp. 23–51, Jan. 2009.

[11] U. J. Mönich and H. Boche, "Non-equidistant sampling for bounded bandlimited signals," *Signal Processing*, vol. 90, no. 7, pp. 2212–2218, Jul. 2010.

[12] H. Boche and U. J. Mönich, "The structure of bandlimited BMO-functions and applications," *Journal of Functional Analysis*, vol. 264, no. 12, pp. 2637–2675, Jun. 2013.

[13] H. Boche and U. J. Mönich, "Convergence behavior of non-equidistant sampling series," *Signal Processing*, vol. 90, no. 1, pp. 145–156, Jan. 2010.

Bandlimited Signal Reconstruction From the Distribution of Unknown Sampling Locations

Animesh Kumar
Department of Electrical Engineering
Indian Institute of Technology Bombay
Mumbai, India – 400076
Email: animesh@ee.iitb.ac.in

***Abstract*—We study the reconstruction of bandlimited fields from samples taken at unknown but statistically distributed sampling locations. The setup is motivated by distributed sampling where precise knowledge of sensor locations can be difficult.**

Periodic one-dimensional bandlimited fields are considered for sampling. Perfect samples of the field at independent and identically distributed locations are obtained. The statistical realization of sampling locations is *not known*. First, it is shown that a bandlimited field cannot be uniquely determined with samples taken at statistically distributed but unknown locations, even if the number of samples is infinite. Next, it is assumed that the order of sample locations is known. In this case, using insights from order-statistics, an estimate for the field with useful asymptotic properties is designed. Distortion (mean-squared error) and central-limit are established for this estimate.

I. Introduction

In the smart-dust paradigm [1], consider a distributed field sampling problem where sensors are deployed without precise control on the sensor-locations. One method for distributed field sampling is to learn the location of these individual sensors, and then reduce field acquisition to the well-studied non-uniform sampling problem [2]. However, localization of individual sensors in a wireless sensor network can be difficult [3]. In light of these issues, the reconstruction of a physical field from samples taken at *unknown* but statistically distributed locations is studied in this work.

Assuming that the field has a finite support, sensors will have to be deployed in the finite region where the field is non-zero. The smoothness of the physical field can be modeled by bandlimitedness. In this work, it will be assumed that the field is spatially periodic and bandlimited. Only *one-dimensional fields* will be considered. The lack of control in sensor deployment is modeled by a uniform-distribution on the sensor or sampling-locations. It is assumed that sensors are deployed (or scattered) independent of each other. Thus, perfect samples of the field at independent and identically distributed (i.i.d.) but also *unknown* locations are obtained by the sampling method outlined above. From these samples the field has to be estimated. This work focuses on a *consistent estimate*, that is, an estimate which converges to the true underlying field when the number of samples is infinite.

This work has been supported by grant no. P09IRCC039, IRCC, IIT Bombay.

The key results shown in this work are as follows:

1) It will be shown that a bandlimited field *cannot* be uniquely determined with perfect samples obtained at statistically distributed locations, even if the number of samples is infinite.
2) If the order of sample locations is known, then using insights from classical order-statistics, a consistent estimate for the spatial field is presented. Distortion (average mean-squared error) and a central-limit type weak convergence result are established for this estimate.

Prior art: Recovery of discrete-time bandlimited signals from samples taken at unknown locations was first studied by Marziliano and Vetterli [4]. Recovery of a bandlimited signal from a finite number of ordered nonuniform samples at unknown sampling locations has been studied by Browning [5]. Estimation of periodic bandlimited signals in the presence of random sampling location under two models has been studied by Nordio et al. [6]. Their first model studies reconstruction of bandlimited signal affected by noise at random but known locations. Their second model studies estimation of bandlimited signal from noisy samples on a location set obtained by random perturbation of equi-spaced deterministic grid.

In contrast, this work presents the estimation of a bandlimited field from i.i.d. distributed but unknown samples in an asymptotic setting (where the number of samples increases to infinity). Asymptotic consistency (convergence in probability), mean-squared error bounds, and central-limit type weak law are the focus of this work. The first key-result of this work is absent in related work due to difference in the sensing model. *Organization:* In Section II the field model, distortion, sensor deployment model, and useful statistical theory are outlined. In Section III asymptotic consistency, mean-squared error, and weak convergence aspects of field estimate are discussed. Finally, conclusions will be presented in Section IV.

II. Problem setup and useful classical results

This section will review the field model, the distortion, and some useful mathematical results. Field model appears first.

A. Field model and associated properties

The field of interest $g(t)$ is periodic, real-valued, and bandlimited. Without loss of generality, the period is assumed to be $T = 1$. It is also assumed that the field $|g(t)| \leq 1$ is

bounded. Bandlimitedness implies that some $b > 0$ coefficients are non-zero in the Fourier series. Thus,

$$g(t) = \sum_{k=-b}^{b} a_k \exp(j2\pi kt). \quad (1)$$

Real-valued $g(t)$ implies conjugate symmetry in the Fourier domain, that is, $a_k = a_{-k}^*$; however, this symmetry will not be utilized in this work. The $(b+1)$ Fourier coefficients constitute the degrees of freedom for this signal. With $||g||_\infty \leq 1$, using Bernstein's inequality [7], we get

$$|g'(t)| \leq 2\pi b, \quad (2)$$

where $2\pi b$ rad/s is the bandwidth of the signal. For simplicity of notation, define $s_b := 1/(2b+1)$ as a spacing parameter and $\phi_k := \exp(j2\pi k/(2b+1)) = \exp(j2\pi k s_b)$. By using $(2b+1)$ samples of the field $g(t)$, its Fourier coefficients can be computed as follows:

$$\begin{bmatrix} g(0) \\ g(s_b) \\ \vdots \\ g(2bs_b) \end{bmatrix} = \begin{bmatrix} 1 & \dots & 1 \\ \phi_{-b} & \dots & \phi_b \\ \vdots & & \vdots \\ (\phi_{-b})^{2b} & \dots & (\phi_b)^{2b} \end{bmatrix} \begin{bmatrix} a_{-b} \\ a_{-b+1} \\ \vdots \\ a_b \end{bmatrix}$$

or more simply

$$\vec{g} = \Phi_b \vec{a}, \quad (3)$$

where the vector matrix notation is obvious. The columns of Φ_b are orthogonal with a norm-square $(2b+1)$ under the standard inner-product. The relation in (3) is inverted to obtain

$$\vec{a} = (\Phi_b)^{-1}\vec{g} = \frac{1}{(2b+1)}\Phi_b^\dagger \vec{g}, \quad (4)$$

where $\Phi_b^\dagger$ is the conjugate transpose of Φ_b. The expression in (4) will be used to obtain an estimate for $\vec{a}$ as discussed later.

B. Sensor deployment model and reconstruction distortion

Denote any sequence as $x_l^m := (x_l, x_{l+1}, \dots, x_m)$ for $m \geq l$. It will be assumed that sensors are deployed at random locations U_1^n in the interval of interest $[0, 1]$. The locations U_1^n are i.i.d. random variables with uniform distribution and probability density function $f(u) = 1$ for $0 \leq u \leq 1$. The locations U_1^n are *not known* in our model. An asymptotic number of samples and limiting distribution of U_1^n will be used for field estimation. The average mean-squared error will be used as a distortion metric. If $\widehat{G}(t)$ is any estimate of $g(t)$, then the distortion is defined as

$$D := \mathbb{E}(||\widehat{G} - g||_2^2) := \mathbb{E}\left[\int_0^1 |\widehat{G}(t) - g(t)|^2 \mathrm{d}t\right]. \quad (5)$$

C. Useful mathematical results

For estimation of field from the statistical properties of U_1^n, the following convergence results will be useful. These results for order-statistics and quantiles are a counterpart to the strong-law of large numbers (see [8, Ch. 10]). The ordered version of U_1^n will be denoted by $U_{1:n}^{n:n} := \{U_{1:n}, U_{2:n}, \dots, U_{n:n}\}$ where $U_{n:n}$ is the largest and $U_{1:n}$ is the smallest [8].

For uniform distribution, the p-th population quantile q_p is equal to p. Then with $r = [np]+1$, it is known that [8, pg. 285]

$$U_{r:n} - p = -(F_n(p) - p) + R_n, \quad (6)$$

where $F_n(u) := \frac{1}{n}\sum_{i=1}^n \mathbb{1}(U_i \leq u)$ is the empirical distribution of U_1^n. The remainder term R_n decreases to 0 almost surely,

$$R_n = O\left(n^{-3/4}(\log n)^{1/2}(\log\log n)^{1/4}\right) \quad \text{as } n \to \infty. \quad (7)$$

By the strong law of large numbers [9], we know that $F_n(p) \xrightarrow{a.s.} p$; thus, $U_{r:n} \xrightarrow{a.s.} p$ from (7). Analogous to the central limit theorem, the following fact is noted.

Fact 2.1: [8, Theorem 10.3] Let $0 < p_1 < p_2 < \dots < p_{2b+1} < 1$ and assume that $(r_i/n - p_i) = o(1/\sqrt{n}), i = 1, 2, \dots, 2b+1$. Then the following result holds:

$$\sqrt{n}[U_{r_1:n} - p_1, \dots, U_{r_{2b+1}:n} - p_{2b+1}]^T \xrightarrow{d} \mathcal{N}\left(\vec{0}, K_U\right),$$

where $[K_U]_{j,j'} = p_j(1 - p_{j'})$ for $j \leq j'$.

All the moments of U are finite since it is bounded (by definition). The second moment of $U_{r:n} - p$, with $r \approx [np]$ is bounded by.

$$\begin{aligned} n\mathbb{E}(U_{r:n} - p)^2 &= p(1-p)\mathbb{E}(Z^2) + O(\sqrt{1/n}), \\ &\leq \frac{1}{4} + O(\sqrt{1/n}). \end{aligned} \quad (8)$$

where $Z \sim \mathcal{N}(0, 1)$ is a normalized Gaussian random variable.

The following fact relates consistency and $\mathcal{L}^2$ convergence.

Fact 2.2: [9] If $X_n \xrightarrow{a.s.} X$ and $Y_n \xrightarrow{a.s.} Y$, then $aX_n + bY_n \xrightarrow{a.s.} aX + bY$ for any constants $a, b \in \mathbb{R}$. If X_n is bounded and $X_n \xrightarrow{a.s.} X$, then $X_n \xrightarrow{\mathcal{L}^2} X$.

We now proceed to the main results of this paper.

III. Sampling and estimation with random samples

In this section, the key results of this work are presented. It will be shown that the field $g(t)$ cannot be inferred uniquely from samples collected at U_1^∞, where sample-locations are unknown. Further, with order information on sample-locations, consistent estimation of the field is presented.

A. It is impossible to infer $g(t)$ uniquely from U_1^∞

It will be shown that if $g(U_1), \dots, g(U_n)$ is available without the knowledge of U_1^n, then $g(t)$ cannot be inferred uniquely as $n \to \infty$. Consider the statistic

$$F_{g,n}(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}(g(U_i) \leq x), \quad (9)$$

where $\mathbb{1}()$ are the indicator random variables. Then $F_{g,n}(x), x \in [-1, 1]$ completely characterizes the field values $g(U_1), \dots, g(U_n)$ and vice-versa. By Glivenko-Cantelli theorem, the right hand limit in (9) converges almost surely to $\mathbb{P}(g(U) \leq x)$ for all $x \in [-1, 1]$ as $n \uparrow \infty$ [10]. This limit is explained using Fig. 1. For any $x \in [-1, 1]$ the set of points where $g(t) \leq x$ can be marked on the t-axis. The length or measure of this set is equal to $\mathbb{P}(g(U) \leq x)$.

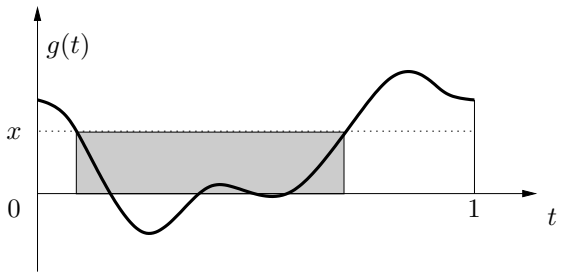


Fig. 1. *For any $x \in [-1,1]$ the set of points where $g(t) \leq x$ can be marked on the t-axis. The length or measure of this set is equal to $\mathbb{P}(g(U) \leq x)$.*

For $0 < \theta < 1$, let $g_\theta(t) = g(t-\theta)$, i.e., $g_\theta(t)$ is the shifted version of $g(t)$. Since $g(t)$ is periodic, its shifts will be cyclic in nature in the period $[0,1]$. Thus, the level-sets of $g(t-\theta)$ will be cyclic (in θ) and its measure $\{u : g_\theta(u) \leq x\}$ will be independent of θ for every $x \in [-1,1]$. Therefore, $\mathbb{P}(g_\theta(U) \leq x)$ will be independent of θ for every $x \in [-1,1]$. Thus, by only using $F_{g,n}(x)$, which converges to $\mathbb{P}(g(U) \leq x), x \in [-1,1]$, the field $g(t)$ cannot be inferred uniquely. This completes the discussion of this subsection.

B. Consistent estimation of $g(t)$ from $U_{1:n}^{n:n}$

From now on, it will be assumed that order information of samples is available. That is, samples $g(U_{1:n}), \ldots, g(U_{n:n})$ are available and $g(t)$ has to be estimated. Using (4) and the convergence results in Sec. II-C, the following estimate for the Fourier series coefficients of $g(t)$ is proposed:

$$\vec{A} := [\widehat{A}_{-b}, \widehat{A}_{-b+1}, \ldots, \widehat{A}_b]^T := \frac{1}{(2b+1)} \Phi_b^\dagger \vec{G}. \tag{10}$$

where $\vec{G} = [g(U_{1:n}), g(U_{[ns_b]+1:n}), \ldots, g(U_{[n2bs_b]+1:n})]^T$. With (6) and the smoothness properties (continuity) of $g(t)$, this estimate is obtained by substitution method in (4). Using $\vec{A}$, an estimate for $g(t)$ is obtained as follows

$$\widehat{G}(t) = \sum_{k=-b}^{b} \widehat{A}_k \exp(j2\pi kt) = \Phi(t)^T \vec{A} \tag{11}$$

where $\Phi(t)^T = \begin{bmatrix} \exp(-j2\pi bt) & \ldots & \exp(j2\pi bt) \end{bmatrix}$. Intuitively, $g(t)$ has a finite degrees of freedom. This enables a procedure to estimate the Fourier series coefficients (the degrees of freedom) from a finite number of sample estimates of $g(t)$. Using these estimates of the Fourier series coefficients, the entire field of interest $g(t)$ can be estimated. For distortion calculation, the Parseval's theorem [11] will be useful,

$$||\widehat{G} - g||_2^2 = \sum_{k=-b}^{b} |\widehat{A}_k - a_k|^2. \tag{12}$$

A bound on $\mathbb{E}(|\widehat{A}_k - a_k|^2)$ will result in a bound on the expected mean-squared error $\mathbb{E}(||\widehat{G} - g||_2^2)$.

We state our first result now.

Theorem 3.1 (Consistency of $\vec{A}$): Let $U_{1:n}^{n:n}$ be ordered i.i.d. Uniform$[0,1]$ random variables. Define $\vec{A}$ and $\widehat{G}(t)$ as in (10) and (11). Then the estimates $\vec{A}$ and $\widehat{G}(t)$ are consistent in almost-sure and $\mathcal{L}^2$ sense to their respective limits, i.e.,

$$\vec{A} \xrightarrow{a.s.} \vec{a}, \widehat{G}(t) \xrightarrow{a.s.} g(t) \text{ and } \vec{A} \xrightarrow{\mathcal{L}^2} \vec{a}, \widehat{G}(t) \xrightarrow{\mathcal{L}^2} g(t). \tag{13}$$

Proof: Only a sketch is provided due to space constraints. First note that $U_{[nis_b]+1:n} \xrightarrow{a.s.} is_b$ for each $i = 0, 1, \ldots, 2b$. Since $g(t)$ is continuous by assumption, $g(U_{[nis_b]+1:n}) \xrightarrow{a.s.} g(is_b)$ for each $i = 0, 1, \ldots, 2b$. Let $\vec{G} := [g(U_{1:n}), g(U_{[ns_b]+1:n}), \ldots, g(U_{[n2bs_b]+1:n})]^T$ and $\vec{g} := [g(0), g(s_b), \ldots, g(2bs_b)]^T$. By repeated use of Fact 2.2, any finite linear combination $\vec{c}^T\vec{G}$ converges almost-surely to $\vec{c}^T\vec{g}$. Thus, from (10), each element of $\vec{A}$ converges almost surely to $\vec{a}$. Hence, $\vec{A} \xrightarrow{a.s.} \vec{a}$.

Next, $\widehat{G}(t)$ is a finite linear combination of $\vec{A}$. Since $\vec{A} \xrightarrow{a.s.} \vec{a}$, therefore, $\widehat{G}(t) \xrightarrow{a.s.} g(t)$ in a similar fashion as above.

For $\mathcal{L}^2$-convergence, note that $\vec{G}$ is bounded in each coordinate since $|g(t)| \leq 1$ for all t. Each element of the matrix Φ_b has a magnitude one. Thus, by (10) and the triangle inequality, $|\widehat{A}_i| \leq ||g||_\infty \leq 1$ for every $i = -b, -b+1, \ldots, b$. Thus, each $\widehat{A}_i$ is a bounded random variable. For bounded random sequences, from Fact 2.2, $\vec{A} \xrightarrow{a.s.} \vec{a}$ implies that $\vec{A} \xrightarrow{\mathcal{L}^2} \vec{a}$. Similarly, $|\widehat{G}(t)| \leq \sum_{k=-b}^{b} |\widehat{A}_k| \leq (2b+1)$ from (11). Thus, by Fact 2.2, $\widehat{G}(t) \xrightarrow{a.s.} g(t)$ implies $\widehat{G}(t) \xrightarrow{\mathcal{L}^2} g(t)$, since $\widehat{G}(t)$ is bounded. ■

The second result establishes the scaling of distortion for the proposed estimate in (11).

Theorem 3.2: Let $U_{1:n}^{n:n}$ be ordered i.i.d. Uniform$[0,1]$ random variables. Define $\vec{A}$ and $\widehat{G}(t)$ as in (10) and (11). Then,

$$n\mathbb{E}\left[||\widehat{G} - g||_2^2\right] \leq \pi^2 b^2 (2b+1)\left[1 + O(\sqrt{1/n})\right], \tag{14}$$

that is, the expected distortion decreases as $O(1/n)$.

Proof: The proof is presented in two parts. First, using the smoothness properties of $g(t)$, the norm $||\widehat{G} - g||_2^2$ will be bounded using the error in quantiles $U_{[np]+1:n} - p$. Next, the convergence rate of $U_{[np]+1:n} - p$ as in (8) will be utilized to upper-bound the distortion. First note that

$$\begin{aligned}
||\widehat{G} - g||_2^2 &= \sum_{k=-b}^{b} |\widehat{A}_k - a_k|^2 && (15)\\
&= \frac{1}{(2b+1)^2} ||\Phi_b^\dagger(\vec{G} - \vec{g})||_2^2 && (16)\\
&= \frac{1}{(2b+1)^2} \sum_{k=-b}^{b} \left| \sum_{l=0}^{2b} [\phi_k^l]^* (\widehat{G}(ls_b) - g(ls_b)) \right|^2 \\
&\overset{(a)}{\leq} \frac{(2b+1)}{(2b+1)^2} \sum_{k=-b}^{b} \sum_{l=0}^{2b} |\phi_k^l| |\widehat{G}(ls_b) - g(ls_b))|^2 \\
&\overset{(b)}{=} \frac{1}{(2b+1)} \sum_{k=-b}^{b} \sum_{l=0}^{2b} |\widehat{G}(ls_b) - g(ls_b))|^2 && (17)\\
&\overset{(c)}{\leq} \sum_{l=0}^{2b} |\widehat{G}(ls_b) - g(ls_b))|^2. && (18)
\end{aligned}$$

$$= \sum_{l=0}^{2b} |g(U_{[nls_b]+1:n}) - g(ls_b))|^2. \quad (19)$$

$$\leq ||g'||_\infty^2 \sum_{l=0}^{2b} |U_{[nls_b]+1:n} - ls_b|^2. \quad (20)$$

where (a) follows by $(a_1 + a_2 + \ldots + a_n)^2 \leq n(a_1^2 + a_2^2 + \ldots + a_n^2)$, (b) follows by $|\phi_k| = 1$ for all k, and (c) follows since the summation does not depend on k. Using (8), and taking expectations on both sides

$$n\mathbb{E}\left(||\widehat{G} - g||_2^2\right) \leq ||g'||_\infty^2 \sum_{l=0}^{2b} n\mathbb{E}\left(|U_{[nls_b]+1:n} - ls_b|^2\right)$$

$$\leq ||g'||_\infty^2 \sum_{l=0}^{2b} \left[\frac{1}{4} + O(\sqrt{1/n})\right] \quad (21)$$

$$\leq (2\pi b)^2(2b+1)\frac{1}{4} + O(\sqrt{1/n}) \quad (22)$$

$$= \pi^2 b^2(2b+1)[1 + O(\sqrt{1/n})]. \quad (23)$$

This completes the proof. ■

The third result establishes the weak-convergence of $\widehat{G}(t)$.

Theorem 3.3 (Central limit for $\widehat{G}(t)$): Let $U_{1:n}^{n:n}$ be ordered i.i.d. Uniform$[0,1]$ random variables and $\vec{u} = (0, s_b, 2s_b, \ldots 2bs_b)^T$. Define $\vec{A}$ and $\widehat{G}(t)$ as in (10) and (11). Then the estimate $\vec{A}$ and $\widehat{G}(t)$ satisfy the following central limits:

$$\sqrt{n}(\vec{A} - \vec{a}) \xrightarrow{d} \mathcal{N}\left(\vec{0}, K_A\right). \quad (24)$$

where $K_G = \nabla g^T(\vec{u}) K_U \nabla g(\vec{u})$ and $K_{\vec{A}}$ is independent of n and given in terms of K_G and Φ_b. Further,

$$\sqrt{n}(\widehat{G}(t) - g(t)) \xrightarrow{d} \mathcal{N}\left(\vec{0}, K_G(t)\right). \quad (25)$$

where $K_G(t)$ is independent of n and given in terms of K_G and Φ_b.

Proof: From Fact 2.1, we know that $\vec{U} := [U_{1:n}, U_{[ns_b]+1:n}, \ldots, U_{[n2bs_b]+1:n}]^T$ is asymptotically normal. That is, $\sqrt{n}(\vec{U} - \vec{u}) \xrightarrow{d} \mathcal{N}(\vec{0}, K)$, where

$$[K]_{i,i'} = (i-1)s_b[1 - (i'-1)s_b] \text{ for } i \leq i'. \quad (26)$$

Note that $[K]_{i,i'} = [K]_{i',i}$ by the symmetry of a covariance matrix. Recall $\vec{G}$ from (10). Since $g(t)$ is a differentiable field, by the delta-method [10],

$$\sqrt{n}(\vec{G} - \vec{g}) \xrightarrow{d} \mathcal{N}(\vec{0}, K_{\vec{G}}), \quad (27)$$

where $K_{\vec{G}} = \nabla g(\vec{u})^T K \nabla g(\vec{u})$. Observe that the matrix $K_{\vec{G}}$ depends on the field $g(t)$. However, by smoothness of $g(t)$, the vector $\nabla g(\vec{u})$ is bounded and K is independent of n. Thus, $K_{\vec{G}}$ is independent of n. From (10), since $\vec{A}$ is obtained from $\vec{G}$ by a complex-valued linear transformation, we get

$$\sqrt{n}(\vec{A} - \vec{a}) \xrightarrow{d} \mathcal{CN}(\vec{0}, K_{\vec{A}}). \quad (28)$$

Observe that the limit is a complex normal Gaussian vector. In general, the covariance properties of a zero-mean complex random vector $\vec{S}$ are determined by $\mathbb{E}(\vec{S}\vec{S}^\dagger)$ and $\mathbb{E}(\vec{S}\vec{S}^T)$. Thus, $K_{\vec{A}}$ is determined by the two matrices $\frac{1}{(2b+1)^2}\Phi_b^\dagger K_{\vec{G}} \Phi_b$ and $\frac{1}{(2b+1)^2}\Phi_b^\dagger K_{\vec{G}} \Phi_b^T$. The covariance matrix $K_{\vec{G}}$ is independent of n; therefore, $K_{\vec{A}}$ is also independent of n and well defined.

Finally, $\widehat{G}(t)$ is obtained from $\vec{A}$ by a t-dependent inner product. From (11), we get $\widehat{G}(t) = \Phi(t)^T \vec{A}$. Therefore, $\widehat{G}(t)$ is a complex normal Gaussian vector. Its variance can be determined by $\mathbb{E}(\widehat{G}(t)^2)$ and $\mathbb{E}(|\widehat{G}(t)|^2)$ which are equal to $\frac{1}{(2b+1)^2}\Phi(t)^T \Phi_b^\dagger K_{\vec{G}} \Phi_b^T \Phi(t)$ and $\frac{1}{(2b+1)^2}\Phi(t)^T \Phi_b^\dagger K_{\vec{G}} \Phi_b \Phi(t)^\dagger$, respectively. Thus the proof is complete. ■

This completes our technical result section. The estimation technique outlined in this section holds well for noise-free setting. If there is additive noise affecting the samples, then more involved estimation techniques will be required. Obtaining consistent estimates for $g(ls_b), l = 0, \ldots, (2b+1)$ is more challenging in the presence of noise.

IV. Conclusions

The reconstruction of bandlimited fields from samples taken at unknown but statistically distributed sampling locations was studied. Periodic one-dimensional bandlimited fields were considered for sampling. Perfect samples of the field at i.i.d. uniform locations were used for the reconstruction. It was shown that a bandlimited field cannot be uniquely determined only with samples taken at statistically distributed locations, even if the number of samples is infinite. Using order information on the sample locations, a consistent estimate was proposed for the underlying field. It was shown that this estimate converges in the mean-squared error sense and almost-sure sense. Further, the mean-squared error asymptotically decreases as $O(1/n)$, where n is the number of obtained field samples.

References

[1] J. M. Kahn, R. H. Katz, and K. S. J. Pister, "Next century challenges: Mobile networking for "smart dust"," in *ACM International Conference on Mobile Computing and Networking (MOBICOM)*, Aug 1999, pp. 271–278. [Online]. Available: citeseer.nj.nec.com/kahn99next.html

[2] Farokh Marvasti (ed.), *Nonuniform Sampling*. New York, USA: Kluwer Academic Publishers, 2001.

[3] N. Patwari, J. N. Ash, S. Kyperountas, A. O. H. III, R. L. Moses, and N. S. Correal, "Location the nodes: Cooperative localization in wireless sensor networks," *IEEE Signal Processing Magazine*, vol. 22, no. 4, pp. 54–69, Jul. 2005.

[4] P. Marziliano and M. Vetterli, "Reconstruction of irregularly sampled discrete-time bandlimited signals with unknown sampling locations," *IEEE Transactions on Signal Processing*, vol. 48, no. 12, pp. 3462–3471, Dec. 2000.

[5] J. Browning, "Approximating signals from nonuniform continuous time samples at unknown locations," *IEEE Transactions in Signal Processing*, vol. 55, no. 4, pp. 1549–1554, Apr. 2007.

[6] A. Nordio, C.-F. Chiasserini, and E. Viterbo, "Performance of linear field reconstruction techniques with noise and uncertain sensor locations," *IEEE Transactions on Signal Processing*, vol. 56, no. 8, pp. 3535–3547, Aug. 2008.

[7] G. H. Hardy, J. E. Littlewood, and G. Polya, *Inequalities*. London, UK: Cambridge University Press, 1959.

[8] H. A. David and H. N. Nagaraja, *Order Statistics*, 3rd ed. New York, NY: John Wiley & Sons, 2003.

[9] R. Durrett, *Probability: Theory and Examples*, 2nd ed. Belmont, CA: Duxbury Press, 1996.

[10] A. W. van der Vaart, *Asymptotic Statistics*. Cambridge, UK: Cambridge University Press, 1998.

[11] Alan Oppenheim and Alan Willsky and Hamid Nawab, *Signals and Systems*, 2nd ed. USA: Prentice Hall, 1996.

Sampling aspects of approximately time-limited multiband and bandpass signals

Joseph Lakey
Department of Mathematical Sciences
New Mexico State University
Las Cruces, NM 88003–8001
Email: jlakey@nmsu.edu

Jeffrey A. Hogan
School of Mathematical and Physical Sciences
University of Newcastle
Callaghan, NSW 2308
Australia
Email: Jeff.Hogan@newcastle.edu.au

***Abstract*—We provide an overview of recent progress regarding the role of sampling in the study of signals that are in the image of a bandpass or multiband frequency limiting operation and have most of their energies concentrated in a given time interval. We finish considering a means to approximate essentially time-limited bandpass signals. In this case we present a new phase-locking metric that arises in the study of EEG signals.**

I. Introduction

We discuss relationships between time and band limiting and sampling, leading also to numerical computation of essentially time-limited multiband and bandpass signals. As an application we propose a method to analyze phase synchrony of bandpass projections of signals, illustrating a particular case of electroencephalographic (EEG) signals. In this introduction we briefly review basic elements of the theory of time and band limiting. In Section II we discuss connections between sampling and time and band limiting. In Section III we present a method to construct time- and multiband-limited signals from eigenfunctions for time and band limiting to separate bands and a numerical technique that takes advantage of sampling. In Section IV we provide a method to approximate essentially time-limited bandpass signals. We use this approach in Section V to provide a new method to study phase differences of bandpass projections of signals. The method is illustrated in the context of study of EEG signals. Relatively constant phase lag among two EEG channels can indicate recruitment of the corresponding cortical regions in distributed cognition.

A. Time and band limiting

Set $(Q_T)(f)(t) = \mathbb{1}_{[-T,T]}(t)\, f(t)$ where $\mathbb{1}_S$ denotes the function equal to one on $S \subset \mathbb{R}$ and zero outside S. Let $Q = Q_1$. Also let $(P_\Sigma)(f)(t) = (\mathbb{1}_\Sigma \widehat{f})^\vee(t)$ where $\widehat{f}(\xi) = \int_{-\infty}^{\infty} f(t)\, \mathrm{e}^{-2\pi i t\xi}\, \mathrm{d}t$. We write $P_\Omega = P_{[-\Omega/2,\Omega/2]}$ and $P = P_1$. The *Paley–Wiener* space PW_Σ is of the image of $L^2(\mathbb{R})$ under the orthogonal projection P_Σ. We write PW_Ω instead when $\Sigma = [-\Omega/2, \Omega/2]$ and simply PW when $\Omega = 1$. For compact Σ, the operator $P_\Sigma Q_T$ is compact and its trace is equal to $2T|\Sigma|$ where $|\Sigma|$ denotes the Lebesgue measure of $\Sigma \subset \mathbb{R}$. It is also self adjoint on PW_Σ while $P_\Sigma Q_S P_\Sigma$ is self adjoint on $L^2(\mathbb{R})$. Since functions in PW_Σ are real analytic, $P_\Sigma Q_S$ has no unit eigenfunctions and the discrete spectrum of $P_\Sigma Q_T$ is contained in $[0, 1)$.

B. Prolate functions and their properties

The operator $P_{c/\pi}Q$ commutes with a certain self-adjoint second-order differential operators whose eigenfunctions, and hence those of $P_{c/\pi}Q$, are the *prolate spheroidal wave functions*, which form a complete orthogonal basis for $\mathrm{PW}_{c/\pi}$. They are also eigenfunctions of the integral operator

$$(F_c f)(t) = \int_{-1}^{1} \mathrm{e}^{\mathrm{i}cst}\, f(s)\, \mathrm{d}s = \widehat{Qf}(-ct/2\pi)\,. \qquad (1)$$

The eigenvalues $P_{c/\pi}Q$ are non degenerate. Denote by $\lambda_0(c) > \lambda_1(c) > \ldots$ the nth eigenvalue of $P_{c/\pi}Q$ and φ_n^c the corresponding prolate eigenfunction. That φ_n^c is an eigenfunction of (1) and other basic properties imply that

$$D_{c/\pi}\widehat{\varphi_n^c} = \frac{\mathrm{i}^n}{\sqrt{\lambda_n}} Q\varphi_n^c \qquad (2)$$

where D_a is the unitary dilation $(D_a f)(t) = \sqrt{a} f(at)$, $a > 0$. When $L^2(\mathbb{R})$-normalized, the prolates $\{\varphi_n^c\}$ form an orthonormal basis for $\mathrm{PW}_{c/\pi}$ as well as a complete, orthogonal set in $L^2[-1, 1]$ with $\lambda_n(c) = \int_{-1}^{1} |\varphi_n^c|^2$. As such, any $f \in \mathrm{PW}_{c/\pi}$ can be expanded in the form $f = \sum_{n=0}^{\infty} \alpha_n\, \varphi_n^c$ with $\|f\|^2_{L^2(\mathbb{R})} = \sum \alpha_n^2$ and $\int_{-1}^{1} |f|^2 = \sum \lambda_n \alpha_n^2$. The prolates are real valued and φ_n^c is even (odd) if n is even (odd). Further properties of prolates and justification of the facts just mentioned, which were established in the Bell Labs papers, especially [1], [2], can be found in [3].

C. The $2\Omega T$ theorem

Suppose that Σ is a union of M pairwise disjoint frequency intervals of unit length so that the total time–bandwidth product corresponding to $P_\Sigma Q_T$ is $2MT$. Denote by $\mathcal{N}(2MT, \alpha)$ the number of eigenvalues of $P_\Sigma Q_T$ larger than α. The following is a special case of a version of the "$2\Omega T$" theorem proved by Landau and Widom in [4].

Theorem 1 (Landau–Widom, 1980): As $T \to \infty$ the number of eigenvalues of $P_\Sigma Q_T$ exceeding $\alpha \in (0, 1)$ satisfies

$$\mathcal{N}(2MT, \alpha) = 2MT + \frac{M}{\pi^2} \log 2T \log\Big(\frac{\alpha}{1-\alpha}\Big) + o(\log 2MT)\,.$$

II. Sampling and time and band limiting

Walter and Shen [5] and Khare and George [6] observed

$$(PQ_Tf)(t) = \sum_{n=0}^{\infty} \lambda_n \sum_{k=-\infty}^{\infty} f(k)\varphi_n(k)\,\varphi_n(t)$$

where φ_n are eigenfunctions of PQ_T. Oscillatory behavior of the prolates near the endpoints of $[-T,T]$ prohibits an estimate $\sum_{|k|>T}\varphi_n^2(k) \le C(T)(1-\lambda_n)$. However, in [7] the estimate

$$\sum_{|k|>\pi^2 T(1+\log^\gamma(T))} \varphi_n^2(k) \le C(1-\lambda_n), \tag{3}$$

was proved for any $\gamma > 1$. It was conjectured in [7] that the log factor in the sum index is not necessary. The following consequence of (3) regarding approximation of Q_Tf from samples of f near $[-T,T]$ was also established in [7].

Theorem 2: Let $f \in \operatorname{span}\{\varphi_n\}_{n=0}^N$, with φ_n the nth eigenfunction of PQ_T. Define $\varphi_n^T = \sum_{|k|<M(T)} \varphi_n(k)\mathrm{sinc}\,(t-k)$ with $M(T)$ as in (3). Then

$$\|Q_T(f - \sum_{n=0}^{N}\langle f, \varphi_n^T\rangle\varphi_n^T)\| \le C\|f\| \sum_{n=0}^{N}\lambda_n(1-\lambda_n)\,.$$

A method to obtain accurate numerical estimates of integer samples of prolates is outlined in Hogan et al., [7].

III. Time- and multiband-limited signals

This section reviews techniques underlying numerical computation of certain time- and multiband-limited signals. We start with a method to build eigenfunctions for the case Σ is a finite union of intervals from appropriately modulated prolates.

A. Eigenfunctions for unions

If Σ is a finite union of pairwise disjoint intervals $I_1,\dots,I_M$ then we can denote $P_\Sigma = \sum_{k=1}^M P_{I_k}$. Unlike PQ_T, the operator $P_\Sigma Q_T$ does not commute with a finite order differential operator with polynomial coefficients when Σ is a union of two or more intervals. This important fact, established by Morrison in [8], bars us from using power series methods to compute eigenfunctions.

The following results were established in [9] in a more general setting. If J is a frequency interval of unit length then the orthogonal projection onto PW_J, the Paley–Wiener subspace of $L^2(\mathbb{R})$ of functions frequency supported in J, has the form $M_{m_J}PM_{-m_J}$ where, as before, $P = P_{[-1/2,1/2]}$ and $(M_uf)(t) = \mathrm{e}^{2\pi itu}f(t)$ with m_{J_k} the midpoint of J_k. Suppose that one has M pairwise disjoint frequency intervals $J_1,\dots,J_M$ each of unit length and set $\Sigma = \cup_k J_k$. Set $m_k = m_{J_k}$. Since the J-prolates $\varphi_n^J = M_{m_J}\varphi_n$, with φ_n the corresponding eigenfunction of PQ_T, form a complete family for PW_J, any function in PW_Σ has an orthogonal decomposition $f = \sum_{k=1}^M\sum_{n=0}^\infty \langle f,\, M_{m_k}\varphi_n\rangle M_{m_k}\varphi_n$. Consider now the problem of finding an eigenvalue–eigenfunction pair $(\lambda,\,\psi)$ for $P_\Sigma Q_T$. Expanding ψ in terms of the modulated prolates $M_{m_k}\varphi_n$ and applying $P_\Sigma Q_T$ to these, one sees that one must identify the coefficients $\Gamma_{nm}^{k,\ell} = \langle Q_TM_{m_k}\varphi_n,\, M_{m_\ell}\varphi_m\rangle$. Note that $\Gamma_{mn}^{\ell,k} = \overline{\Gamma_{mn}^{k,\ell}}$, that is, if $\Gamma^{k,\ell}$ is the matrix with entries $\Gamma_{nm}^{k,\ell}$ then $\Gamma^{\ell,k} = \overline{\Gamma^{k,\ell}}$. The eigenvalue–eigenfunction pairs $(\lambda,\,\psi)$ for $P_\Sigma Q_T$ are produced as follows.

Proposition 3: Suppose that $J_1,\dots,J_M$ are pairwise disjoint unit intervals with union $\Sigma = \cup_{k=1}^M J_k$. Let Λ denote the diagonal matrix with nth diagonal entry $\lambda_n(PQ_T)$ and let $\Gamma^{k,\ell}$ be the matrix with entries $\gamma_{nm}^{k,\ell} = \langle Q_TM_{m_k-m_\ell}\varphi_n,\,\varphi_m\rangle$, $k<\ell$. Then any eigenvector–eigenvalue pair ψ and λ for $P_\Sigma Q_T$ can be expressed as $\psi = \sum_{k=1}^M\sum_{n=0}^\infty \alpha_n^k M_{m_k}\varphi_n$ where the vectors $\boldsymbol{\alpha}_k = \{\alpha_n^k\}$ together form a discrete eigenvector for the block matrix eigenvalue problem

$$\lambda\begin{pmatrix}\boldsymbol{\alpha}_1\\ \boldsymbol{\alpha}_2\\ \vdots\\ \boldsymbol{\alpha}_M\end{pmatrix} = \begin{pmatrix}\Lambda & \bar\Gamma^{12} & \cdots & \bar\Gamma^{1M}\\ (\Gamma^{12})^T & \Lambda & \bar\Gamma^{23} & \cdots\\ \vdots & \vdots & \ddots & \vdots\\ (\Gamma^{1M})^T & \cdots & \cdots & \Lambda\end{pmatrix}\begin{pmatrix}\boldsymbol{\alpha}_1\\ \boldsymbol{\alpha}_2\\ \vdots\\ \boldsymbol{\alpha}_M\end{pmatrix}.$$

In order to turn the method into a means to compute eigenvalues and eigenfunctions of $P_\Sigma Q_T$ numerically, one needs to estimate the coefficients

$$\Gamma_{nm}^{k,\ell} = \int_{-T}^{T} \mathrm{e}^{2\pi\mathrm{i}(m_k-m_\ell)t}\varphi_n(t)\,\varphi_m(t)\,\mathrm{d}t$$

and to justify truncating the matrices Λ and $\Gamma^{k,\ell}$. The matrix truncations are justified by Theorem 1.

The corresponding Γ-matrix entries can be expressed as

$$\langle Q_TM_{m_I}\varphi_n,\, M_{m_J}\varphi_m\rangle = \sum_k\sum_\ell \varphi_n(k)\varphi_m(\ell)A(T;I,J)_{k\ell};$$

$$A(T;I,J)_{k\ell} = \int_{-T}^{T} \mathrm{e}^{2\pi\mathrm{i}(m_I-m_J)t}\,\mathrm{sinc}\,(t-k)\mathrm{sinc}\,(t-\ell)\,\mathrm{d}t\,.$$

The inner products are computed using the following, see [9].

Lemma 4: As a bilinear form acting on the pair of sequences $\{\varphi_n(k)\}$, $\{\varphi_m(\ell)\}$, the matrix $A(T;I,J)_{k\ell}$ coincides with $\mathrm{i}^{n+m}\sqrt{\lambda_m\lambda_n}\,\mathrm{sinc}\,(2T(m_J-m_I)+k-\ell)$.

An eigenfunction ψ of $P_\Sigma Q_T$ will be called a *time- and multiband-limiting eigenfunction* (TMBLE). If ψ is a TMBLE with eigenvalue $\lambda > 1/2$ then ψ should be, at least nearly, in the span of those eigenfunctions φ_n^I, where $\Sigma = \cup I$, corresponding to the eigenvalues of P_IQ_T larger than $1/2$, hence, of eigenfunctions φ_n^I corresponding to $n \le 2T$. In this case, φ_n^I can be approximated accurately on $[-T,T]$ by sinc interpolating its samples $\varphi_n^I(k)$ where $|k| \le M(T)$ above.

B. Numerical estimation of TMBLEs

Accurate numerical estimation of the TMBLEs is obtained via estimation of the entries of suitable truncations of the Γ matrices and eigenvectors of the corresponding truncation of the eigenproblem in Proposition 3. Details are given in [9]. Figure 1 illustrates the case with three frequency intervals.The corresponding eigenfunctions are plotted in Fig. 2.

IV. Time- and bandpass-limited signals

Given $0 < c' < c$ denote by $\mathrm{PW}_{c',c}^\pi$ the orthogonal complement of $\mathrm{PW}_{c'/\pi}$ inside $\mathrm{PW}_{c/\pi}$, that is, the closed subspace of $L^2(\mathbb{R})$ of functions whose Fourier transforms $\widehat{f}(\xi)$ are supported in $c'/\pi \le |\xi| \le c/\pi$, and by $P_{c'c}^\pi$ the orthogonal projection onto $\mathrm{PW}_{c'c}^\pi$. The eigenfunctions of the operator

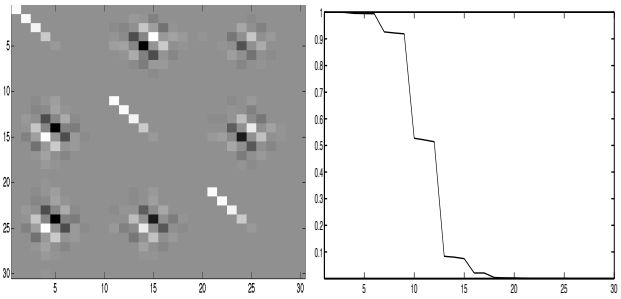


Fig. 1. *Matrix in Proposition 3 for* $T = 2$, $I = [-1/2, 1/2]$, $J = [2, 3]$, *and* $K = [5, 6]$. Intensity plot of the real part of the matrix in Proposition 3. Each $\Gamma^{\mu\nu}$ term is truncated to size 10×10. On the right is a plot of the moduli of the eigenvalues of the same matrix.

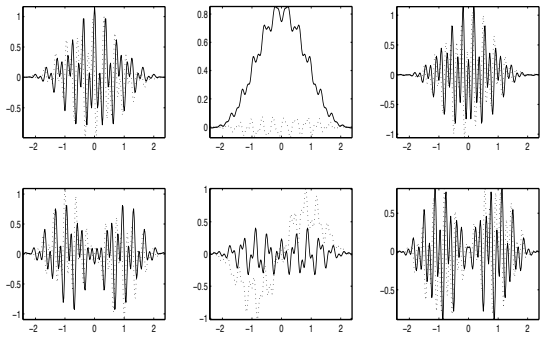


Fig. 2. *TMBLEs for* $T = 2$, $I = [-1/2, 1/2]$, $J = [2, 3]$ *and* $K = [5, 6]$. Plotted are the TMBLMS corresponding to $n = 0, 1, 2, 3, 4, 5$ respectively for three frequency bands. Real parts solid, imaginary parts dashed.

$P^{\pi}_{c'c}Q$ corresponding to time truncation of a function in $L^2(\mathbb{R})$ to a finite interval—$[-1, 1]$ in this work—followed by frequency limiting to frequencies $|\omega| \in [c', c]/\pi$ will be called *bandpass prolates* here. Numerical approximation of the most time concentrated bandpass limited signals (called *bandpass prolates* here) was studied recently by SenGupta et al., [10] by expressing the kernel of the bandpass limiting operator in terms of Legendre polynomials, then identifying the bandpass prolates through their Legendre coefficients. Alternatively, Proposition 5 proved in [11], produces the coefficients of the bandpass prolates, expressed as superpositions of full-band prolates, from *partial inner products* of full-band prolates. As explained below, these partial inner products can be computed directly from pointwise values of φ^c_n and $\varphi^{c'}_n$ where, as before, φ^c_n is the nth eigenfunction of $P_{c/\pi}Q$.

Denote by $R = R(c', c)$ the matrix with entries $R_{jk} = \frac{\mathrm{i}^{k-j}}{\sqrt{\lambda_j \lambda_k}} \int_{-c'/c}^{c'/c} \varphi^c_k(\xi)\, \varphi^c_j(\xi)\, \mathrm{d}\xi$. The matrix R is real symmetric, a consequence of the parity properties of the φ^c_n. Let $\Lambda = \Lambda(c)$ be the diagonal matrix with nth diagonal entry $\lambda_n(c)$.

Proposition 5: If $\psi = \sum \alpha_n \varphi^c_n \in \mathrm{PW}_{c/\pi}$ then

$$P^{\pi}_{c'c}\, Q\, \psi = \sum_k \alpha_k \lambda_k \left(\varphi^c_k - \sum_j R_{jk}\, \varphi^c_j \right).$$

In particular, if $\psi = \sum \alpha_n \varphi^c_n$ is an eigenfunction of $P^{\pi}_{c'c}Q$ with eigenvalue λ then, with $\boldsymbol{\alpha} = \{\alpha_n\}_{n=0}^{\infty}$,

$$\lambda \alpha_n = \lambda_n \alpha_n - \sum_k \lambda_k\, \alpha_k\, R_{nk} \quad i.e. \quad \lambda \boldsymbol{\alpha} = (I - R)\Lambda \boldsymbol{\alpha}\,.$$

The discrete eigenvectors $\boldsymbol{\alpha}$ of the matrix $(I - R)\Lambda$ thus give rise to eigenfunctions of $P^{\pi}_{c'c}Q$ and the eigenvalue λ measures the concentration of ψ in $[-1, 1]$ just as in the case of standard prolates. The proof uses the identities (1) and (2).

The partial inner products can be calculated by virtue of the prolate differential equation and integration by parts. If $n \neq m$ then, with χ_n as in (**??**) and $-1 \leq a \leq b \leq 1$,

$$\big(\chi_n - \chi_m\big) \int_a^b \varphi_n(t)\varphi_m(t)\mathrm{d}t = \Big[(t^2 - 1)(\varphi'_n \varphi_m - \varphi'_m \varphi_n)(t) \Big|_a^b .$$

Approximate bandpass prolates are obtained from finite size truncations of the eigenproblem in Proposition 5, see [11]. Khare [12] also considered the problem of numerical evaluation of bandpass prolates, focusing instead on the role of the interpolating function (sinc multiplied by a suitably dilated cosine) and establishing that the bandpass prolate samples form a discrete eigenvector of the matrix of partial integrals on $[-1, 1]$ of shifts of the interpolating kernel, cf. also Hogan et al., [7]. Khare did not investigate dependence on c'/c.

V. Phase synchrony and an application to EEG

We discuss briefly an application of bandpass prolates to study phase synchrony—nearly constant average instantaneous phase difference—particularly of EEG signals. It is believed that communication between different regions of neural cortex in attention focusing tasks is manifest in phase synchrony of neural firing patterns, e.g., [13], [14], particularly in the *gamma band*, e.g., [15]. Measuring band specific synchrony between EEG channels requires (i) a means to associate instantaneous phase to a given frequency band and (ii) a method to measure temporal phase locking between a pair of signals in a given band by averaging instantaneous phase difference for enough oscillations that average phase difference makes sense—say three to five—but not so many that synchronous epochs are indistinguishable from asynchronous ones.

Proposed methods include filtered analytic signals and convolutions with modulated Gaussians [16], [17], and empirical mode decomposition methods [18], [19] among others. In each case, the instantaneous phase is defined as the log of the complex valued signal divided by its modulus. The instantaneous phase difference of two such signals is the log of the product of the first unimodular signal and the conjugate of the second. To quantify phase locking of two signals one takes a time average of the instantaneous phase difference over a period that amounts to several oscillations.

In the case of analytic extensions of signals filtered over a short duration, aliasing is a concern. In the case of convolution with modulated Gaussians, insufficiently many degrees of freedom are being employed. The empirical mode decomposition provides a data- and algorithmic-driven definition of phase. However, it can be impossible to physical from algorithmic factors underlying the measured phase.

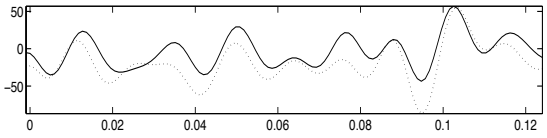


Fig. 3. *EEG channel data* $1/8$ second record of two concurrent EEG channel measurements, digitally sampled at 1024 samples per second.

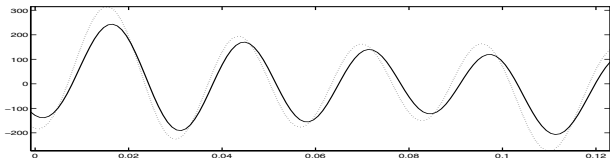


Fig. 4. *Approximate γ projections.* Projections of channel measurements onto the span of the six top eigenvectors of time-limiting to $1/8$ second and bandpass limiting to 24–40 Hz.

We consider here a new phase-locking metric computed through the following steps. *Step 1*: define the duration and frequency band for which synchrony is to be measured. *Step 2*: define the projection onto the span of the bandpass prolates whose eigenvalues are close to one or, at least, not much smaller than one half. *Step 3*: compute the analytic signal for this projection, and divide by its amplitude to get its unimodular factor. *Step 4*: For a pair of such signals, multiply the unimodular part of one by the conjugate of that of the other, integrate over the given duration, and compute the modulus. This is the *phase locking value* (PLV).

We implemented this algorithm as follows to produce Fig. 5. To analyze the gamma band of EEG signals, we chose the frequency range from 24 to 40 Hz. In order to compute the PLV over 3 to 5 oscillations of signals in this range, we took the duration of interest to be $1/8$ second. The time bandwidth product in this case is $2(40-24)/8 = 4$. The corresponding time- and bandpass-limiting operator has six eigenvalues "not much smaller than $1/2$." We successively chose $1/8$ second blocks of the EEG channels and computed the projections onto the span of the first six eigenfunctions. We then computed the analytic signal using the `matlab` builtin `hilbert`. A PLV was computed for each successive $1/8$-second segment of the two EEG channels.

Fig. 5 shows PLVs of projections of $1/8$-seconds of the two EEG channels onto the space generated by the six eigenfunctions of time limiting to $1/8$-second duration and bandpass limiting to 24–40 Hz most concentrated to the given duration. The PLVs were computed for $1/8$-second duration. In the data presented, a visual stimulus was shown to the subject after a half second. An initial interval of synchrony then presumably reflects response of the visual cortex. The subsequent interval of synchrony after "$t = 0$" presumably then corresponds to the subject maintaining a mental representation of the stimulus.

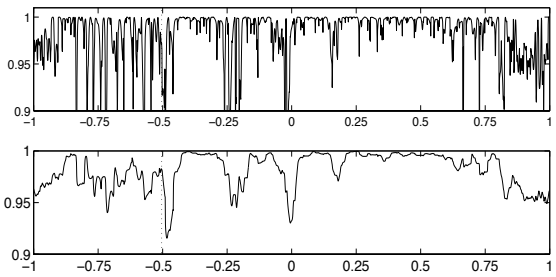


Fig. 5. *Phase locking.* Top shows PLVs of projections in Fig. 4 computed over the full two second record. Bottom shows PLVs time averaged over 20 consecutive time shifts.

REFERENCES

[1] D. Slepian and H. Pollak, "Prolate spheroidal wave functions, Fourier analysis and uncertainty. I," *Bell System Tech. J.*, vol. 40, pp. 43–63, 1961.

[2] H. Landau and H. Pollak, "Prolate spheroidal wave functions, Fourier analysis and uncertainty. II," *Bell System Tech. J.*, vol. 40, pp. 65–84, 1961.

[3] J. Hogan and J. Lakey, *Duration and Bandwidth Limiting*. Birkhäuser/Springer, New York, 2012.

[4] H. Landau and H. Widom, "Eigenvalue distribution of time and frequency limiting," *J. Math. Anal. Appl.*, vol. 77, pp. 469–481, 1980.

[5] G. Walter and X. Shen, "Sampling with prolate spheroidal wave functions," *Sampl. Theory Signal Image Process.*, vol. 2, pp. 25–52, 2003.

[6] K. Khare and N. George, "Sampling theory approach to prolate spheroidal wavefunctions," *J. Phys. A*, vol. 36, pp. 10 011–10 021, 2003.

[7] J. Hogan, S. Izu, and J. Lakey, "Sampling approximations for time- and bandlimiting," *Sampl. Theory Signal Image Process.*, pp. 91–117, 2010.

[8] J. A. Morrison, "On the eigenfunctions corresponding to the bandpass kernel, in the case of degeneracy," *Quart. Appl. Math.*, vol. 21, pp. 13–19, 1963.

[9] J. D. Lakey and J. A. Hogan, "On the numerical computation of certain eigenfunctions of time and multiband limiting." *Numer. Funct. Anal. Optim.*, vol. 33, no. 7-9, pp. 1095–1111, 2012.

[10] I. SenGupta, B. Sun, W. Jiang, G. Chen, and M. Mariani, "Concentration problems for bandpass filters in communication theory over disjoint frequency intervals and numerical solutions," *J. Fourier Anal. Appl.*, vol. 18, no. 1, pp. 182–210, 2012.

[11] J. Hogan and J. Lakey, "Letter to the editor: On the numerical evaluation of bandpass prolates," *J. Fourier Anal. Appl.*, pp. 1–8, 2013.

[12] K. Khare, "Bandpass sampling and bandpass analogues of prolate spheroidal functions," *Signal Process.*, vol. 86, no. 7, pp. 1550–1558, 2006.

[13] D. Hansel, G. Mato, and C. Meunier, "Synchrony in excitatory neural networks." *Neural Computation*, vol. 7, no. 2, pp. 307–337, 1995.

[14] F. Varela, J. P. Lachaux, E. Rodriguez, and J. Martinerie, "The brainweb: phase synchronization and large-scale integration." *Nature reviews. Neuroscience*, vol. 2, pp. 229–239, 2001.

[15] C. Tallon-baudry, O. Bertrand, F. Peronnet, and J. Pernier, "Induced gamma-band activity during the delay of a visual short-term memory task in humans," *J. Neurosci*, vol. 18, pp. 4144–4154, 1998.

[16] J.-P. Lachaux, E. Rodriguez, J. Martinerie, and F. J. Varela, "Measuring phase synchrony in brain signals," *Human Brain Mapping*, vol. 8, pp. 194–208, 1999.

[17] M. L. V. Quyen, J. Foucher, J. Lachaux, E. Rodriguez, A. Lutz, J. Martinerie, and Varela, "Comparison of Hilbert transform and wavelet methods for the analysis of neuronal synchrony," *J Neurosci Methods*, no. 111, pp. 83–98, 2001.

[18] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, and H. H. Liu, "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, vol. 454, no. 1971, pp. 903–995, 1998.

[19] C. Sweeney-Reed and S. Nasuto, "A novel approach to the detection of synchronisation in EEG based on empirical mode decomposition," *Journal of Computational Neuroscience*, vol. 23, pp. 79–111, 2007.

Recovery of Bandlimited Signal Based on Nonuniform Derivative Sampling

Dominik Rzepka[1], Marek Miśkowicz[1], Anna Gryboś[2], Dariusz Kościelnik[1]
[1]Department of Electronics, [2]Faculty of Applied Mathematics,
AGH University of Science and Technology, al. Mickiewicza 30, 30-059 Cracow, Poland
Email: {drzepka, miskow, grybos, koscieln}@agh.edu.pl

***Abstract*—The paper focuses on the perfect recovery of bandlimited signals from nonuniform samples of the signal and its derivatives. The main motivation to address signal recovery using nonuniform derivative sampling is a reduction of mean sampling frequency under Nyquist rate which is a critical issue in event-based signal processing chains with wireless link. In particular, we introduce a set of reconstructing functions for nonuniform derivative sampling as an extension of relevant set of reconstructing functions derived by Linden and Abramson for uniform derivative sampling. An example of signal recovery using the first derivative is finally reported.**

I. Introduction

Modern sensor and control systems require advanced irregular sampling algorithms to represent continuous-time signals by a sequence of discrete-time samples taken at right time. A special class of irregular observations is constituted by the event-based sampling schemes. This class is characterized by the functional relationship between sampling instants and temporal signal behavior so the samples are captured when it is required. Event-based sampling, known since the 50s [15] experiences increasing interest because processing of events defined as a *significant change* of a selected signal parameter is the objective of various signal processing or monitoring and control systems [1], [9], [15], [19], [20], [26], [27]. All the event-based sampling schemes produce samples irregularly in time according to temporal signal variations.

Several signal-dependent sampling criteria have been proposed and investigated in recent years [15], [16], [25], [17], [11], [27] including the algorithms based on controlling the linear intersampling error [1], [5], [6], [11], [15], [21], integral error [16], [18], and the energy of intersampling error [17]. The most natural signal-dependent sampling scheme is based on the send-on-delta principle and consists in keeping the linear intersampling error bounded [1], [5], [6], [11], [15], [21].

The send-on-delta scheme is known in the literature also as Lebesgue sampling in the context of control systems theory [1], [6] and derivations from Lebesgue integral [6], or level-crossing sampling especially in the context of signal conversion and processing [5], [12], [20].

In [25], the send-on-delta/level-crossing sampling with prediction as an enhanced version of the pure send-on-delta/level-crossing principle has been introduced. The send-on-delta/level-crossing scheme with prediction is a sampling algorithm that employs the prediction to approximate the sampled signal between sampling instants.

The prediction is based on a belief that the sampled signal will vary according to the first-order (linear) or second-order (quadratic) approximation by the truncated Taylor series expanded at the instant of the most recent sample. The next sample is captured when the predicted signal value deviates from the real signal value by an interval of confidence [25], [23]. In particular, in the send-on-delta/level-crossing sampling with prediction, either signal or its time-derivatives are sampled and transmitted irregularly in time via communication channel for possible processing and/or reconstruction.

The present paper deals with involving signal derivatives to nonuniform sampling. More specifically, we examine the problem of recovery of original signal based on non-uniform discrete-time representation of the signal and its derivatives.

The present paper deals with involving signal derivatives to nonuniform sampling. More specifically, we examine the problem of recovery of original signal based on non-uniform discrete-time representation of the signal and its derivatives. The primary goal of adopting derivative sampling to irregular discrete-time signal representation is a desire to reduce the sampling rate below the Nyquist rate. Decreasing the mean rate of data records is an issue of primary importance in signal processing systems with wireless links since wireless communication is a major source of energy consumption. In the paper, we provide a procedure for perfect recovery of bandlimited signals for non-uniform derivative sampling. The original contribution of the paper is a formulation of a set of reconstructing functions for non-uniform derivative sampling as the extension of relevant set of reconstructing functions derived by Linden and Abramson for uniform derivative sampling in their classical paper on generalized sampling theorem [14]. Finally, we illustrate the reconstruction procedure on the example of signal recovery using the first derivative.

II. Problem formulation

Let us assume that a signal $x(t)$ of finite energy is bandlimited, i.e. $X(\omega) = 0$ for $\omega \notin (-\Omega, \Omega)$. Suppose that the signal $x(t)$ and its first $(m-1)$ time-derivatives are sampled irregularly in time which results in producing a set of samples $\{x^{(0)}(t_n), x^{(1)}(t_n), \ldots, x^{(m-1)}(t_n)\}$ taken at the instants t_n,

This work was supported by the National Center of Science under Grant DEC-2012/05/E/ST7/01143.

$n \in \mathbb{Z}$. The aim of the present study is to recover the original signal $x(t)$ using the given samples.

A. Recovery of signal from nonuniform samples

Recovering bandlimited signal from its samples taken at nonuniform time instants is possible on the basis of theory of frames and non-harmonic Fourier series. Both concepts were introduced by Duffin and Schaeffer in [2]. The frame $\{g_n\}$ generalize the idea of a basis in a Hilbert space H in the sense that it allows representing an arbitrary element $x \in H$ as long as there exist the *frame bounds* $A, B > 0$ such that the following *frame condition* is fulfilled

$$0 < A\|x\|^2 < \sum_{n=-\infty}^{+\infty} |\langle g_n, x\rangle|^2 < B\|x\|^2 < \infty \tag{1}$$

If the set of functions $\{g_n(t)\}$ is a frame or a basis, there exists a set of coefficients c_n which allows to represent a function $x(t)$ as

$$x(t) = \sum_{n=-\infty}^{+\infty} c_n g_n(t) \tag{2}$$

The Shannon uniform sampling theory uses the basis of functions $g_n(t) = \mathrm{sinc}(\Omega(t - nT))$ and samples $x(nT)$ as coefficients c_n where $\mathrm{sinc}(t) := \sin(t)/t$. Frame theory allows obtaining coefficients c_n for frame composed of functions $g_n(t) = \mathrm{sinc}(\Omega(t - t_n))$, where t_n are sampling instants. The set t_n is not arbitrary and must obey certain conditions: $|t_n - n| < 1/4$ [10] and for finite subset of t_n it is allowed that $|t_n - n| = \mathcal{O}(n^{-\gamma}), n \to \infty, \gamma > 1$ [3]. To obtain values of coefficients c_n, we insert the known time instants t_n as t (we mark them by t_l to avoid confusion) in (2), getting

$$x(t_l) = \sum_{n=-\infty}^{+\infty} c_n g_n(t_l) \tag{3}$$

where $g_n(t_l) = \mathrm{sinc}(\Omega(t_l - t_n))$. This may be written in matrix form

$$\mathbf{x} = \mathbf{G}\mathbf{c} \tag{4}$$

where $\mathbf{x^T} = [\ldots, x(t_{n-1}), x(t_n), x(t_{n+1}), \ldots]$, and $[\mathbf{G}]_{i,j} = g_i(t_j)$. The matrix $\mathbf{G}$ is infinite dimensional, so to apply practical recovery algorithm a truncated matrix is used [24]. The values of c_n can be then calculated on the basis of computation of the pseudo-inverse matrix

$$\mathbf{c} = (\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T\mathbf{x} \tag{5}$$

The recovery method stated above has been used for recovery of amplitude information from time-encoded signals and proposed for applications in neurocomputing and time-mode signal processing systems [13].

B. Derivative sampling

The problem of reconstructing bandlimited signal from the samples of the signal and its time derivative(s) has been studied in the context of uniform sampling for decades [8]. The significant benefit of including samples of time derivatives to procedure of signal recovery is a reduction of the sampling frequency. In particular, a possibility to recover the signal based on knowledge of the samples of the signal and its first time-derivative was mentioned by Shannon in one of his milestone papers [22]. This idea was further developed by Jagerman and Vogel [4], [7] for first derivative. The derivative sampling theorem was generalized for arbitrary number of derivatives by Linden [14]. Sampling of m derivatives at once (in our notation signal itself is zero-order derivative) allows for m-fold decrease of sampling frequency, so the sampling period becomes $T_m = m\pi/\Omega$. The reconstruction formula [14] is given by

$$\begin{aligned} x(t) = \sum_{n=-\infty}^{+\infty} & x^{(0)}(nT_m)g_0(t - nT_m) + \\ & + x^{(1)}(nT_m)g_1(t - nT_m) + \\ & + \ldots + \\ & + x^{(m-1)}(nT_m)g_{m-1}(t - nT_m) \end{aligned} \tag{6}$$

with reconstruction functions

$$g_k(t) = \frac{t^k}{k!}\,\mathrm{sinc}^m\left(\frac{\Omega}{m}t\right) \tag{7}$$

for $k \in (0, \ldots, m-1)$. Note that $g_k(t)$ is a function corresponding not to a single sample $x(kT)$ but to the infinite set of samples $x^{(k)}(nT)$.

C. Nonuniform derivative sampling

Since reconstruction formulas (2) and (6) are both based on convergence of Fourier series, then we can write also the nonuniform analogue of (6)

$$\begin{aligned} x(t) = \sum_{n=-\infty}^{+\infty} & c_{0,n}g_0(t - t_n) + \\ & + c_{1,n}g_1(t - t_n) + \\ & + \ldots + \\ & + c_{m-1,n}g_{m-1}(t - t_n) \end{aligned} \tag{8}$$

assuming that all derivatives are sampled nonuniformly, at the same instants t_n. Using a matrix notation (4) with $[\mathbf{G}^k]_{i,j} = g_k(t_i - t_j)$ for $k \in (0, \ldots, m-1)$ we obtain

$$\mathbf{x} = \mathbf{G}_0\mathbf{c}_0 + \mathbf{G}_1\mathbf{c}_1 + \ldots + \mathbf{G}_{m-1}\mathbf{c}_{m-1} \tag{9}$$

The vector $\mathbf{x}$ contains samples of $x(t)$ taken with frequency m-times lower than Nyquist frequency. For this reason the formula (9) is not sufficient to obtain coefficients $\{\mathbf{c}_0, \mathbf{c}_1, \ldots, \mathbf{c}_{m-1}\}$. To make use of derivatives samples, we differentiate r-times both sides of (8), which yields

$$x^{(r)}(t) = \sum_{n=-\infty}^{+\infty} c_{0,n} g_0^{(r)}(t-t_n) + \\ + c_{1,n} g_1^{(r)}(t-t_n) + \\ + \ldots + \\ + c_{m-1,n} g_{m-1}^{(r)}(t-t_n) \tag{10}$$

$$g_k^{(r)}(t) = \frac{d^r}{dt^r}\left(\frac{t^k}{k!}\operatorname{sinc}^m\left(\frac{\Omega}{m}t\right)\right) \tag{11}$$

In particular, for $m = 0$, the set of reconstructing functions given by (11) is reduced to (7) which represents the classical nonuniform signal recovery from based on (2) without the use of derivative sampling. The equation (10) can be arranged for each $r \in (0, \ldots, m-1)$ into system of equations, which can be also written in block-matrix form

$$\begin{bmatrix} \mathbf{x}^{(0)} \\ \mathbf{x}^{(1)} \\ \vdots \\ \mathbf{x}^{(m-1)} \end{bmatrix} = \begin{bmatrix} \mathbf{G}_0^{(0)} & \mathbf{G}_1^{(0)} & \ldots & \mathbf{G}_{m-1}^{(0)} \\ \mathbf{G}_0^{(1)} & \mathbf{G}_1^{(1)} & \ldots & \mathbf{G}_{m-1}^{(1)} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{G}_0^{(m-1)} & \mathbf{G}_1^{(m-1)} & \ldots & \mathbf{G}_{m-1}^{(m-1)} \end{bmatrix} \cdot \begin{bmatrix} \mathbf{c}_0 \\ \mathbf{c}_1 \\ \vdots \\ \mathbf{c}_{m-1} \end{bmatrix} \tag{12}$$

Solving this system gives coefficients $\mathbf{c}_0, \mathbf{c}_1, \ldots, \mathbf{c}_{m-1}$ for reconstruction with (8). Therefore computational complexity of the reconstruction procedure corresponds to classic matrix inversion complexity $\mathcal{O}(n^3)$ and it is dependent on the number of samples used. Summing up, the proposed procedure of signal recovery is based on computation of the coefficients $\mathbf{c}_0, \mathbf{c}_1, \ldots, \mathbf{c}_{m-1}$ to the set of reconstructing functions defined by (7). The number of reconstructing functions depends on the number of derivatives used for signal recovery.

III. Simulations

As an example illustrating the procedure of signal recovery based on nonuniform derivative sampling, we present a reconstruction from samples of the signal and its first derivative, i.e. for $m = 2$. In this case we have the following reconstructing functions on the basis of (11)

$$g_0^{(0)}(t) = \frac{4\sin^2(\Omega t/2)}{\Omega^2 t^2}$$
$$g_1^{(0)}(t) = \frac{4\sin^2(\Omega t/2)}{\Omega^2 t}$$
$$g_0^{(1)}(t) = \frac{2\left(-2 + 2\cos(\Omega t) + \Omega t \sin(\Omega t)\right)}{\Omega^2 t^3}$$
$$g_1^{(1)}(t) = \frac{2\left(-1 + \cos(\Omega t) + \Omega t \sin(\Omega t)\right)}{\Omega^2 t^2}$$

The coefficients $\mathbf{c}_0, \mathbf{c}_1$, are computed on the basis of the following reconstruction equation:

$$\begin{bmatrix} \mathbf{x}^{(0)} \\ \mathbf{x}^{(1)} \end{bmatrix} = \begin{bmatrix} \mathbf{G}_0^{(0)} & \mathbf{G}_1^{(0)} \\ \mathbf{G}_0^{(1)} & \mathbf{G}_1^{(1)} \end{bmatrix} \cdot \begin{bmatrix} \mathbf{c}_0 \\ \mathbf{c}_1 \end{bmatrix}$$

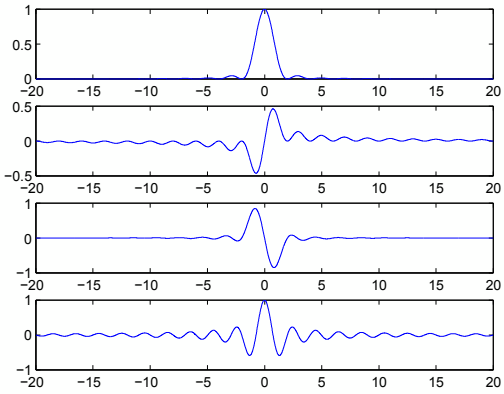


Fig. 1. Reconstruction functions $g_0^{(0)}$, $g_1^{(0)}$, $g_0^{(1)}$, $g_1^{(1)}$, when signal is recovered from the samples of $x(t)$ and its derivative $x'(t)$

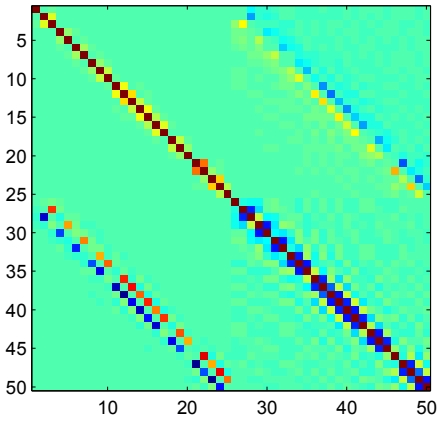


Fig. 2. Reconstruction matrix, composed of matrices $\mathbf{G}_0^{(0)}$, $\mathbf{G}_1^{(0)}$, $\mathbf{G}_0^{(1)}$, $\mathbf{G}_1^{(1)}$ corresponding to respective reconstructon functions $g_0^{(0)}$, $g_1^{(0)}$, $g_0^{(1)}$, $g_1^{(1)}$.

The signal $x(t)$ used for exemplified recovery was generated using Shannon-Whittaker reconstruction formula

$$x(t) = \sum_{n=1}^{40} x_n \frac{\sin(\pi(t-n))}{\pi(t-n)} \tag{13}$$

where $x_1, \ldots, x_{40}$ were selected as independent realizations of random variable with normal distribution. Thus, the signal $x(t)$ is bandlimited to $\Omega = \pi$. The signal derivative $x'(t)$ was computed using differentiated $\operatorname{sinc}(\cdot)$ function. The signal $x(t)$ has been sampled using send-on-delta sampling scheme with linear prediction [23], [25], resulting in 25 samples of the signal and 25 samples of its derivative. The reconstruction block-matrix $\mathbf{G}$ is depicted in the Fig. 2, where red corresponds to higher, and blue to the lower values. The number of each type of samples required for reconstruction is 20 so signal is slightly oversampled. As stated in Introduction, in the send-on-delta scheme with linear prediction, the sampling operation is triggered when the predicted signal based on linear prediction

deviates from the real signal value by an interval of confidence [23]. The original and the reconstructed signal are depicted in the Fig. 3. The absolute linear error of reconstruction is presented in the Fig. 4.

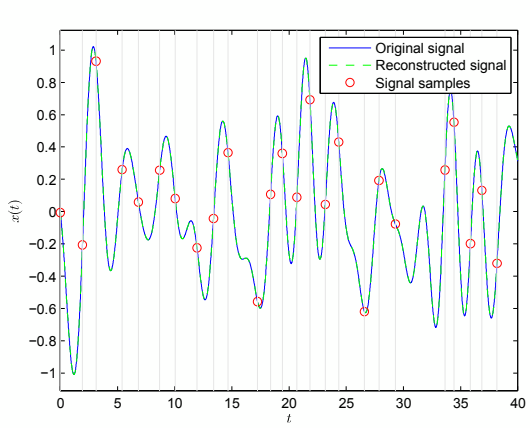


Fig. 3. Original bandlimited signal $x(t)$ and its reconstruction $\tilde{x}(t)$ from the nonuniform samples of $x(t)$ and its first derivative $x'(t)$.

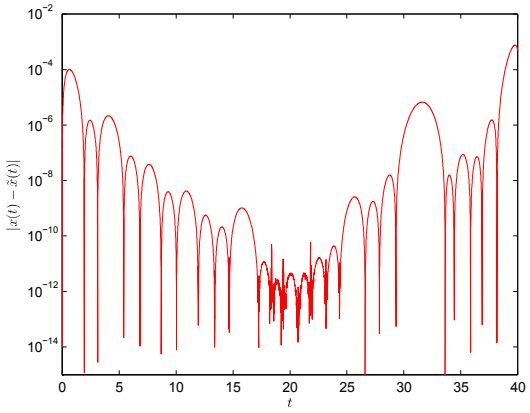


Fig. 4. Error of the reconstruction of the signal from the nonuniform samples of $x(t)$ and its derivative $x'(t)$. The lowest values occure at sampling instants t_n (Compare Fig. 3).

IV. Conclusions

The paper focuses on the perfect reconstruction of bandlimited signal from the nonuniformly spaced samples of the signal and its derivatives. The principal benefit of signal recovery using nonuniform derivative sampling is a reduction of mean sampling frequency under Nyquist rate which is a critical issue in signal processing chains with wireless link based on event-based sampling. The computational complexity of the proposed recovery procedure is connected with the matrix inversion needed to calculate the coefficients $\mathbf{c}_0, \mathbf{c}_1, \ldots, \mathbf{c}_{m-1}$.

References

[1] K.J. Astrm, B. Bernhardsson, *"Comparison of periodic and event based sampling for first-order stochastic systems"*, Proceedings of IFAC World Congress, pp. 301-306, 1999.

[2] R. J. Duffin, A. C. Schaeffer *"Class of Nonharmonic Fourier Series"* , Transactions of the American Mathematical Society, Vol. 72, No. 2 (Mar., 1952), pp. 341-366

[3] J.R. Higgins, *"Completeness and Basic Properties of Sets of Special Functions"* , Cambridge University Press, 1977, pp. 81-83

[4] L. Fogel, *"A note on the sampling theorem"* , IRE Transactions on Information Theory, vol.1, no.1, pp.47-48, March 1955

[5] K.M. Guan, S.S Kozat, and A.C. Singer, *"Adaptive reference levels in a level-crossing analog-to-digital converter"*, EURASIP Journal on Advances in Signal Processing, vol. 2008, Article ID 513706, 2008.

[6] E. Gluskin, *"Nonlinear sampling and Lebesgue's integral sums"*, Proceedings of IEEE Convention of Electrical and Electronics Engineers in Israel IEEEI 2010, pp. 736-740, 2010.

[7] D. Jagerman, L. Fogel, *"Some general aspects of the sampling theorem"*, IRE Transactions on Information Theory, vol.2, no.4, pp.139-146, December 1956

[8] A.J. Jerri, *"The Shannon sampling theorem-Its various extensions and applications: A tutorial review"*, Proceedings of the IEEE , vol.65, no.11, pp. 1565- 1596, Nov. 1977

[9] W. Heemels, M. Donkers, A Teel, *"Periodic event-triggered control for linear systems"*, IEEE Transactions on Automatic Control, vol. 58, no. 4, pp. 847-861, April 2013.

[10] M. I. Kadec, *"The exact value of the Paley-Wiener constant"*, Dokl. Akad. Nauk SSSR 155, 1253-1254, 1964

[11] E. Kofman, J.H. Braslavsky, *"Level crossing sampling in feedback stabilization under data-rate constraints"*, Proceedings of IEEE Conference on Decision and Control CDC 2006, pp. 4423-4428, 2006.

[12] M. Kurchuk, Y. Tsividis, *"Signal-dependent variable-resolution clockless A/D conversion with application to continuous-time digital signal processing"*, IEEE Transactions on Circuits and Systems-I: Regular Papers, vol. 57, pp. 982-991, 2010.

[13] A.A Lazar, L.T. Tth, *"Perfect recovery and sensitivity analysis of time encoded bandlimited signals"*, IEEE Transactions on Circuits and Systems-I: Regular Papers, vol. 51, no. 10, pp. 2060-73, 2004.

[14] D.A. Linden, N.M. Abramson, *"A generalization of the sampling theorem"*, Information and Control, Volume 3, Issue 1, pp. 26-31, March 1960

[15] M. Miśkowicz, *"Send-on-delta concept: an event-based data reporting strategy"*, Sensors, vol. 6, pp. 49-63, 2006.

[16] M. Miśkowicz, *"Asymptotic effectiveness of the event-based sampling according to the integral criterion"* , Sensors, vol. 7, pp. 16-37, 2007.

[17] M. Miśkowicz, *"Efficiency of event-based sampling according to error energy criterion"*, Sensors, vol. 10, pp. 2242-2261, 2010.

[18] V.H. Nguyen and Y.S. Suh, *"Networked estimation with an area-triggered transmission method"*, Sensors, vol. 8, pp. 897-909, 2008.

[19] J. Sánchez, A. Visioli, S. Dormido, *"A two-degree-of-freedom PI controller based on events"*, Journal of Process Control, vol. 21, no. 4, pp. 639-651, 2011.

[20] M. De la Sen, *"On Chebyshevs systems and non-uniform sampling related to Caputo fractional dynamic time-invariant systems"*, Discrete Dynamics in Nature and Society, vol. 2010, Article ID 846590, 2010.

[21] S. Senay, L.F. Chaparro, M. Sun, and R.J. Sclabassi, *"Adaptive level-crossing sampling and reconstruction"*, Proceedings of European Signal Processing Conference EUSIPCO 2010, pp. 1296-1300, August 2010.

[22] C. E. Shannon, *"Communication in the Presence of Noise"*, Proceedings of the IRE, Vol. 37, No. 1. (January 1949), pp. 10-21

[23] K. Staszek, S. Koryciak, M. Miśkowicz *"Performance of send-on-delta sampling schemes with prediction"*, Industrial Electronics (ISIE), 2011 IEEE International Symposium on , vol., no., pp.2037-2042, 27-30 June 2011

[24] T. Strohmer. *"Approximation of dual Gabor frames, window decay, and wireless communications"*, Appl. Comput. Harmon. Anal., 11(2):243–262, 2001.

[25] Y.S. Suh, *"Send-on-delta sensor data transmission with a linear predictor"*, Sensors, vol. 7, pp. 537-547, 2007.

[26] Y. Tsividis, *"Event-driven data acquisition and digital signal processing: a tutorial"* , IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 57, no. 8, pp. 577-582, 2010.

[27] V. Vasyutynskyy, K. Kabitzsch, *"Comparative study of energy-efficient sampling approaches for wireless control networks"* , IEEE Transactions on Industrial Informatics, vol. 6, no. 3, pp. 416-424, 2010.

[28] J. M. Whittaker , *"Interpolatory Function Theory"* , Cambridge University Press, vol. 33, 1935

Approximation by Shannon sampling operators in terms of an averaged modulus of smoothness

Andi Kivinukk
Dept. of Mathematics
Tallinn University
Narva mnt 25
Tallinn 10120, Estonia
Email: andik@tlu.ee

Gert Tamberg
Dept. of Mathematics
Tallinn University of Technology
Ehitajate tee 5
Tallinn 19086, Estonia
Email: gtamberg@staff.ttu.ee

***Abstract*—The aim of this paper is to study the approximation properties of generalized sampling operators in $L^p(\mathbb{R})$-space in terms of an averaged modulus of smoothness.**

I. Introduction

For the uniformly continuous and bounded functions $f \in C(\mathbb{R})$ the generalized sampling series are given by ($t \in \mathbb{R}$; $w > 0$)

$$(S_w f)(t) := \sum_{k=-\infty}^{\infty} f(\frac{k}{w}) s(wt - k), \tag{1}$$

where the condition for the operator $S_w : C(\mathbb{R}) \to C(\mathbb{R})$ to be well-defined is

$$\sum_{k=-\infty}^{\infty} |s(u-k)| < \infty \quad (u \in \mathbb{R}), \tag{2}$$

the absolute convergence being uniform on compact intervals of $\mathbb{R}$.

If the kernel function is

$$s(t) = \operatorname{sinc}(t) := \frac{\sin \pi t}{\pi t},$$

we get the classical (Whittaker-Kotel'nikov-)Shannon operator,

$$(S_w^{\operatorname{sinc}} f)(t) := \sum_{k=-\infty}^{\infty} f(\frac{k}{w}) \operatorname{sinc}(wt - k).$$

A systematic study of sampling operators (1) for arbitrary kernel functions s with (2) was initiated at RWTH Aachen by P. L. Butzer and his students since 1977 (see [1], [2], [3] and references cited there).

Since in practice signals are however often discontinuous, this paper is concerned with the convergence of $S_w f$ to f in the $L^p(\mathbb{R})$-norm for $1 \leqslant p < \infty$, the classical modulus of continuity being replaced by the averaged modulus of smoothness $\tau_k(f; 1/w)p$. For the classical (Whittaker-Kotel'nikov-Shannon) operator this approach was introduced by P. L. Butzer, C. Bardaro, R. Stens and G. Vinti (2006) in [4] (see also [5]) for $1 < p < \infty$. For time-limited kernels s this approach was applied for $1 \leqslant p < \infty$ in [6] and [7]. In this paper we use this approach for band-limited kernels for $1 \leqslant p < \infty$.

In this paper we study an even band-limited kernel s, defined by an even window function $\lambda \in C_{[-1,1]}$, $\lambda(0) = 1$, $\lambda(u) = 0$ ($|u| \geqslant 1$) by the equality

$$s(t) := s(\lambda; t) := \int_0^1 \lambda(u) \cos(\pi t u)\, du. \tag{3}$$

We first used the band-limited kernel in general form (3) in [8], see also [9], [10]. We studied the generalized sampling operators $S_W : C(\mathbb{R}) \to C(\mathbb{R})$ with the kernels in form (3) in [11]-[12]. We computed exact values of operator norms

$$\|S_w\| := \sup_{\|f\|_C \leqslant 1} \|S_w f\|_C = \sup_{u \in \mathbb{R}} \sum_{k=-\infty}^{\infty} |s(u-k)| \tag{4}$$

and estimated the order of approximation in terms of the classical modulus of smoothness. In this paper we give similar results for $L^p(\mathbb{R})$ norm in terms of the averaged modulus of smoothness. The main result of this paper, Theorem 2, was proved for $f \in C(\mathbb{R})$ in [11].

II. Preliminary results

A. Averaged modulus of smoothness

In this section we follow the approach of Butzer et al [4] of convergence problems of Shannon sampling series in a suitable subspace of $L^p(\mathbb{R})$.

Let $f \in M(\mathbb{R})$ be measurable and bounded on $\mathbb{R}$, and $\delta \geqslant 0$. The k-th averaged τ-modulus of smoothness for $1 \leqslant p \leqslant \infty$ is defined as ([4], Def. 1)

$$\tau_k(f; \delta)p := \|\omega_k(f; \cdot; \delta)\|_p, \tag{5}$$

where $\omega_k(f; t; \delta)$ is a local modulus of smoothness of order $k \in \mathbb{N}$ at $t \in \mathbb{R}$,

$$\omega_k(f; t; \delta) := \\ := \sup\{|\Delta_h^k f(x)|; x, x + kh \in [t - \frac{k\delta}{2}, t + \frac{k\delta}{2}]\},$$

where the classical finite forward difference is given by

$$\Delta_h^k f(x) = \sum_{\ell=0}^{k} (-1)^{k-\ell} \binom{k}{\ell} f(x + \ell)h). \tag{6}$$

The classical modulus of smoothness can be estimated via the τ-modulus (see [4], Proposition 4)

$$\omega_k(f;\delta)_p \leqslant \tau_k(f;\delta)_p \quad (1 \leqslant p < \infty).$$

B. The space Λ^p

Since the sampling series $S_w f$ of (1) of an arbitrary L^p-function f may be divergent, we have to restrict the matter to a suitable subspace. Further, since we want to use the τ-modulus as a measure for the approximation error, we have to ensure that it is finite for all functions under consideration. In [4] it was proved that we can define a suitable subspace as follows

Definition 1 ([4], Def. 10, [6], Def. 2.1):
(a) A sequence $\Sigma := (x_j)_{j\in\mathbb{Z}} \subset \mathbb{R}$ is called an admissible partition of $\mathbb{R}$ or an admissible sequence, if it satisfies

$$0 < \inf_{j\in\mathbb{Z}} \Delta_j \leqslant \sup_{j\in\mathbb{Z}} \Delta_j < \infty, \quad \Delta_j := x_j - x_{j-1}.$$

(b) Let $\Sigma := (x_j)_{j\in\mathbb{Z}} \subset \mathbb{R}$ be an admissible partition of $\mathbb{R}$. The discrete $\ell^p(\Sigma)$-seminorm of a sequence of function values f_Σ on Σ of a function $f : \mathbb{R} \to \mathbb{C}$ is defined for $1 \leqslant p < \infty$ by

$$\|f\|_{\ell^p(\Sigma)} := \left\{ \sum_{j\in\mathbb{Z}} |f(x_j)|^p \Delta_j \right\}^{1/p}.$$

(c) The space Λ^p for $1 \leqslant p < \infty$ is defined by

$$\Lambda^p := \{f \in M(\mathbb{R}); \|f\|_{\ell^p(\Sigma)} < \infty \text{ for each admissible sequence } \Sigma\}.$$

It can be shown (see [4], Proposition 18) that if $f \in \Lambda^p \cap R_{loc}(\mathbb{R})$ for $1 \leqslant p < \infty$ we have

$$\lim_{\delta\to 0} \tau_k(f;\delta)_p = 0, \tag{7}$$

where

$$R_{loc}(\mathbb{R}) := \{f : \mathbb{R} \to \mathbb{C}, \text{ is locally Riemann integrable on } \mathbb{R}\}.$$

The assumption $f \in R_{loc}(\mathbb{R})$ is related to the fact that the τ-modulus on $[a,b]$ tends to zero (with $\delta \to 0+$) if and only if when f is Riemann integrable on $[a,b]$ (see [13], Th. 1.2 and [4], Proposition 6.).

We have for $1 \leqslant p < \infty$ that $B^p_w \subsetneq W^r_p \subsetneq \Lambda^p \subsetneq L^p$, where B^p_w is the Bernstein class (e.g. [14], Def. 6.5) and

$$W^r_p := \{f \in L^p; f \in AC^r_{loc}, f^{(r)} \in L^p\}$$

is the classical Sobolev space.

In the following we consider the uniform partitions $\Sigma_w := (j/w)_{j\in\mathbb{Z}} \subset \mathbb{R}$ for $w > 0$ only. For these partitions we have ([6], Proposition 2.2)

$$\|f\|_{\ell^p(w)} \leqslant \|f\|_p + \frac{1}{w}\|f'\|_p, \quad f \in W^r_p. \tag{8}$$

Proposition 1 ([6], Th. 2.8): Let $(L_w)_{w>0}$ be a family of linear operators mapping Λ^p into L^p, $1 \leqslant p < \infty$, satisfying the properties

$$(i) \quad \|L_w f\|_p \leqslant K\|f\|_{\ell^p(w)}, \quad f \in \Lambda^p, \tag{9}$$

$$(ii) \quad \|L_w g - g\|_p \leqslant K_r \frac{1}{w^s}\|g^{(r)}\|_p, \quad g \in W^r_p, \tag{10}$$

for some fixed $r, s \in \mathbb{N}$, $(s \leqslant r)$ and a constant K_r depending only on r. Then for each $f \in \Lambda^p$ there holds the estimate

$$\|L_w f - f\|_p \leqslant c\,\tau_r(f; \frac{1}{W^{s/r}})_p, \quad W > 0, \tag{11}$$

the constant c depending only on r, K and K_r.

To use Proposition 1 for Shannon sampling operators we need the following proposition.

Proposition 2 (cf. [4], Proposition 25): For $1 \leqslant p \leqslant \infty$, for some $r \in \mathbb{N}$ and $s = 0, 1, \ldots, r$ there exists a constant $c_r > 0$ such that for each $f \in W^r_p$ and $w > 0$ one can find a function $g_w \in B^p_{\pi w}$ satisfying

$$\|f^{(s)} - g^{(s)}_w\|_p \leqslant c_r \frac{1}{w^{r-s}}\|f^{(r)}\|_p.$$

C. Sampling operators

The kernel for the sampling operators S_w in (1) is defined in the following way.

Definition 2 ([3], Def. 6.3): If $s : \mathbb{R} \to \mathbb{C}$ is a bounded function such that

$$\sum_{k=-\infty}^{\infty} |s(u-k)| < \infty \quad (u \in \mathbb{R}), \tag{12}$$

the absolute convergence being uniform on compact subsets of $\mathbb{R}$, and

$$\sum_{k=-\infty}^{\infty} s(u-k) = 1 \quad (u \in \mathbb{R}), \tag{13}$$

then s is said to be a kernel for sampling operators (1).

For $f \in \Lambda^p$ we have:

Proposition 3 ([6], Proposition 3.2): Let $s \in M(\mathbb{R}) \cap L^1(\mathbb{R})$ be a kernel. Then $\{S_w\}_{w>0}$ defines a family of bounded linear operators from Λ^p into L^p, $1 \leqslant p < \infty$ (and also from $C(\mathbb{R})$ into $C\mathbb{R}$ with the norm (4)), satisfying $(1/p + 1/q = 1)$

$$\|S_w f\|_p \leqslant \|S_w\|^{1/q}\|s\|_1^{1/p}\|f\|_{\ell^p(w)} \quad (w > 0). \tag{14}$$

If the kernel s is time-limited, i.e. there exists $T_0, T_1 \in \mathbb{R}$, $T_0 < T_1$ such that $s(t) = 0$ for $t \notin [T_0, T_1]$, then in case $f \in \Lambda^p \cap R_{loc}(\mathbb{R})$ for $1 \leqslant p < \infty$, we have (see [6], Th. 4.4)

$$\lim_{w\to\infty} \|S_w f - f\|_p = 0. \tag{15}$$

I this paper we prove analogous result for band-limited kernels.

D. Band-limited kernels

In the following we assume that our kernel (3) belongs to B^1_π. For the band-limited functions $s \in B^p_\pi \subset L^p(\mathbb{R})$ the operator norm $\|S_w\|$ is related to the norm $\|s\|_p$ by Nikolskii's inequality.

Proposition 4 (Nikolskii inequality; [14], Th. 6.8): Let $1 \leqslant p \leqslant \infty$. Then, for every $s \in B^p_\sigma$,

$$\|s\|_p \leqslant \sup_{u\in\mathbb{R}} \left\{ \sum_{k=-\infty}^{\infty} |s(u-k)|^p \right\}^{1/p} \leqslant (1+\sigma)\|s\|_p.$$

From the Nikolskii's inequality we see that our assumption $s \in L^1(\mathbb{R})$ is sufficient for (12) and thus s in (3) is indeed a kernel in the sense of Definition 2.

These types of kernels arise in conjunction with window functions widely used in applications (e.g. [15], [16], [17], [18]), in Signal Analysis in particular. Unfortunately bandlimited kernels do not have compact support. Many kernels can be defined by (3), e.g.

1) $\lambda(u) = 1$ defines the sinc function;

2) $\lambda_j(u) := \cos \pi(j+1/2)u$, $j = 0, 1, 2, \ldots$ defines the Rogosinski-type kernel (see [9]) in the form

$$r_j(t) := \frac{1}{2}\left(\operatorname{sinc}(t+j+\frac{1}{2}) + \operatorname{sinc}(t-j-\frac{1}{2})\right) \tag{16}$$

3) $\lambda_H(u) := \cos^2 \frac{\pi u}{2} = \frac{1}{2}(1+\cos \pi u)$ defines the Hann kernel (see [12])

$$s_H(t) := \frac{1}{2}\frac{\operatorname{sinc} t}{1-t^2}; \tag{17}$$

III. Subordination by typical (Zygmund) sampling operators

In [11] we introduced typical (Zygmund) sampling series $Z^r_w f$ for $f \in C(\mathbb{R})$ with kernels $z_r \in B^1_\pi$ defined via (3) using the window function

$$\lambda_{Z,r}(u) := 1 - u^{2r}, \quad r > 0.$$

We proved an estimate ([11], Th. 1)

$$\|Z^r_w\| \leqslant \frac{2}{\pi}\log r + C \tag{18}$$

Consider now an even bandlimited kernel $s_r \in B^1_\pi$ defined via (3) using the window function λ_r, which has a representation

$$\lambda_r(u) := 1 - \sum_{j=r}^{\infty} c_j\, u^{2j}, \quad r \geqslant 1. \tag{19}$$

The condition (19) is satisfied for many kernels $s \in B^1_\pi$.

If $\sum_{j=r}^{\infty} |c_j| \log j < \infty$ then substituting (19) in (3) and the last one into (1) gives a double series, where interchanging of the order of summation is justified. Therefore, for generalized sampling series in (1) defined by the kernel s_r one has the subordination equalities

$$S^r_w f = \sum_{j=r}^{\infty} c_j Z^j_w f \tag{20}$$

$$S^r_w f - f = \sum_{j=r}^{\infty} c_j (Z^j_w f - f). \tag{21}$$

Theorem 1: Let $f \in \Lambda^p$ for $1 \leqslant p < \infty$, $r \in \mathbb{N}$. Then

$$\|Z^r_w f - f\|_p \leqslant M_r \tau_{2r}(f; \frac{1}{w})_p. \tag{22}$$

The constants M_r are independent of f and w. Moreover, if $f \in \Lambda^p \cap R_{loc}(\mathbb{R})$ for $1 \leqslant p < \infty$, we have

$$\lim_{w\to\infty} \|Z^r_w f - f\|_p = 0. \tag{23}$$

Proof: We apply Proposition 1. For (9) in Proposition 1 we have for $f \in \Lambda^p$ by Proposition 3, (18) and Nikolski inequality

$$\|Z^r_w f\|_p \leqslant \|Z^r_w\|^{1/q} \|z_r\|_1^{1/p} \|f\|_{\ell^p(w)} \leqslant \|Z^r_w\| \|f\|_{\ell^p(w)}.$$

Now we show that (10) in Proposition 1 holds. Let $g \in B^p_{\pi w}$. For $f \in W^{2r}_p$ we have

$$\|Z^r_w f - f\|_p \leqslant \|Z^r_w (f-g)\|_p + \|Z^r_w g - g\|_p + \|f-g\|_p \tag{24}$$

By Proposition 3 and (8) we have

$$\begin{aligned}\|Z^r_w (f-g)\|_p &\leqslant \|Z^r_w\|^{1/q} \|z_r\|_1^{1/p} \|f-g\|_{\ell^p(w)} \\ &\leqslant \|Z^r_w\|^{1/q} \|z_r\|_1^{1/p} (\|f-g\|_p + \frac{1}{w}\|f'-g'\|_p). \end{aligned} \tag{25}$$

If $g \in B^p_{\pi w}$, then $S^{sinc}_w g = g$ i.e.

$$g(t) = \sum_{k\in\mathbb{Z}} g\Big(\frac{k}{w}\Big) \int_0^1 \cos(\pi(wt-k)u)\, du.$$

Hence on the right hand side the series is uniformly convergent and after term-by-term differentiation we get also a uniformly convergent series (cf. [2], Th. 3.3). Therefore for $r \in \mathbb{N}$

$$\frac{(-1)^r}{(\pi w)^{2r}} g^{(2r)}(t) = \sum_{k\in\mathbb{Z}} g\Big(\frac{t}{w}\Big) \int_0^1 u^{2r} \cos(\pi(wt-k)u)\, du \tag{26}$$

Now by the definition of Z^r_w it follows

$$\begin{aligned}\|Z^r_w g - g\|_p &= \frac{1}{(\pi w)^{2r}} \|g^{(2r)}\|_p \\ &\leqslant \frac{1}{(\pi w)^{2r}} (\|f^{(2r)} - g^{(2r)}\|_p + \|f^{(2r)}\|_p). \end{aligned} \tag{27}$$

Substituting (25) and (27) in (24) and choosing finally the function g as $g_w \in B^p_{\pi w}$ from Proposition 2 it follows

$$\|Z^r_w f - f\|_p \leqslant K_r \frac{1}{w^{2r}} \|f^{(2r)}\|_p$$

and (10) is fullfilled. Proposition 1 yields (22). The last assertion (23) follows from (22) and (7). ∎

Theorem 2: Let sampling operator S^r_w ($w > 0$) be defined by the kernel (3) with $\lambda = \lambda_r$ and for some $r \in \mathbb{N}$ let

$$\lambda_r(u) := 1 - \sum_{j=r}^{\infty} c_j u^{2j}, \quad \sum_{j=r}^{\infty} |c_j| \log j \leqslant \infty. \tag{28}$$

Then for $f \in \Lambda^p$ $(1 \leqslant p < \infty)$

$$\|S^r_w f - f\|_p \leqslant M_r \tau_{2r}(f; \frac{1}{w})_p. \tag{29}$$

The constants M_r are independent of f and w. Moreover, if $f \in \Lambda^p \cap R_{loc}(\mathbb{R})$ for $1 \leqslant p < \infty$, we have

$$\lim_{w\to\infty} \|S_w^r f - f\|_p = 0. \tag{30}$$

PROOF: We apply Proposition 1. For (9) in Proposition 1 we have for $f \in \Lambda^p$ by (20), (18), Proposition 3 and Nikolski inequality

$$\|S_w^r f\|_p \leqslant \|f\|_{\ell^p(w)} \sum_{j=r}^{\infty} |c_j| \log j$$

Now we show that (10) in Proposition 1 holds. Let $g \in B^p_{\pi w}$. For $f \in W_p^{2r}$ we have

$$\|S_w^r f - f\|_p \leqslant \|S_w^r (f-g)\|_p + \|S_w^r g - g\|_p + \|f-g\|_p \tag{31}$$

The subordination equality (21) gives the estimate

$$\|S_w^r g - g\|_p \leqslant \sum_{j=r}^{\infty} |c_j| \|Z_w^j g - g\|_p$$

Now we show that for $g \in B^p_{\pi w}$ and $s \leqslant r$ there holds the estimate $\|Z_w^r g - g\|_p \leqslant \|Z_w^s g - g\|_p$. Using (26) and the definition of Z_w^r we have

$$Z_w^j g(t) - g(t) = -(\pi w)^{-2} \Big((Z_w^{j-1} g)''(t) - g''(t) \Big) \tag{32}$$

Applying ([14], Th. 6.11 and Lemma 6.6) we have $Z_w^j g \in B^1_{\pi w} \subset B^p_{\pi w}$, hence $(Z_w^j g - g) \in B^p_{\pi w}$ and we can use the Bernstein inequality for $1 \leqslant p \leqslant \infty$

$$\|(Z_w^{j-1} g)'' - g''\|_p \leqslant (\pi w)^2 \|Z_w^{j-1} g - g\|_p,$$

hence

$$\|Z_w^j g - g\|_p \leqslant \|Z_w^{j-1} g - g\|_p,$$

and we have

$$\|S_w^r g - g\|_p \leqslant \|Z_w^r g - g\|_p \sum_{j=r}^{\infty} |c_j|.$$

Finally we use (27) and substitute the resulting estimate in (31). The rest of the proof is the same as for Theorem 1. ∎

IV. EXAMPLES

Now we apply Theorem 2 for some sampling operators.

Theorem 3: Let the Rogosinski-type sampling operator $R_{w,j}$ $(j = 0, 1, 2, \ldots)$ be defined by the kernel (16). Then for $f \in \Lambda^p$ $(1 \leqslant p < \infty)$

$$\|R_{w,j} f - f\|_p \leqslant M_j \tau_2(f; \frac{1}{w})_p.$$

The constants M_j are independent of f and w.

PROOF: We have for the Rogosinski-type window function

$$\lambda_j(u) = \cos\pi\big(j+\frac{1}{2}\big)u = 1 - \sum_{k=1}^{\infty} (-1)^{k+1} \frac{\pi^{2k}(j+1/2)^{2k}}{(2k)!} u^{2k}$$

and obviously

$$\sum_{k=1}^{\infty} \frac{\pi^{2k}(j+1/2)^{2k}}{(2k)!} \log k < \infty. \quad ∎$$

Theorem 4: Let the Hann sampling operator H_w be defined by the kernel (17). Then for $f \in \Lambda^p$ $(1 \leqslant p < \infty)$

$$\|H_w f - f\|_p \leqslant M \tau_2(f; \frac{1}{w})_p.$$

The constant M is independent of f and w.

PROOF: We have for the Hann window function

$$\lambda_H(u) = \frac{1}{2}(1 + \cos\pi u) = 1 - \sum_{k=1}^{\infty} (-1)^{k+1} \frac{\pi^{2k}}{2(2k)!} u^{2k}. \quad ∎$$

ACKNOWLEDGMENT

This research was partially supported by the Estonian Sci. Foundation, grants 8627 and 9383, and by the Estonian Min. of Educ. and Research, project SF0140011s09.

REFERENCES

[1] P. L. Butzer, G. Schmeisser, and R. L. Stens, "An introduction to sampling analysis," in *Nonuniform Sampling, Theory and Practice*, F. Marvasti, Ed. New York: Kluwer, 2001, pp. 17–121.

[2] P. L. Butzer, W. Splettster, and R. L. Stens, "The sampling theorems and linear prediction in signal analysis." *Jahresber. Deutsch. Math-Verein*, vol. 90, pp. 1–70, 1988.

[3] R. L. Stens, "Sampling with generalized kernels," in *Sampling Theory in Fourier and Signal Analysis: Advanced Topics*, J. R. Higgins and R. L. Stens, Eds. Oxford: Clarendon Press, 1999.

[4] P. L. Butzer, C. Bardaro, R. L. Stens, and G. Vinti, "Approximation error of the Whittaker cardinal series in terms of an averaged modulus of smoothness covering discontinuous signals." *J. Math. Anal. Appl.*, vol. 316, pp. 269–306, 2006.

[5] P. Butzer, J. Higgins, and R. Stens, "Classical and approximate sampling theorems; studies in the $L^p(\mathbb{R})$ and the uniform norm," *Journal of Approximation Theory*, vol. 137, pp. 250–263, 2005.

[6] P. L. Butzer and R. L. Stens, "Reconstruction of signals in $L^p(\mathbb{R})$-space by generalized sampling series based on linear combinations of B-splines." *Integral Transforms Spec. Funct.*, vol. 19, pp. 35–58, 2008.

[7] P. L. Butzer, C. Bardaro, R. L. Stens, and G. Vinti, "Prediction by samples from the past with error estimates covering discontinious signals." *IEEE Transactions on Information Theory*, vol. 56, pp. 614–633, 2010.

[8] A. Kivinukk, "Approximation of continuous functions by Rogosinski-type sampling series," in *Modern Sampling Theory: Mathematics and Applications*, J. Benedetto and P. Ferreira, Eds. Birkhuser Verlag, 2001, pp. 229 – 244.

[9] A. Kivinukk and G. Tamberg, "On sampling series based on some combinations of sinc functions." *Proc. of the Estonian Academy of Sciences. Physics Mathematics*, vol. 51, pp. 203–220, 2002.

[10] Z. Burinska, K. Runovski, and H.-J. Schmeisser, "On the approximation by generalized sampling series in L_p-metrics." *Sampling Theory in Signal and Image Processing*, vol. 5, pp. 59–87, 2006.

[11] A. Kivinukk, "Approximation by typical sampling series," in *Proc. 1999 Intern. Workshop on Sampling Theory and Applications, Loen, Norway*. Norwegian Univ. Sci. and Technology, 1999, pp. 161–166.

[12] A. Kivinukk and G. Tamberg, "On sampling operators defined by the Hann window and some of their extensions." *Sampling Theory in Signal and Image Processing*, vol. 2, pp. 235–258, 2003.

[13] B. Sendov and V. Popov, *The Averaged Moduli of Smoothness*. Chichester: Wiley, 1988.

[14] J. R. Higgins, *Sampling Theory in Fourier and Signal Analysis*. Oxford: Clarendon Press, 1996.

[15] H. H. Albrecht, "A family of cosine-sum windows for high resolution measurements," in *IEEE International Conference on Acoustics, Speech and Signal Processing, Salt Lake City, Mai 2001*. Salt Lake City: IEEE, 2001, pp. 3081–3084.

[16] R. B. Blackman and J. W. Tukey, *The measurement of power spectra*. New York: Wiley-VCH, 1958.

[17] F. J. Harris, "On the use of windows for harmonic analysis." *Proc. of the IEEE*, vol. 66, pp. 51–83, 1978.

[18] H. D. Meikle, *A New Twist to Fourier Tansforms*. Berlin: Dover, 2004.

Sparse Recovery with Fusion Frames via RIP

Ulaş Ayaz
Hausdorff Center for Mathematics
Institute for Numerical Simulation
University of Bonn
Endenicher Allee 60, 53115 Bonn, Germany
Email: ulas.ayaz@hcm.uni-bonn.de

Holger Rauhut
Hausdorff Center for Mathematics
Institute for Numerical Simulation
University of Bonn
Endenicher Allee 60, 53115 Bonn, Germany
Email: rauhut@hcm.uni-bonn.de

***Abstract*—We extend ideas from compressed sensing to a structured sparsity model related to fusion frames. We present theoretical results concerning the recovery of sparse signals in a fusion frame from undersampled measurements. We provide both nonuniform and uniform recovery guarantees. The novelty of our work is to exploit an incoherence property of the fusion frame which allows us to reduce the number of measurements needed for sparse recovery.**

I. Introduction

Compressed sensing (CS) predicts that one can efficiently recover a sparse vector from few measurements by solving a convex optimization problem [1]–[3]. Often signals possess more structure than mere sparsity, and exploiting such structure often allows to further reduce the amount of required measurements, see, e.g., [4]. In this paper, we investigate a structured sparsity model related to fusion frames. These were introduced as generalizations of classical frames, in order to better capture the richness of multidimensional signals with an inherent structure [5]. Here, subspaces take the role of the frame vectors.

We investigate sufficient conditions in order to recover a sparse signal in a fusion frame via mixed ℓ_1/ℓ_2 minimization. We both give nonuniform and uniform recovery guarantees. The uniform recovery result is based on the fusion RIP introduced in [6]. Hereby, we improve the recovery conditions given in [6] by exploiting the additional information inherent in the fusion frame structure.

II. Fusion Frames

A *fusion frame* for $\mathbb{R}^d$ is a collection of N subspaces $W_j \subset \mathbb{R}^d$ and associated weights v_j that satisfies

$$A\|x\|_2^2 \leq \sum_{j=1}^{N} v_j^2 \|P_j x\|_2^2 \leq B\|x\|_2^2$$

for all $x \in \mathbb{R}^d$ and for some universal fusion frame bounds $0 < A \leq B < \infty$, where $P_j \in \mathbb{R}^{d\times d}$ denotes the orthogonal projection onto the subspace W_j. For simplicity we assume that the dimensions of the W_j are equal, $\dim(W_j) = k$.

For a fusion frame $(W_j)_{j=1}^N$, let us define the Hilbert space $\mathcal{H}$ as

$$\mathcal{H} = \{(x_j)_{j=1}^N : x_j \in W_j,\ \forall j \in [N]\} \subset \mathbb{R}^{d\times N},$$

where we denote $[N] = \{1, \dots, N\}$. The *mixed* $\ell_{2,1}$*-norm* of a vector $\mathbf{x} \equiv (x_j)_{j=1}^N \in \mathcal{H}$ is defined as

$$\|(x_j)_{j=1}^N\|_{2,1} \equiv \sum_{j=1}^{N} \|x_j\|_2.$$

Furthermore, the 'ℓ_0-norm' (which is actually not even a quasi-norm) is defined as

$$\|\mathbf{x}\|_0 = \sharp\{j \in [N] : x_j \neq 0\}.$$

We call a vector $\mathbf{x} \in \mathcal{H}$ s-sparse, if $\|\mathbf{x}\|_0 \leq s$. Our sparsity model requires that the 'blocks' x_j are either zero or nonzero as a whole.

A. Sparse Recovery Problem

We take m linear combinations of an s-sparse vector $\mathbf{x}^0 = (x_j^0)_{j=1}^N \in \mathcal{H}$, i.e.,

$$\mathbf{y} = (y_i)_{i=1}^m = \left(\sum_{j=1}^{N} a_{ij} x_j^0\right)_{i=1}^m, \quad y_i \in \mathbb{R}^d.$$

Let us denote the block matrices $\mathbf{A_I} = (a_{ij} I_d)_{i\in[m], j\in[N]}$ and $\mathbf{A_P} = (a_{ij} P_j)_{i\in[m], j\in[N]}$ that consist of the blocks $a_{ij} I_d$ and $a_{ij} P_j$ respectively. Here I_d is the identity matrix of size $d \times d$. Then we can formulate this measurement scheme as

$$\mathbf{y} = \mathbf{A_I}\mathbf{x}^0 = \mathbf{A_P}\mathbf{x}^0.$$

We can replace $\mathbf{A_I}$ by $\mathbf{A_P}$ since the relation $P_j x_j = x_j$ holds for all $\mathbf{x} \in \mathcal{H}$ and $j \in [N]$. We wish to recover $\mathbf{x}^0$ from $\mathbf{y}$. This task can be stated as

$$(L0) \quad \hat{\mathbf{x}} = \text{argmin}_{x\in\mathcal{H}} \|\mathbf{x}\|_0 \quad s.t. \quad \mathbf{A_P}\mathbf{x} = \mathbf{y}.$$

This optimization problem is NP-hard. Therefore, we instead propose the following program

$$(L1) \quad \hat{\mathbf{x}} = \text{argmin}_{x\in\mathcal{H}} \|\mathbf{x}\|_{2,1} \quad s.t. \quad \mathbf{A_P}\mathbf{x} = \mathbf{y}.$$

B. Relation with Previous Work

A special case of the sparse recovery problem above appears when all subspaces coincide with the ambient space $W_j = \mathbb{R}^d$ for all j. Then the problem reduces to the well studied *joint sparsity setup* [7] in which all the vectors have the same sparsity structure.

Furthermore, our problem is itself a special case of the *block sparsity setup* [8], with significant additional structure that allows us to enhance existing results. In fact, the fusion frame model assumes the additional prior knowledge that the x_j's are contained in the fusion frame subspaces W_j.

Finally in the case $d = 1$, the projections equal 1, and hence the problem reduces to the *classical recovery problem* $Ax = y$ with $x \in \mathbb{R}^N$ and $y \in \mathbb{R}^m$.

C. Incoherence Parameter

We define the parameter λ as a measure of the coherence of the fusion frame subspaces as

$$\lambda = \max_{i \neq j} \|P_i P_j\|_{2\to 2}, \quad i, j \in [N].$$

Note that $\|P_i P_j\|_{2\to 2}$ equals the largest absolute value of the cosines of the principle angles between W_i and W_j. Observe that if the subspaces are all orthogonal to each other, i.e., $\lambda = 0$, then only one measurement suffices to recover $\mathbf{x}^0$ as $y_1 = \sum_j a_{1j} x_j^0$ is an orthogonal decomposition. This observation suggests that fewer measurements are necessary when λ gets smaller. In this work our goal is to provide a solid theoretical understanding of this observation.

D. A Nonuniform Result

We first consider the recovery of a fixed sparse signal from random measurements. To this end, we introduce the Gaussian matrix whose entries consist of independent standard normal distributed random variables and the Bernoulli matrix where the entries are independent random variables taking the values ± 1 with equal probability.

Theorem II.1. *Let $(W_j)_{j=1}^N$ be given with parameter $\lambda \in [0, 1]$ and $\mathbf{x} \in \mathcal{H}$ be s-sparse. Let $A \in \mathbb{R}^{m\times N}$ be a Bernoulli or Gaussian matrix. Assume that*

$$m \geq C(1 + \lambda s) \ln^{\alpha}(\max\{N, sd\}) \ln(\varepsilon^{-1}), \tag{1}$$

where $C > 0$ is a universal constant. Then with probability at least $1 - \varepsilon$, $(L1)$ recovers $\mathbf{x}$ from $\mathbf{y} = \mathbf{A_P x}$. Here $\alpha = 1$ in the Bernoulli case and $\alpha = 2$ in the Gaussian case.

We provide an outline of the proof in [9]. We remark that Theorems II.1 is also shown to be stable with respect to noise on the measurements and under passing to approximately sparse signals.

III. Sparse recovery using "fusion" RIP

In this section we study uniform recovery of sparse fusion frame signals from their random measurements. One common way to study such recovery conditions is via the restricted isometry property (RIP). A version adapted to fusion frames has been introduced in [6].

Definition III.1 (Fusion RIP)**.** Let $A \in \mathbb{R}^{m\times N}$ and $(W_j)_{j=1}^N$ be a fusion frame for $\mathbb{R}^d$. The fusion restricted isometry constant δ_s is the smallest constant such that

$$(1 - \delta_s)\|\mathbf{x}\|_2^2 \leq \|\mathbf{A_P x}\|_2^2 \leq (1 + \delta_s)\|\mathbf{x}\|_2^2 \tag{2}$$

for all $\mathbf{x} \in \mathcal{H}$ of sparsity $\|\mathbf{x}\|_0 \leq s$.

The following result was also shown in [6].

Proposition III.2 (Fusion RIP implies recovery)**.** *Let $(A, (W_j)_{j=1}^N)$ with fusion RIP constant $\delta_{2s} < 1/3$. Then $(L1)$ recovers all s-sparse $\mathbf{x}$ from $\mathbf{y} = \mathbf{A_P x}$.*

This result shows that given a fusion frame $(W_j)_{j=1}^N$ and matrix A, for uniform recovery it is enough to check whether the block matrix $\mathbf{A_P}$ satisfies the fusion RIP. Recovery is also stable under noise and passing to compressible signals. Another result from [6] tells us that if the underlying random measurement matrix A satisfies the classical RIP, $\mathbf{A_P}$ satisfies fusion RIP with same constants. This suggests that $m \gtrsim s \ln(N/s)$ is sufficient for many random measurement ensembles (up to some log factors). However, the following main result of our work shows that the inherent structure of fusion frames provides additional information that can be exploited to derive stronger recovery conditions.

Theorem III.3. *Let $(W_j)_{j=1}^N$ be given with $\dim(W_j) = k$ and parameter $\lambda \in [0, 1]$. Let $A \in \mathbb{R}^{m\times N}$ be a Bernoulli matrix and $\delta \in (0, 1)$. Assume that*

$$m \geq C\delta^{-2} k \sqrt{\lambda s^2 + s} \ln^4(\max\{N, d\}). \tag{3}$$

Then with probability at least $1 - 2e^{-c\delta^2 m}$, the fusion RIP constant δ_s of $\tilde{\mathbf{A}}_\mathbf{P} = \frac{1}{\sqrt{m}}\mathbf{A_P}$ satisfies $\delta_s \leq \delta$. Above $C, c > 0$ are universal constants.

Theorem III.3 can be extended for the random matrices with independent subgaussian entries. Presently the uniform result (3) behaves slightly worse than the nonuniform one (1) for small λ and suffers from additional log-terms. On the other hand, we gain uniformity and stronger stability.

IV. Proof outline

Due to lack of space, we only present the outline of the proof of Theorem III.3. The detailed proof will appear in a forthcoming journal publication. Let us first give a characterization of the fusion RIP constant. The definition (2) implies that

$$\delta_s = \sup_{\mathbf{x}\in D_{s,N}} \left| \|\tilde{\mathbf{A}}_\mathbf{P}\mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2 \right|,$$

where $D_{s,N} := \{\mathbf{x} \in \mathcal{H} : x_i \in W_i, \|\mathbf{x}\|_2 \leq 1, \|\mathbf{x}\|_0 \leq s\}$. Next we derive an estimate for the expectation of δ_s. To this end, we denote $\mathbf{E}_{ij}(Y) \in \mathbb{R}^{md\times Nd}$, $i \in [m], j \in [N]$, as the block matrix (consisting of $m \times N$ blocks) with a single block entry $Y \in \mathbb{R}^{d\times d}$ at position (i, j) and the entry $0 \in \mathbb{R}^{d\times d}$ elsewhere. Let ϵ_{ij} be the entries of A and observe that

$$\tilde{\mathbf{A}}_\mathbf{P}\mathbf{x} = \frac{1}{\sqrt{m}} \sum_{i\in[m], j\in[N]} \epsilon_{ij}(\mathbf{Q}_{ij}\mathbf{x}),$$

where $\mathbf{Q}_{ij} := \mathbf{E}_{ij}(P_j)$. We define the matrix $V_\mathbf{x}$ whose columns are $\frac{1}{\sqrt{m}}\mathbf{Q}_{ij}\mathbf{x}$ for all i, j, i.e.,

$$V_\mathbf{x} = \frac{1}{\sqrt{m}}(\mathbf{Q}_{11}\mathbf{x}|\mathbf{Q}_{12}\mathbf{x}|\dots|\mathbf{Q}_{mN}\mathbf{x}).$$

Then we can write $\tilde{\mathbf{A}}_{\mathbf{P}}\mathbf{x} = V_{\mathbf{x}}\boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ is a Bernoulli vector of length mN. Denoting the set $\mathcal{A} = \{V_{\mathbf{x}} : \mathbf{x} \in D_{s,N}\}$, we have

$$\delta_s = \sup_{\mathbf{x}\in D_{s,N}} \left| \|V_{\mathbf{x}}\boldsymbol{\epsilon}\|_2^2 - \|\mathbf{x}\|_2^2 \right| = \sup_{A\in\mathcal{A}} \left| \|A\boldsymbol{\epsilon}\|_2^2 - \mathbb{E}\|A\boldsymbol{\epsilon}\|_2^2 \right|.$$

Following Krahmer et al. [10] where they use chaining methods in order to get bounds for this type of random variables, we obtain

$$\mathbb{E}\sup_{A\in\mathcal{A}} \left| \|A\boldsymbol{\epsilon}\|_2^2 - \mathbb{E}\|A\boldsymbol{\epsilon}\|_2^2 \right| \lesssim d_F(\mathcal{A}) d_{2\to 2}(\mathcal{A}) + \left(d_F(\mathcal{A})\gamma_2(\mathcal{A}, \|\cdot\|_{2\to 2}) + \gamma_2(\mathcal{A}, \|\cdot\|_{2\to 2})^2 \right). \quad (4)$$

Here, $d_F(\mathcal{A})$ and $d_{2\to 2}(\mathcal{A})$ denote the radius of $\mathcal{A}$ in the Frobenius and the operator norms, respectively. For the definition of Talagrand's γ_2-functional we refer to [11]. It is easy to check that

$$d_{2\to 2}(\mathcal{A}) = \sup_{\mathbf{x}\in D_{s,N}} \|V_{\mathbf{x}}\|_{2\to 2} \le 1/\sqrt{m} \quad \text{and} \quad d_F(\mathcal{A}) = 1.$$

The γ_2-functional can be estimated by the well-known Dudley integral [11]

$$\gamma_2(\mathcal{A}, \|\cdot\|_{2\to 2}) \lesssim \int_0^{d_{2\to 2}(\mathcal{A})} \sqrt{\ln \mathcal{N}(\mathcal{A}, \|\cdot\|_{2\to 2}, u)}\, du, \quad (5)$$

where the covering number $\mathcal{N}(T, d, u)$ is defined as the smallest number of open balls of radius u in (T, d) needed to cover T. Therefore estimating the expectation in (4) amounts to estimating covering numbers which will perform in two different ways similar to [12].

a) Small values of u: For $S \subset [N]$ we introduce the set $B_S^2 := \{\mathbf{x} : \operatorname{supp}(\mathbf{x}) \subset S, \|\mathbf{x}\|_2 \le 1\}$. Furthermore define the norm $|||\mathbf{x}||| := \|V_{\mathbf{x}}\|_{2\to 2}$. Observe that $|||\mathbf{x}||| \le \frac{1}{\sqrt{m}}\|\mathbf{x}\|_2$. Then using subadditivity of covering numbers and a standard volumetric argument (see, e.g., [13, Chapter 8.4]) we obtain

$$\begin{aligned}
&\mathcal{N}(\mathcal{A}, \|\cdot\|_{2\to 2}, u) = \mathcal{N}(D_{s,N}, |||\cdot|||, u) \\
&\le \sum_{\substack{S\subset[N]\\|S|=s}} \mathcal{N}\left(B_S^2, |||\cdot|||, u\right) \le \sum_{\substack{S\subset[N]\\|S|=s}} \mathcal{N}\left(B_S^2, \frac{\|\cdot\|_2}{\sqrt{m}}, u\right) \\
&= \sum_{\substack{S\subset[N]\\|S|=s}} \mathcal{N}\left(B_S^2, \|\cdot\|_2, u\sqrt{m}\right) \le \left(\frac{eN}{s}\right)^s \left(1 + \frac{2}{u\sqrt{m}}\right)^{sk}.
\end{aligned}$$

For $u > 0$, it thus holds

$$\ln \mathcal{N}(\mathcal{A}, \|\cdot\|, u) \le s\ln(eN/s) + sk\ln\left(1 + \frac{2}{u\sqrt{m}}\right). \quad (6)$$

b) Large values of u: We define the set

$$B_{2,1} := \left\{ \mathbf{x} \in \mathcal{H} : \|\mathbf{x}\|_{2,1} \le 1, \|\mathbf{x}\|_2 \le \frac{1}{\sqrt{s}} \right\}.$$

Then it is evident that $D_{s,N} \subset \sqrt{s}B_{2,1}$. Therefore,

$$\begin{aligned}
&\mathcal{N}(\mathcal{A}, \|\cdot\|_{2\to 2}, u) = \mathcal{N}(D_{s,N}, |||\cdot|||, u) \\
&\le \mathcal{N}(\sqrt{s}B_{2,1}, |||\cdot|||, u) = \mathcal{N}\left(B_{2,1}, |||\cdot|||, \frac{u}{\sqrt{s}}\right). \quad (7)
\end{aligned}$$

For the task of estimating $\mathcal{N}(B_{2,1}, |||\cdot|||, u)$, we invoke the so-called *empirical method of Maurey*. We fix $u > 0$ and $\mathbf{x} \in B_{2,1}$. The idea is to approximate $\mathbf{x}$ by a finite set of very sparse vectors of ℓ_2-norm 1. In order to construct this set, we discretize the unit sphere of each frame subspace W_j. Denote $S_j = \{\mathbf{y} \in \mathcal{H} : \|y_j\|_2 = 1; y_i = 0,\ i \ne j\}$. A volumetric argument yields that

$$\mathcal{N}(S_j, \|\cdot\|_2, \tilde{\varepsilon}) \le \left(1 + \frac{2}{\tilde{\varepsilon}}\right)^k.$$

For each j, let $T_j \subset S_j$ be the covering set of S_j with this cardinality. We will use 1-sparse elements from the set $\mathcal{T} = \bigcup_{j\in[N]} T_j$ in order to find a vector $\mathbf{z}$ that is close to $\mathbf{x}$. To this end, we define a random vector $\tilde{\mathbf{Z}}$ as follows

$$\mathbb{P}\left(\tilde{\mathbf{Z}} = \overrightarrow{\mathbf{E}}_j\left(\frac{x_j}{\|x_j\|_2}\right)\right) = \|x_j\|_2 \quad \text{for } j \in [N],$$

and $\tilde{\mathbf{Z}} = 0$ with probability $1 - \|\mathbf{x}\|_{2,1}$. Here the notation $\overrightarrow{\mathbf{E}}_j(x)$ corresponds to the block column vector of size N with the vector x in j-th position and 0 elsewhere. Observe that $\mathbb{E}\tilde{\mathbf{Z}} = \mathbf{x}$. Let M be a number to be determined later. Let $\tilde{\mathbf{Z}}_1, \ldots, \tilde{\mathbf{Z}}_M$ be independent copies of $\tilde{\mathbf{Z}}$, and put

$$\tilde{\mathbf{z}} = \frac{1}{M}\sum_{\ell=1}^{M} \tilde{\mathbf{Z}}_\ell.$$

We now denote $\mathbf{Z}_\ell \in \mathcal{T}$ as the closest vector to $\tilde{\mathbf{Z}}_\ell$ in the set $\mathcal{T}$ for all ℓ. Then we have $\|\tilde{\mathbf{Z}}_\ell - \mathbf{Z}_\ell\|_2 \le \tilde{\varepsilon}$. The M-sparse vector $\mathbf{z} = \frac{1}{M}\sum_{\ell=1}^{M} \mathbf{Z}_\ell$ will be our candidate to approximate $\mathbf{x}$. By the triangle inequality

$$|||\mathbf{z} - \mathbf{x}||| \le |||\mathbf{z} - \tilde{\mathbf{z}}||| + |||\tilde{\mathbf{z}} - \mathbf{x}|||. \quad (8)$$

With the choice $\tilde{\varepsilon} = \frac{u\sqrt{m}}{2}$, it is not hard to deduce $|||\mathbf{z} - \tilde{\mathbf{z}}||| \le u/2$. It remains to show that $|||\tilde{\mathbf{z}} - \mathbf{x}||| \le u/2$ with nonzero probability for large enough M. Since $|||\tilde{\mathbf{z}} - \mathbf{x}||| = \left\| \frac{1}{M}\sum_{\ell=1}^{M} (V_{\tilde{\mathbf{Z}}_\ell} - V_{\mathbf{x}}) \right\|_{2\to 2}$ is a sum of centered random matrices, we may invoke the noncommutative Bernstein inequality due to Tropp [14, Theorem 1.6] in order the bound the tail probability of this norm. This leads to the condition

$$M \ge \ln(md + mN)\left(\frac{16\sqrt{\lambda + 1/s}}{mu^2} + \frac{3}{\sqrt{m}u}\right), \quad (9)$$

which implies the existence of a realization of the vector $\tilde{\mathbf{z}}$ for which $|||\tilde{\mathbf{z}} - \mathbf{x}||| \le u/2$. Together with (8) this yields $|||\mathbf{z} - \mathbf{x}||| \le u$. Since each $\mathbf{Z}_\ell \in \mathcal{T}$ takes at most

$$|\mathcal{T}| = \bigcup_{j\in[N]} |T_j| \le N\left(1 + \frac{4}{u\sqrt{m}}\right)^k$$

many values, $\mathbf{z}$ can take at most $N^M\left(1 + \frac{4}{u\sqrt{m}}\right)^{kM}$ values. Setting M to the least integer that satisfies (9), we deduce that the covering numbers can be estimated by

$$\sqrt{\ln \mathcal{N}(B_{2,1}, |||\cdot|||, u)} \le \sqrt{\ln\left[N^M\left(1 + \frac{4}{u\sqrt{m}}\right)^{kM}\right]}$$

$$\leq \sqrt{\frac{16\sqrt{\lambda+1/s}}{mu^2}+\frac{3}{\sqrt{m}u}}\sqrt{k\ln(D)\ln\left[N\left(1+\frac{4}{u\sqrt{m}}\right)\right]}, \tag{10}$$

where $D := md + mN$. Finally we estimate the Dudley integral (5) by integrating (6) from 0 to a suitable $\kappa \in (0, 1/\sqrt{m})$ and (10) from κ to $1/\sqrt{m}$ with replacing u by $u/\sqrt{s}$ due to (7). Plugging all estimates derived for $d_{2\to 2}(\mathcal{A})$, $d_F(\mathcal{A})$ and $\gamma_2(\mathcal{A}, \|\cdot\|_{2\to 2})$ into (4), we obtain $\mathbb{E}\delta_s \leq \delta$, provided Condition (3) of Theorem III.3 holds with an appropriate constant.

The probability estimate for δ_s is derived by applying a concentration inequality provided also in [10] having all complexity parameters at hand. This completes the proof.

V. Numerical Experiments

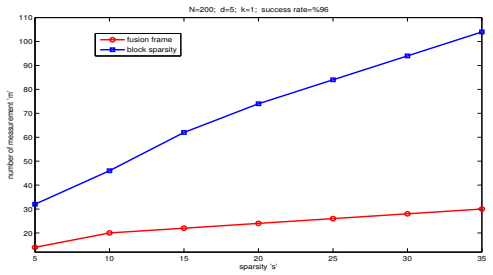
In this section, we compare two sparsity models: Fusion frame and block sparsity. We present numerical experiments that illustrate that the additional knowledge about the fusion frame subspaces, that is $\mathbf{x} \in \mathcal{H}$, significantly improves the recovery compared to the block sparsity case where we do not assume such a knowledge. (See Section II-B.) In all of our experiments, we use SPGL1 [15], [16] to solve the minimization problems.

a) In Fig.1a, we fix a fusion frame with $N = 200$ subspaces in $\mathbb{R}^d$, $d = 5$ with $k = 1$. Then we vary the sparsity level s from 5 to 35, and generate an s-sparse vector $\mathbf{x}$ in the fusion frame. We form $\mathbf{y} = \mathbf{A_P x}$ with a randomly generated Gaussian matrix $A \in \mathbb{R}^{m\times N}$ for different values of m and solve the minimization problem $(L1)$ with and without the constraint that $\mathbf{x} \in \mathcal{H}$. Repeating this test 50 times for each s for both cases, we record the values of m which yield a recovery success rate of at least $\%96$.

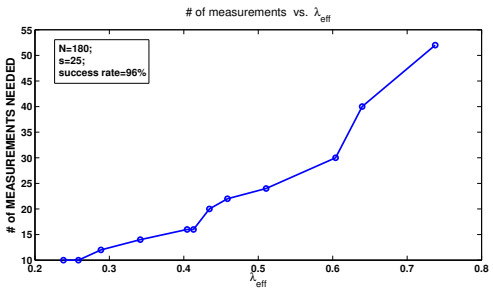
b) Fig.1b depicts a relation between the number of measurements needed m and the incoherence parameter λ_{eff} where

$$\lambda_{\text{eff}} = \frac{1}{s}\max_{i\in[N]}\sum_{j\in S}\|P_iP_j\|_{2\to 2}.$$

In the Bernoulli case, the parameter λ in (1) can be replaced by λ_{eff} which is smaller. To that end, we fix the sparsity level to $s = 25$ and generate various fusion frames with $N = 180$ and different values of λ_{eff}. Then we generate an s-sparse vector in each fusion frame and find the number of measurements m which yields an empirical recovery rate of 96%.



(a)



(b)

Fig. 1. (a) $N = 200, d = 5$ fixed, we vary s and plot the number of measurements needed m for the cases where we assume the knowledge the subspaces (fusion frame) and the general block sparsity case. (b) $N = 180, s = 25$ fixed, we plot the number of measurements m vs. λ_{eff}.

Acknowledgment

The authors would like to thank the Hausdorff Center for Mathematics for support, and acknowledge funding through the WWTF project SPORTS (MA07-004) and the ERC Starting Grant StG 258926.

References

[1] E. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 489–509, 2006.

[2] E. Candès and T. Tao, "Near optimal signal recovery from random projections: universal encoding strategies?" *IEEE Trans. Inf. Theory*, vol. 52, no. 12, pp. 5406–5425, 2006.

[3] D. Donoho, "Compressed sensing," *IEEE Trans. Inform. Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[4] Y. Eldar and H. Rauhut, "Average case analysis of multichannel sparse recovery using convex relaxation," *IEEE Trans. Inform. Theory*, vol. 56, no. 1, pp. 505–519, 2010.

[5] P. G. Casazza, G. Kutyniok, and S. Li, "Fusion frames and distributed processing," *Appl. Comput. Harmon. Anal.*, vol. 254, no. 1, pp. 114–132, 2008.

[6] P. Boufounos, G. Kutyniok, and H. Rauhut, "Sparse recovery from combined fusion frame measurements," *IEEE Trans. Inform. Theory*, vol. 57, no. 6, pp. 3864–3876, 2011.

[7] M. Fornasier and H. Rauhut, "Recovery algorithms for vector valued data with joint sparsity constraints," *SIAM J. Numer. Anal.*, vol. 46, no. 2, pp. 577–613, 2008.

[8] Y. C. Eldar and H. Bölcskei, "Block-sparsity: Coherence and efficient recovery," *CoRR*, vol. abs/0812.0329, 2008.

[9] U. Ayaz and H. Rauhut, "Nonuniform sparse recovery with fusion frames," in *Submitted to Proc. of SPARS'13.*, 2013.

[10] F. Krahmer, S. Mendelson, and H. Rauhut, "Suprema of chaos processes and the restricted isometry property," *Comm. Pure Appl. Math.*, to appear.

[11] M. Talagrand, *The Generic Chaining*, ser. Springer Monographs in Mathematics. Springer-Verlag, 2005.

[12] M. Rudelson and R. Vershynin, "On sparse reconstruction from Fourier and Gaussian measurements," *Comm. Pure Appl. Math.*, vol. 61, pp. 1025–1045, 2008.

[13] H. Rauhut, "Compressive sensing and structured random matrices," in *Theoretical Foundations and Numerical Methods for Sparse Recovery*, ser. Radon Series Comp. Appl. Math., M. Fornasier, Ed. deGruyter, 2010, vol. 9, pp. 1–92.

[14] J. A. Tropp, "User-friendly tail bounds for sums of random matrices," *Foundations of Computational Mathematics*, vol. 12, no. 4, pp. 389–434, 2012.

[15] E. van den Berg and M. P. Friedlander, "SPGL1: A solver for large-scale sparse reconstruction," June 2007, http://www.cs.ubc.ca/labs/scl/spgl1.

[16] ——, "Probing the pareto frontier for basis pursuit solutions," *SIAM Journal on Scientific Computing*, vol. 31, no. 2, pp. 890–912, 2008.

Blind Sensor Calibration in Sparse Recovery Using Convex Optimization

Çağdaş Bilen*, Gilles Puy†, Rémi Gribonval* and Laurent Daudet‡
* INRIA, Centre Inria Rennes - Bretagne Atlantique, 35042 Rennes Cedex, France.
† Institute of Electrical Engineering, Ecole Polytechnique Federale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland
‡ Institut Langevin, CNRS UMR 7587, UPMC, Univ. Paris Diderot, ESPCI, 75005 Paris, France

***Abstract*—We investigate a compressive sensing system in which the sensors introduce a distortion to the measurements in the form of unknown gains. We focus on *blind* calibration, using measures performed on a few unknown (but sparse) signals. We extend our earlier study on real positive gains to two generalized cases (signed real-valued gains; complex-valued gains), and show that the recovery of unknown gains together with the sparse signals is possible in a wide variety of scenarios. The simultaneous recovery of the gains and the sparse signals is formulated as a convex optimization problem which can be solved easily using off-the-shelf algorithms. Numerical simulations demonstrate that the proposed approach is effective provided that sufficiently many (unknown, but sparse) calibrating signals are provided, especially when the sign or phase of the unknown gains are not completely random.**

I. Introduction

Compressed sensing theory shows that K-sparse signals can be sampled at much lower rate than required by the Nyquist-Shannon theorem [1]. More precisely, if $\mathbf{x} \in \mathbb{C}^N$ is a K-sparse source vector then it can be captured by collecting only $M \ll N$ linear measurements

$$y_i = \mathbf{m}_i' \mathbf{x}, \qquad i = 1, \dots, M \tag{1}$$

In the above equation, $\mathbf{m}_1, \dots, \mathbf{m}_M \in \mathbb{C}^N$ are *known* measurement vectors, and $.'$ denotes the conjugate transpose operator. Under certain conditions on the measurement vectors, the signal can be accurately reconstructed by solving, e.g.,

$$\mathbf{x}^*_{\ell_1} = \arg\min_{\mathbf{z}} \|\mathbf{z}\|_1 \quad \text{subject to} \quad y_i = \mathbf{m}_i' \mathbf{z}, \quad i = 1, \dots, M$$

where $\|\cdot\|_1$ denotes the ℓ_1-norm, which favors the selection of sparse signals among the ones satisfying the measurement constraints. It has been shown that the number of measurements needed for accurate recovery of $\mathbf{x}$ scales only linearly with K [1]. Note that the above minimization problem can easily be modified to handle the presence of additive noise on the measurements.

Unfortunately, in some practical situations, it is sometimes not possible to perfectly know the measurement vectors $\mathbf{m}_1, \dots, \mathbf{m}_M$. In many applications dealing with distributed sensors or radars, the location or intrinsic parameters of the sensors are not exactly known, which in turn results in unknown phase shifts and/or gains at each sensor [2], [3]. Similarly, applications with microphone arrays are shown to require calibration of each microphone to account for the unknown gain and phase shifts introduced [4]. Unlike additive perturbations in the measurement matrix, this multiplicative perturbation may introduce significant distortion if ignored [5], [6].

In this paper, we investigate the problem of estimating the unknown gains introduced by the sensors when multiple unknown but sparse input signals are measured. We extend the convex optimization approach dealing with positive real gains proposed in [7] to the case of signed real-valued and complex-valued gains which is more realistic from the application perspective. In addition to identifying the additional challenges introduced by the more difficult problem, we further demonstrate the performance of the proposed algorithms in cases where the unknown phase shifts (or sign changes) introduced by the sensors are not completely random.

II. Problem Definition

Suppose that the measurement system in (1) is perturbed by complex gains at each sensor i and there are multiple sparse input signals, $\mathbf{x}_l \in \mathbb{C}^N, \quad l = 1 \dots L$, applied to the system such that

$$y_{i,l} = d_i e^{j\theta_i} \mathbf{m}_i' \mathbf{x}_l \qquad i = 1 \dots M,\ \theta_i \in [0, 2\pi),\ d_i \in \mathbb{R}^+ \tag{2}$$

For a real valued system, the phase term $e^{j\theta_i}$ is replaced by ∓ 1 (or $\theta_i \in \{0, \pi\}$). We focus only on the noiseless case for the sake of simplicity.

It should be noted that, unlike the case with positive real gains, ignoring the unknown gains during recovery is not a viable option when dealing with signed real or complex gains even when the magnitude of the gains are constant. This is due to the significant distortion introduced by the change in sign (and phase). Therefore it is essential to employ a reconstruction approach that deals with the unknown gains.

In a traditional recovery strategy, one can enforce the sparsity of the input signals while enforcing the measurement constraints in (2). However, when dealing with unknown gains, the measurement constraints are non-linear with respect to the unknowns d_i and $\mathbf{x}_l$. This non-linearity can be dealt

This work was partly funded by the Agence Nationale de la Recherche (ANR), project ECHANGE (ANR-08-EMER-006) and by the European Research Council, PLEASE project (ERC-StG-2011-277906). LD is on a joint affiliation between Univ. Paris Diderot and Institut Universitaire de France.

with by using an alternate minimization strategy where one iteratively estimates $\mathbf{x}$ while keeping d_i fixed and vice-versa [2]. However, the convergence of this alternating optimization to the global minimum is not guaranteed since there is a chance that the algorithm gets stuck in a local minimum.

A. Proposed Approach

The recovery of $\mathbf{x}_l$, $l = 1 \dots L$ and d_i, $i = 1 \dots M$ with convex optimization when $e^{j\theta_i}$ are known has been studied in [7]. In this paper, we extend the same approach to systems with signed real-valued and complex-valued gains. Therefore the term $d_i e^{j\theta_i}$ will henceforth simply be represented as $d_i \in \mathbb{R}$ for real-valued systems and $d_i \in \mathbb{C}$ for complex-valued systems.

As an alternative to the alternating non-linear optimization described above, the measurement equation (2) can be reorganized in a bi-linear fashion such that

$$y_{i,l}\tau_i = \mathbf{m}_i'\mathbf{x}_l \qquad i = 1 \dots M \ , \ l = 1 \dots L \qquad (3)$$
$$\tau_i \triangleq \frac{1}{d_i}$$

assuming that $d_i \neq 0 \ \forall i$. Consequently, one can attempt to recover the sparse signals and the gains with the convex optimization

$$\begin{matrix}\mathbf{x}_1^*, \dots, \mathbf{x}_L^* \\ \tau_1^*, \dots, \tau_M^*\end{matrix} = \underset{\substack{\mathbf{z}_1,\dots,\mathbf{z}_L \\ t_1,\dots,t_M}}{\arg\min} \ \sum_{n=1}^{L} \|\mathbf{z}_n\|_1 \qquad (4)$$
$$\text{subject to} \quad y_{i,l}t_i = \mathbf{m}_i'\mathbf{z}_l \qquad \begin{matrix} l = 1, \dots, L \\ i = 1, \dots, M \end{matrix}$$
$$\sum_{n=1}^{M} t_n = c$$

for an arbitrary constant $c > 0$. The actual gains can be set as $d_i^* = \frac{1}{\tau_i^*}$ after the optimization. Note that even though the minimized objective function is equivalent to the alternating non-linear optimization, the problems with local minimums are now eliminated thanks to the convexity of the formulation.

We can make several observations regarding the optimization in (4):

1) The constraint $\sum_n t_n = c$ ensures that the trivial solution ($\tau_i = 0, \ \mathbf{x}_l = 0, \ \forall i, l$) is excluded from the solution set.
2) The constraint $\sum_n t_n = c$ also excludes the solutions where the sum of the gains are zero. When dealing with signed real or complex valued gains, this may result in excluding the actual solution in rare cases where the sought out gains actually sum up to zero. However, the probability of encountering this phenomena in real applications is often infinitesimally small. For the applications in which this possibility is higher, an alternative approach to deal with this case is discussed in Section III.
3) The measurement constraints are satisfied up to a global scale factor (and phase shift for complex signals), therefore the constant c can be set arbitrarily. Unfortunately, the global scale (and phase) factor cannot be determined with the given optimization approach, although this is often not an issue in practical systems.
4) The successful recovery of the gains and the signals require availability of more than one input signal ($L > 1$). Although this may seem like a restriction, acquiring data from multiple sources is often straightforward in many application fields.

III. Experimental Results

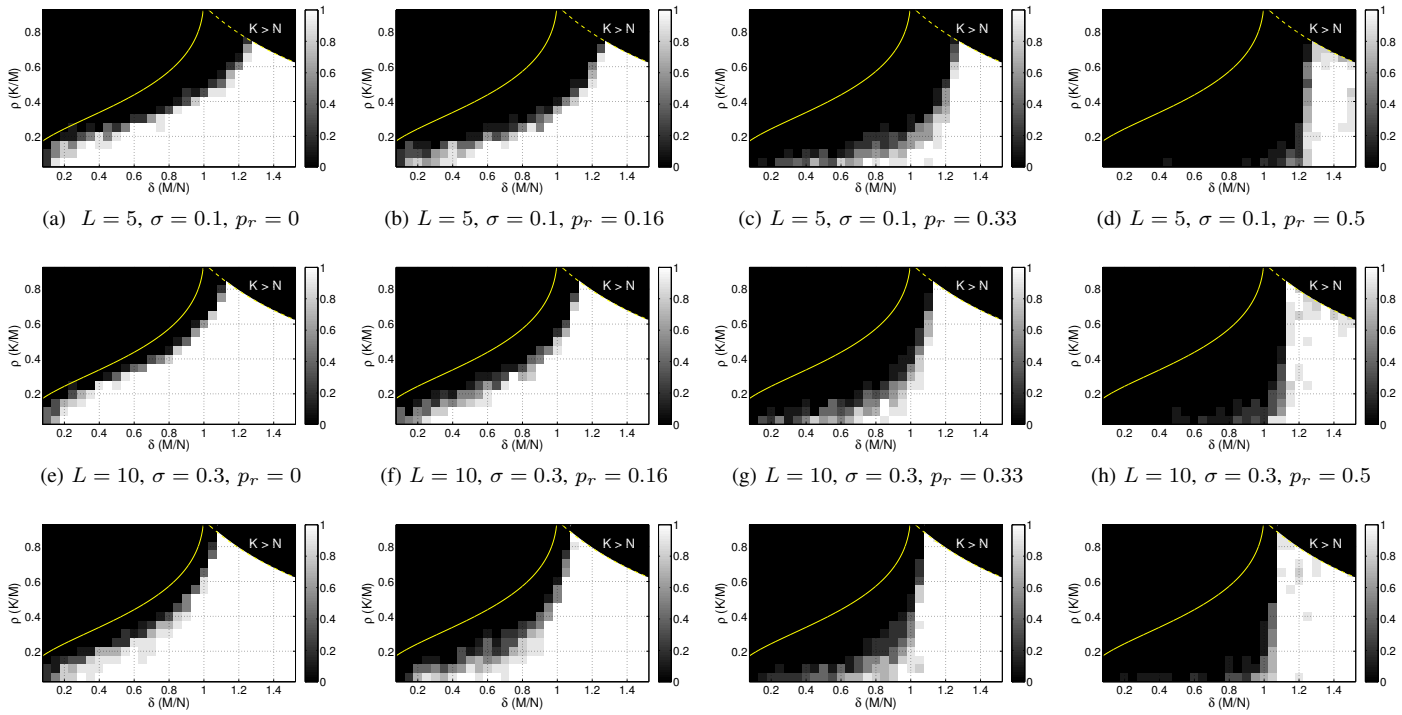
In order to test the performance of the proposed algorithm, phase transition curves as in the compressed sensing recovery are plotted for a signal size $N = 100$ with the measurement vectors, $\mathbf{m}_i$, and all the non zero entries in the input signals, $\mathbf{x}_l$, randomly generated from an i.i.d. normal distribution. The positions of the non-zero coefficients of the input signals, $\mathbf{x}_l$, are chosen uniformly at random in $\{1, \dots, N\}$. The magnitude of the gains were generated using $|d_i| \sim \exp(\mathcal{N}(0, \sigma^2))$, where σ is the parameter governing the amplitude of decalibration. For real valued experiments, the sign of the gains are randomly assigned such that the probability, p_r, of setting a negative gain is adjusted to be $p_r \in \{0, 0.16, 0.33, 0.5\}$. Similarly for complex valued gains, the phase of the gains are chosen uniformly at random from the range $[0, \ 2\pi p_c)$ where $p_c \in \{0, 0.33, 0.66, 1\}$. Note that the parameters p_r and p_c determine the scale of ambiguity in the signs and phases where maximum possible ambiguity is observed when $p_r = 0.5$ and $p_c = 1$ respectively.

The signals (and the gains) are recovered for different amount of decalibration amplitude ($\sigma = 0.1, \ 0.3, \ 1$) with sufficiently high number of input signals ($L = 5, \ 10, \ 20$ respectively). The proposed optimization in (4) is performed using an ADMM [9] based algorithm. The perfect reconstruction criteria is selected as $\frac{1}{L}\sum_l \mu(\mathbf{x}_l, \mathbf{x}_l^*) > 0.9999$, where the absolute correlation factor $\mu(\cdot, \cdot)$ is defined as

$$\mu(\mathbf{x}_1, \mathbf{x}_2) \triangleq \frac{|\mathbf{x}_1'\mathbf{x}_2|}{\|\mathbf{x}_1\|_2\|\mathbf{x}_2\|_2} \qquad (5)$$

so that the global phase and scale difference between the source and recovered signals is ignored.

The probability of recovery (computed through 10 independent simulations for each set of parameters) of the proposed method with respect to $\delta \triangleq M/N$ and $\rho \triangleq K/M$ are shown in Figures 1 and 2 for real valued and complex valued systems respectively. The first thing to notice from the results is that the performance for $p_r = 0$ (or $p_c = 0$) is consistent with the results presented in [7] as expected. The effect of increasing sign (or phase) ambiguity can be observed in the results as p_r (or p_c) increases. Although the performance is acceptable for p_r as high as 0.33 (p_c up to 0.66), there is a significant degradation when dealing very high sign (or phase) ambiguity such that signal recovery is impossible regardless of the sparsity, unless the measurement system is overcomplete ($M > N$). This phenomena can best be observed in the last row of complex-valued results, Figures 2(m)-2(p), where the number of input signals is very large ($L = 50$) with respect to



(a) $L = 5$, $\sigma = 0.1$, $p_r = 0$ (b) $L = 5$, $\sigma = 0.1$, $p_r = 0.16$ (c) $L = 5$, $\sigma = 0.1$, $p_r = 0.33$ (d) $L = 5$, $\sigma = 0.1$, $p_r = 0.5$

(e) $L = 10$, $\sigma = 0.3$, $p_r = 0$ (f) $L = 10$, $\sigma = 0.3$, $p_r = 0.16$ (g) $L = 10$, $\sigma = 0.3$, $p_r = 0.33$ (h) $L = 10$, $\sigma = 0.3$, $p_r = 0.5$

(i) $L = 20$, $\sigma = 1$, $p_r = 0$ (j) $L = 20$, $\sigma = 1$, $p_r = 0.16$ (k) $L = 20$, $\sigma = 1$, $p_r = 0.33$ (l) $L = 20$, $\sigma = 1$, $p_r = 0.5$

Fig. 1: The probability of perfect recovery in the real valued system for $N = 100$ with respect to $\delta \triangleq M/N$ and $\rho \triangleq K/M$. The solid yellow line indicates the Donoho-Tanner phase transition curve for fully calibrated compressed sensing recovery [8]. The dashed yellow line indicates the boundary to the region where $K > N$. Each row of figures display the change in recovery performance with increasing sign ambiguity from left to right for a fixed set of L and σ.

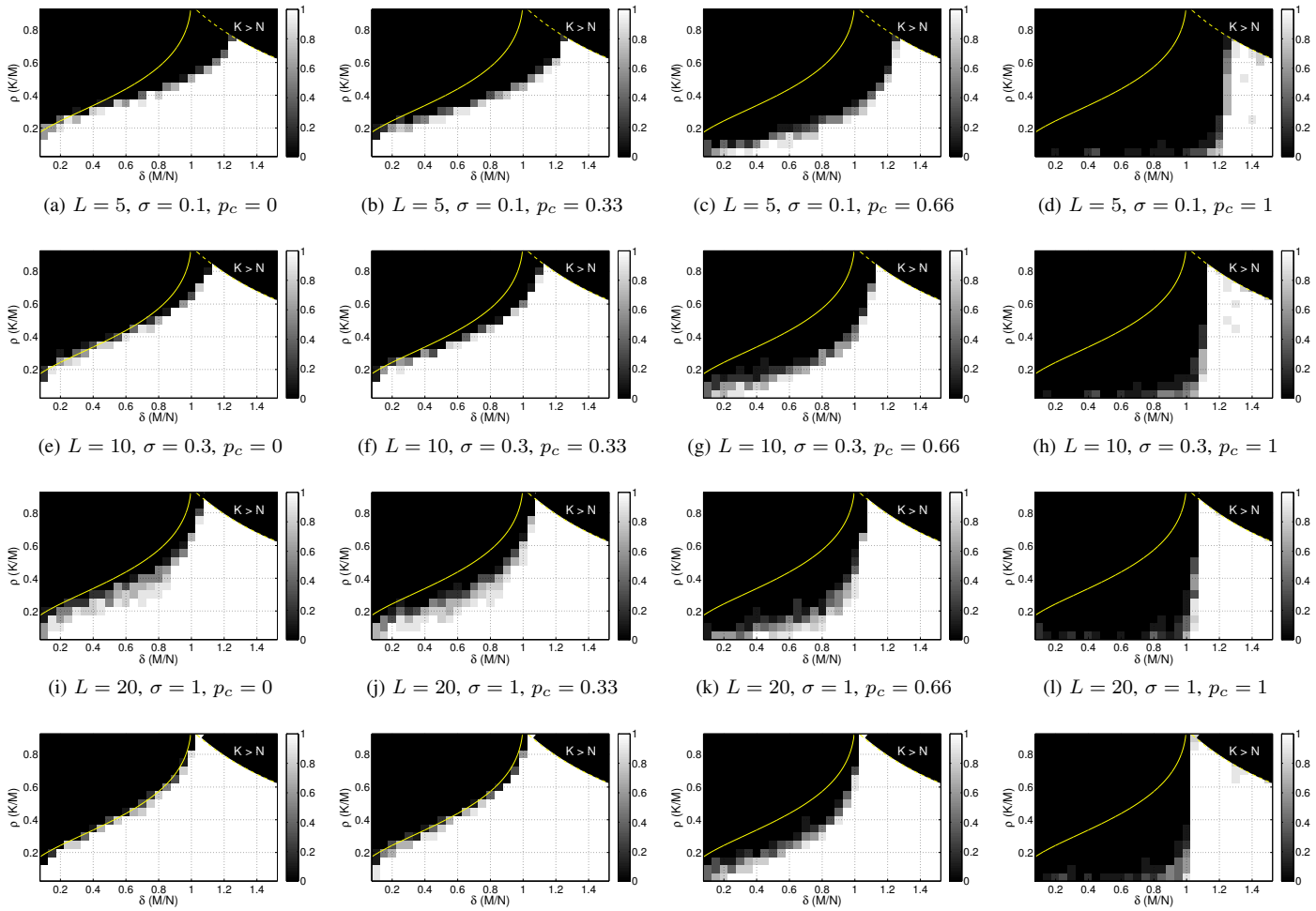
the variance in the gain magnitudes ($\sigma = 0.1$). The degradation in the results can be attributed to the significant increase in the contamination of the information in the measurements as the sign or phase ambiguity increases. Therefore recovery becomes possible only when there are sufficient number of measurements to overcome the high distortion. For the maximally ambiguous case ($p_r = 0.5$, $p_c = 1$), this is only possible for $M > N$. Even though this is a drawback of the presented approach, it should be noted that in many practical systems the sign (or phase) ambiguity is often not as severe as fully random, but within a limited range. Therefore the presented algorithm can still be applied in various scenarios.

As an alternative to the proposed method in this paper, a phase calibration algorithm (in which gain magnitudes are assumed to be known) that can recover the sparse signals along with the unknown phases distributed within the entire $[0,\ 2\pi)$ range has been presented in [10], [11]. This approach for phase calibration can be combined with the proposed method in this paper in order to recover signed real-valued or complex-valued gains with maximum sign and phase ambiguity. It is also possible to use this combined approach for signal recovery in applications where the sum of the gains are likely to be zero.

IV. Conclusions and Future Work

In this paper, we have investigated the problem of estimating the unknown gains at each measurement sensor along with sparse input signals in a compressed sensing measurement system. We have extended the use of convex recovery strategy suggested for positive real gains to the more general cases of signed real-valued and complex-valued gains, and demonstrated the change of recovery performance with increasing sign and phase ambiguity.

The performance of the proposed algorithm is shown to be approaching to that of the unperturbed compressed sensing recovery when there are sufficient number sparse input signals unless the distribution of the sign changes or the phase shifts are maximally varying among the sensors. This drawback of the proposed algorithm can still be ignored for many application fields in which the ambiguity in the sign changes or the phase shifts at the sensors are within a limited range. For other applications, it is possible to combine the proposed method with other approaches employed for phase calibration to improve the recovery performance which is considered as a future work. The theoretical justification of the limitation of the proposed method for maximum sign and phase ambiguity is also a work in progress.



(a) $L = 5$, $\sigma = 0.1$, $p_c = 0$ (b) $L = 5$, $\sigma = 0.1$, $p_c = 0.33$ (c) $L = 5$, $\sigma = 0.1$, $p_c = 0.66$ (d) $L = 5$, $\sigma = 0.1$, $p_c = 1$

(e) $L = 10$, $\sigma = 0.3$, $p_c = 0$ (f) $L = 10$, $\sigma = 0.3$, $p_c = 0.33$ (g) $L = 10$, $\sigma = 0.3$, $p_c = 0.66$ (h) $L = 10$, $\sigma = 0.3$, $p_c = 1$

(i) $L = 20$, $\sigma = 1$, $p_c = 0$ (j) $L = 20$, $\sigma = 1$, $p_c = 0.33$ (k) $L = 20$, $\sigma = 1$, $p_c = 0.66$ (l) $L = 20$, $\sigma = 1$, $p_c = 1$

(m) $L = 50$, $\sigma = 0.1$, $p_c = 0$ (n) $L = 50$, $\sigma = 0.1$, $p_c = 0.33$ (o) $L = 50$, $\sigma = 0.1$, $p_c = 0.66$ (p) $L = 50$, $\sigma = 0.1$, $p_c = 1$

Fig. 2: The probability of perfect recovery in the complex valued system for $N = 100$ with respect to $\delta \triangleq M/N$ and $\rho \triangleq K/M$. The solid yellow line indicates the Donoho-Tanner phase transition curve for fully calibrated compressed sensing recovery [8]. The dashed yellow line indicates the boundary to the region where $K > N$. Each row of figures display the change in recovery performance with increasing phase ambiguity from left to right for a fixed set of L and σ. The last row, (m)-(p) shows the performance limit for very high L.

REFERENCES

[1] David L. Donoho, "Compressed Sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289 – 1306, 2006.

[2] Zai Yang, Cishen Zhang, and Lihua Xie, "Robustly stable signal recovery in compressed sensing with structured matrix perturbation," *Signal Processing, IEEE Transactions on*, vol. 60, no. 9, pp. 4658 – 4671, sept. 2012.

[3] Boon Chong Ng and Chong Meng Samson See, "Sensor-array calibration using a maximum-likelihood approach," *Antennas and Propagation, IEEE Transactions on*, vol. 44, no. 6, pp. 827 –835, jun 1996.

[4] R. Mignot, L. Daudet, and F. Ollivier, "Compressed sensing for acoustic response reconstruction: Interpolation of the early part," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011 IEEE Workshop on*, oct. 2011, pp. 225 –228.

[5] Emmanuel J. Cands, Justin K. Romberg, and Terence Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on Pure and Applied Mathematics*, vol. 59, no. 8, pp. 1207–1223, 2006.

[6] M.A. Herman and T. Strohmer, "General deviants: An analysis of perturbations in compressed sensing," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 4, no. 2, pp. 342 –349, april 2010.

[7] Rémi Gribonval, Gilles Chardon, and Laurent Daudet, "Blind calibration for compressed sensing by convex optimization," in *Acoustics Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, 2012, pp. 2713–2716.

[8] David L. Donoho and Jared Tanner, "Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing.," *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, vol. 367, no. 1906, pp. 4273–93, Nov. 2009.

[9] Stephen Boyd, Neal Parikh, Eric Chu, B Peleato, and Jonathan Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[10] Cagdas Bilen, Gilles Puy, Rémi Gribonval, and Laurent Daudet, "Blind Sensor Calibration in Sparse Recovery," in *International Biomedical and Astronomical Signal Processing (BASP) Frontiers Workshop*, Villars-sur-Ollon, Switzerland, Jan. 2013.

[11] Cagdas Bilen, Gilles Puy, Rémi Gribonval, and Laurent Daudet, "Blind Phase Calibration in Sparse Recovery," in *European Signal Processing Conference (EUSIPCO) (submitted)*, 2013.

Sampling by blocks of measurements in compressed sensing

Jérémie Bigot [1], Claire Boyer [2] and Pierre Weiss [3]

[1]*Institut Supérieur de l'Aéronautique et de l'Espace (ISAE), Département Mathématiques (DMIA), Toulouse, France*
[2]*Institut de Mathématiques de Toulouse, Université Paul Sabatier, Toulouse, France*
[3]*Institut des Technologies Avancées du Vivant (USR 3505), Toulouse, France*

***Abstract*—Various acquisition devices impose sampling blocks of measurements. A typical example is parallel magnetic resonance imaging (MRI) where several radio-frequency coils simultaneously acquire a set of Fourier modulated coefficients. We study a new random sampling approach that consists in selecting a set of blocks that are predefined by the application of interest. We provide theoretical results on the number of blocks that are required for exact sparse signal reconstruction. We finish by illustrating these results on various examples, and discuss their connection to the literature on CS.**

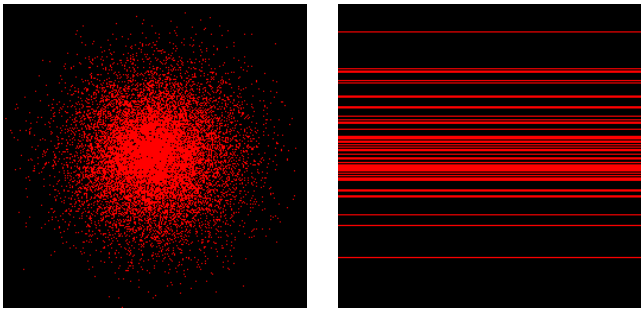
Key-words : compressed sensing, blocks of measurements, sampling continuous trajectory, exact recovery, ℓ^1 minimization.

I. Introduction

In many applications, the sampling strategy imposes to acquire data in the form of blocks of measurements (see Fig. 1(b) for block-structured sampling), instead of isolated measurements (see Fig. 1(a)). For instance, in medical echography, images are sampled along lines in the space domain, while, in magnetic resonance imaging (MRI), acquiring data along radial lines or spiral trajectories is a popular sampling strategy. In compressed sensing (CS), various theoretical conditions have been proposed to guarantee the exact reconstruction of a sparse vector from a small number of isolated measurements that are randomly drawn, see [1], [2], [3], and [4] for a detailed review of the most recent results on this topic.

In a noise-free setting, the focus of the present paper is on studying the problem of exact recovery of a sparse signal in the case where the sampling strategy consists in randomly choosing blocks of measurements. Each block corresponds to a set of rows of an orthogonal sensing matrix. Our approach is more flexible than the angle chosen in [5], while we assert theoretical guarantees on the exact reconstruction of sparse signals from blocks of measurements. Moreover, we assume that physical acquisition devices impose block-structured measurements, whereas in [6], or in [7] the authors consider a block-sparse signal.

[1]jeremie.bigot@isae.fr
[2]claire.boyer@math.univ-toulouse.fr
[3]pierre.armand.weiss@gmail.com



(a) (b)

Fig. 1. **An example of two sampling schemes in the 2D Fourier domain with an undersampling factor** $R = 4$ (a): Isolated points and radial distribution. (b): Corresponding acquisition in the case of block measurements that consist of lines in the 2D Fourier domain.

In this paper, we deal with the case where the blocks are predefined. We give some conditions on the choice of the drawing probability of the blocks and on the number of measurements that are sufficient to obtain an exact recovery by ℓ^1 minimization. We finish by illustrating these results on various examples, and we discuss their connection to the literature on CS.

II. Problem setting

A. Notation

We consider an orthogonal matrix $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ which denotes the full sensing matrix. Matrix $\boldsymbol{A}$ is given a block structure, as follows: $\boldsymbol{A} = \begin{bmatrix} \boldsymbol{B}_1 \\ \vdots \\ \boldsymbol{B}_M \end{bmatrix}$, where the blocks $(\boldsymbol{B}_j)_{1\leq j\leq M}$ are non-overlapping and such that $\boldsymbol{B}_j \in \mathbb{C}^{n_j\times n}$ with $\sum_{j=1}^M n_j = n$. We set $\|\boldsymbol{A}\|_\infty = \max\limits_{1\leq i,j\leq n} |\boldsymbol{A}_{ij}|$.

Let $(\pi_j)_{1\leq j\leq M}$ be positive weights with $\sum_{j=1}^M \pi_j = 1$, and let Π be a discrete probability distribution on the set of integers $\{1,\dots,M\}$, associated to these weights. Throughout $(J_k)_{1\leq k\leq m}$ denotes a sequence of i.i.d. discrete random variables taking their value in $\{1,\dots,M\}$ with distribution Π.

Let $S \subset \{1,\dots,n\}$ be a set of cardinality s. For a matrix $\boldsymbol{M} \in \mathbb{C}^{m\times n}$, we define

$$\boldsymbol{M}^S = (\boldsymbol{M}_{ij})_{1\leq i\leq m, j\in S}\,.$$

B. The sampling strategy

In this paper, we consider the following sampling strategy. We randomly select m blocks among $(\boldsymbol{B}_j)_{1\leq j\leq M}$, according to the discrete probability distribution Π, which leads to consider the sequence of i.i.d. random blocks $(\boldsymbol{X}_k)_{1\leq k\leq m}$ defined by

$$\boldsymbol{X}_k = \frac{1}{\sqrt{\pi_{J_k}}}\boldsymbol{B}_{J_k},\ k = 1\dots m \tag{1}$$

We consider the following random sampling matrix

$$\widetilde{\boldsymbol{A}_m} = \frac{1}{\sqrt{m}}\begin{bmatrix}\boldsymbol{X}_1\\ \vdots \\ \boldsymbol{X}_m\end{bmatrix}. \tag{2}$$

It satisfies $\mathbb{E}\left[\widetilde{\boldsymbol{A}_m}^*\widetilde{\boldsymbol{A}_m}\right] = \mathrm{Id}_n$ by construction.

C. Minimization problem

Let $\boldsymbol{y} = \widetilde{\boldsymbol{A}_m}\boldsymbol{x}$ denote a set of $q = \sum_{k=1}^m n_{J_k}$ linear measurements of a signal $\boldsymbol{x}$. To reconstruct $\boldsymbol{x}$, the following standard ℓ_1-minimization problem is solved:

$$\min_{\boldsymbol{z}\in\mathbb{C}^n} \|\boldsymbol{z}\|_1 \quad \text{subject to} \quad \widetilde{\boldsymbol{A}_m}\boldsymbol{z} = \boldsymbol{y}. \tag{3}$$

III. A non-uniform recovery result

Let us first introduce a new quantity of interest that will be shown to be of primary importance to obtain exact recovery.

Definition III.1

For $S\subset\{1,..,n\}$ we denote by ρ_k^S for $1\leq k\leq M$ any set of positive reals that satisfies

$$\rho_k^S \geq \left\|\left(\boldsymbol{B}_k^S\right)^*\boldsymbol{B}_k^S\right\|_2$$

where $\|\boldsymbol{C}\|_2$ is the spectral norm of a matrix $\boldsymbol{C}$.

The following theorem is the main result of the paper. It gives a set of sufficient conditions for exact recovery of $\boldsymbol{x}$ with large probability.

Theorem III.2

Let $S\subset\{1\dots n\}$, be a set of cardinality $\sharp\{S\} = s$ and let $\boldsymbol{\epsilon} = (\epsilon_\ell)_{\ell\in S}\in\mathbb{C}^s$ be a sequence of independent random variables that are uniformly distributed on $\{-1;1\}$ (or on the torus $\{z\in\mathbb{C}\,,\,|z|=1\}$).
Let $\boldsymbol{x}$ be a sparse vector with support S and $\mathrm{sgn}(\boldsymbol{x}^S) = \boldsymbol{\epsilon}$.
Let $\widetilde{\boldsymbol{A}_m}$ be the sampling matrix built as above (see (2)).
Assume that

$$\begin{cases} m \geq Cs\ln^2\left(\dfrac{2^{3/4}n}{\varepsilon}\right)\max\limits_{1\leq k\leq M}\dfrac{\|\boldsymbol{B}_k^*\boldsymbol{B}_k\|_\infty}{\pi_k} & (4)\\ m \geq C\ln\left(\dfrac{2^{3/4}s}{\varepsilon}\right)\max\limits_{1\leq k\leq M}\dfrac{\rho_k^S}{\pi_k} & (5)\end{cases}$$

with $C = 256\kappa^2$, $C' = 32\kappa^2$ and $\kappa^2 = \left(\frac{\sqrt{17}+1}{4}\right)^2$.
Then with probability at least $1-\varepsilon$ the vector $\boldsymbol{x}$ is the unique solution to the ℓ_1-minimization problem (3).

The proof of Theorem III.2 is too long to be written here. It will appear in a forthcoming preprint. The approach is inspired by the results in [4]. To derive Theorem III.2, we had to extend probabilistic tools such as symmetrization and Rudelson's lemma [4] from the vectorial case to the matricial one.

Remark : We can notice that the bounding above of $\left\|\left(\boldsymbol{B}_k^S\right)^*\boldsymbol{B}_k^S\right\|_2$ by ρ_k^S should not be too coarse, at the risk of making the required number of measurements too large.

IV. Discussion and examples

Conditions (4) and (5) may lead to a different optimal drawing probabilty Π^*, in the sense that they can be used to minimize a lower bound on the number m of block measurements. Indeed

- if the right-hand side (rhs) of Inequality (4) is greater than the rhs of Inequality (5), an optimal drawing probability Π^* is defined as follows: $\forall k\in\{1,\dots,M\}$
$$\pi_k^* = \frac{\|\boldsymbol{B}_k^*\boldsymbol{B}_k\|_\infty}{\sum_{\ell=1}^M\|\boldsymbol{B}_\ell^*\boldsymbol{B}_\ell\|_\infty}.$$
- On the contrary, if the rhs of Inequality (5) prevails, then an optimal drawing probability Π^* turns to be: $\forall k\in\{1,\dots,M\}$
$$\pi_k^* = \frac{\rho_k^S}{\sum_{\ell=1}^M\rho_\ell^S}.$$

Let us illustrate Theorem III.2 on practical examples.

A. One row blocks - the case of isolated measurements

First, let us show that our result matches the standard setting where blocks are made of only one row. This is the case considered e.g. by [2], [4]. Thus $M = n$,

$$\boldsymbol{A} = \begin{pmatrix}\boldsymbol{B}_1\\ \vdots\\ \boldsymbol{B}_M\end{pmatrix} = \begin{pmatrix}\boldsymbol{a}_1^*\\ \vdots\\ \boldsymbol{a}_n^*\end{pmatrix}$$

where $\boldsymbol{a}_1,\dots,\boldsymbol{a}_n$ are vectors of $\mathbb{C}^n$, and $\forall k\in\{1,\cdots,M\}$, $\boldsymbol{B}_k = \boldsymbol{a}_k^*$. We can set

$$\rho_k^S = s\|\boldsymbol{a}_k\|_\infty^2$$

with $\sharp S = s$. Then, the required number of measurements will be minimized for the following drawing probability: $\forall k\in\{1,\dots,M\}$

$$\pi_k^* = \frac{\|\boldsymbol{a}_k\|_\infty^2}{\sum_{\ell=1}^n\|\boldsymbol{a}_\ell\|_\infty^2}.$$

According to Theorem III.2 the number of isolated measurements sufficient to obtain perfect reconstruction with high probability is

$$m \geq Cs\ln^2\left(\frac{2^{3/4}3n}{\varepsilon}\right)\sum_{\ell=1}^n\|\boldsymbol{a}_\ell\|_\infty^2. \tag{6}$$

This condition is consistent with [4] for the non-uniform recovery, up to a constant. This additional factor is not too serious, since Theorem III.2 should be mainly considered as a

guide to construct sampling patterns and not as a requirement for perfect recovery. Surprisingly, a better drawing probability distribution reducing the required number of measurements is not the uniform one, as commonly used in [8], [4], but the one depending on the ℓ_∞-norm of the considered row.

B. Block diagonal case

Let us assume that $\boldsymbol{A}$ is orthogonal, and $\boldsymbol{A}$ can be written as

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{B}_1 \\ \vdots \\ \boldsymbol{B}_M \end{pmatrix} = \begin{pmatrix} \boldsymbol{D}_1 & 0 & 0 & \dots & 0 \\ 0 & \boldsymbol{D}_2 & 0 & \dots & 0 \\ & & & & \\ 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & 0 & \dots & \boldsymbol{D}_M \end{pmatrix}.$$

Then $\left\|\boldsymbol{B}_k^* \boldsymbol{B}_k\right\|_\infty = 1$ and ρ_k^S can be taken equal to 1 for all $k \in \{1, \dots, M\}$, since $\boldsymbol{D}_k$ is orthogonal. Thus, the block diagonal case corresponds to a uniform bound for ρ_k^S. Therefore, both Inequalities (4) and (5) entail a uniform drawing probability as an optimal choice. Here, we see that no matter how large the block is, an optimal drawing probability Π^* is the uniform one: $\forall k \in \{1, \dots, M\}$,

$$\pi_k^* = \frac{1}{M}.$$

Moreover, with such a choice for Π^*, and by Theorem III.2 the number of block measurements sufficient to obtain perfect reconstruction with high probability is

$$m \geq Cs \ln^2\left(\frac{2^{3/4}3n}{\varepsilon}\right) M. \tag{7}$$

C. 2D Fourier matrix

We now turn to a more realistic setting where signals are sparse in the Dirac basis and blocks of frequencies are probed in the 2D Fourier domain. We consider blocks that consist of discrete lines in the 2D Fourier space as in Fig 1(b). This scenario is close to what can be encountered in MRI, echography or some tomographic devices.

We assume that $\sqrt{n} \in \mathbb{N}$ and that $\boldsymbol{A}$ is the 2D Fourier matrix applicable on $\sqrt{n} \times \sqrt{n}$ images. For all $p_1 \in \{1, \dots, \sqrt{n}\}$,

$$\boldsymbol{B}_{p_1} = \left[\frac{1}{\sqrt{n}} \exp\left(2i\pi\left(\frac{p_1\ell_1 + p_2\ell_2}{\sqrt{n}}\right)\right)\right]_{(p_1,p_2)(\ell_1,\ell_2)} \tag{8}$$

with $1 \leq p_2 \leq \sqrt{n}, 1 \leq \ell_1, \ell_2 \leq \sqrt{n}$. Let $S \subset \{1, \dots, \sqrt{n}\} \times \{1, \dots, \sqrt{n}\}$ denote the support of $\boldsymbol{x}$, with $\sharp S = s$. We can write $S = \{(S_{1,1}, S_{1,2}), (S_{2,1}, S_{2,2}), \dots, (S_{s,1}, S_{s,2})\}$, and we call $S_1 = \{S_{1,1}, S_{2,1}, \dots, S_{s,1}\}$ and $S_2 = \{S_{1,2}, S_{2,2}, \dots, S_{s,2}\}$. We can rewrite $\boldsymbol{B}_{p_1}^S$ as

$$\underbrace{\left(\frac{1}{n^{1/4}} e^{-2i\pi p_2 \frac{\ell_2}{\sqrt{n}}}\right)_{\substack{1 \leq p_2 \leq \sqrt{n} \\ \ell_2 \in S_2}}}_{\substack{\boldsymbol{M}^S \\ \sqrt{n} \times s \text{ matrix}}} \underbrace{\begin{pmatrix} \ddots & 0 & 0 \\ 0 & \frac{1}{n^{1/4}} e^{-2i\pi p_1 \frac{\ell_1}{\sqrt{n}}} & 0 \\ 0 & 0 & \ddots \end{pmatrix}_{\ell_1 \in S_1}}_{\substack{\boldsymbol{D}_{p_1} \\ s \times s \text{ diagonal matrix}}}$$

$$\begin{aligned} \text{so} \quad \left\|\left(\boldsymbol{B}_{p_1}^S\right)^* \boldsymbol{B}_{p_1}^S\right\|_2 &= \left\|\boldsymbol{D}_{p_1}^* \boldsymbol{M}^{S*} \boldsymbol{M}^S \boldsymbol{D}_{p_1}\right\|_2 \\ &\leq \left\|\boldsymbol{D}_{p_1}^*\right\|_2 \left\|\boldsymbol{M}^{S*} \boldsymbol{M}^S\right\|_2 \left\|\boldsymbol{D}_{p_1}\right\|_2 \\ &\leq \frac{1}{n^{1/2}} \left\|\boldsymbol{M}^{S*} \boldsymbol{M}^S\right\|_2. \end{aligned}$$

In fact, we can see $\boldsymbol{M}^S$ as 1D Fourier matrix $\boldsymbol{M}$, from which we select columns, eventually repeated, the indexes of which are in S_2. Now we have to evaluate the quantity $\left\|\boldsymbol{M}^{S*} \boldsymbol{M}^S\right\|_2$. To do so, let us denote by $(s_j)_{j=1..\sqrt{n}}$ the number of repetitions of the j-th element of $\{1, \dots, \sqrt{n}\}$ in S_2. We have that $\sum_{j=1}^{\sqrt{n}} s_j = s$, and $0 \leq s_j \leq \sqrt{n}$, $\forall j \in \{1, \dots, \sqrt{n}\}$.

Simple calculation leads to the following upper bound:

$$\left\|\boldsymbol{M}^{S*} \boldsymbol{M}^S\right\|_2 \leq \frac{\max_{j=1,\dots,\sqrt{n}} s_j}{\sqrt{n}} \leq \frac{\min(s, \sqrt{n})}{\sqrt{n}},$$

which leads to the choice

$$\rho_k^S = \frac{\min(s, \sqrt{n})}{n},\ k = 1, \dots, M.$$

By definition of the 2D Fourier matrix of size $n \times n$, $\left\|\boldsymbol{B}_k^* \boldsymbol{B}_k\right\|_\infty = 1/\sqrt{n}$, for all $k \in \{1, \dots, \sqrt{n}\}$. Then, the choice of the optimal drawing probability is given by $\forall k \in \{1, \dots, \sqrt{n}\}$

$$\pi_k^* = \frac{1}{\sqrt{n}}.$$

We deduce that the number of block measurements sufficient to ensure exact recovery with high probability is

$$m \geq Cs \ln^2\left(\frac{2^{3/4}n}{\varepsilon}\right).$$

D. Wavelet Transform

Here, we consider that $\boldsymbol{A}$ is a dyadic wavelet transform matrix, with $n = 2^\alpha$, $\alpha \in \mathbb{N}$. To each resolution level $k \in \{0, \dots, \alpha\}$ ($k = 0$, corresponding to the scaling function), we associate the block $\boldsymbol{B}_k$

$$\boldsymbol{B}_k = \left(\Psi_{k,j}(\ell)\right)_{\substack{j = 1 \dots n_k \\ 1 \leq \ell \leq n}}, \tag{9}$$

where $\Psi_{k,j}$ is the discrete wavelet at scale k and location parameter j, ℓ is the time variable and n_k is the number of wavelets (or scaling function) at scale k defined as follows

$$n_k = \begin{cases} 1 & \text{if} \quad k = 0 \\ 2^{k-1} & \text{if} \quad k \geq 1. \end{cases}$$

Although this example is not realistic in practice, it provides an interesting illustration of Theorem III.2. Let S be a set of indexes of cardinality s. Then $\boldsymbol{B}_k^S$ can be defined by restricting ℓ to belong to S, i.e.

$$\boldsymbol{B}_k^S = \left(\Psi_{k,j}(\ell)\right)_{\substack{j = 1 \dots n_k \\ \ell \in S}}.$$

As a consequence, $\left(\boldsymbol{B}_k^S\right)^* \boldsymbol{B}_k^S$ is an $s \times s$ matrix, and

$$\left[\left(\boldsymbol{B}_k^S\right)^* \boldsymbol{B}_k^S\right]_{(\ell,\ell')\in S^2} = \left(\sum_{j=1}^{n_k} \boldsymbol{\psi}_{k,j}(\ell)\, \boldsymbol{\psi}_{k,j}(\ell')\right)_{(\ell,\ell')\in S^2}. \tag{10}$$

By the results in [9], for wavelets with compact support, such as Haar's wavelets, we obtain that $\left\|\left(\boldsymbol{B}_k^S\right)^* \boldsymbol{B}_k^S\right\|_2 \leq \left\|\left(\boldsymbol{B}_k^S\right)^* \boldsymbol{B}_k^S\right\|_\infty s \leq \frac{n_k}{n} s$. Hence, one can take $\rho_k^S = \frac{n_k}{n} s$, and the required number of measurements satisfies the bounds

$$\begin{cases} m \geq C s \ln^2\left(\dfrac{2^{3/4} n}{\varepsilon}\right) \dfrac{1}{n} \max\limits_{1\leq k\leq K} \dfrac{n_k}{\pi_k} & (11) \\ m \geq C' s \ln\left(\dfrac{2^{3/4} s}{\varepsilon}\right) \dfrac{1}{n} \max\limits_{1\leq k\leq K} \dfrac{n_k}{\pi_k} & (12) \end{cases}$$

that m is still proportional to s. If (12) is the strongest condition on m, then an optimal choice for the drawing probability Π^* is

$$\pi_k^* = \frac{n_k}{\sum_{q=0}^{\alpha} n_q} \quad k \in \{1, \ldots, K\}.$$

In this setting, the drawing probability is growing with the resolution level k and it is proportional to the block size.

V. Conclusion

In this paper, we have introduced some theoretical tools for the study of the exact recovery of sparse signals from blocks of measurements selected randomly from an orthogonal sensing matrix. We introduced the new quantities ρ_k^S and $\|\boldsymbol{B}_k^* \boldsymbol{B}_k\|_\infty$. They play a central role to derive optimal sampling strategies and to assess the number of block measurements that is necessary to exactly reconstruct sparse signals by ℓ_1-minimization. We plan to calibrate their for orthogonal matrices that appear in applications such as the product of a discrete Fourier transform with a wavelet transform. The extension of this work to overlapping blocks, as presented in Figure 2, offers much more versatility in the sampling patterns.

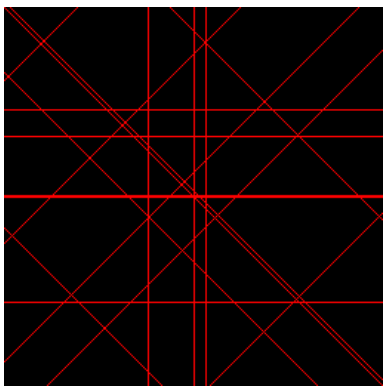


Fig. 2. An example of overlapping blocks of measurements in the 2D Fourier domain.

References

[1] M. Rudelson and R. Vershynin, "On sparse reconstruction from fourier and gaussian measurements," *Communications on Pure and Applied Mathematics*, vol. 61, no. 8, pp. 1025–1045, 2008.

[2] E. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *Information Theory, IEEE Transactions on*, vol. 52, no. 2, pp. 489–509, 2006.

[3] E. Candes and Y. Plan, "A probabilistic and ripless theory of compressed sensing," *Information Theory, IEEE Transactions on*, vol. 57, no. 11, pp. 7235–7254, 2011.

[4] H. Rauhut, "Compressive sensing and structured random matrices," *Theoretical foundations and numerical methods for sparse recovery*, vol. 9, pp. 1–92, 2010.

[5] L. Gan, "Block compressed sensing of natural images," in *Digital Signal Processing, 2007 15th International Conference on*. IEEE, 2007, pp. 403–406.

[6] Y. C. Eldar, P. Kuppinger, and H. Bolcskei, "Block-sparse signals: Uncertainty relations and efficient recovery," *Signal Processing, IEEE Transactions on*, vol. 58, no. 6, pp. 3042–3054, 2010.

[7] N. Rao, B. Recht, and R. Nowak, "Tight measurement bounds for exact recovery of structured sparse signals," *arXiv preprint arXiv:1106.4355*, 2011.

[8] E. Candes and J. Romberg, "Sparsity and incoherence in compressive sampling," *Inverse problems*, vol. 23, no. 3, p. 969, 2007.

[9] S. Mallat, *A wavelet tour of signal processing*. Academic press, 1999.

Travelling salesman-based variable density sampling

Nicolas Chauffert, Philippe Ciuciu
CEA, NeuroSpin center,
INRIA Saclay, PARIETAL Team
145, F-91191 Gif-sur-Yvette, France
Email: firstname.lastname@cea.fr

Jonas Kahn
Laboratoire Painlevé, UMR 8524
Université de Lille 1, CNRS
Cité Scientifique Bât. M2
59655 Villeneuve d'Ascq Cedex, France
Email: jonas.kahn@math.univ-lille1.fr

Pierre Weiss
ITAV, USR 3505
PRIMO Team,
Université de Toulouse, CNRS
Toulouse, France
Email: pierre.weiss@itav-recherche.fr

***Abstract*—Compressed sensing theory indicates that selecting a few measurements independently at random is a near optimal strategy to sense sparse or compressible signals. This is infeasible in practice for many acquisition devices that acquire samples along *continuous* trajectories. Examples include magnetic resonance imaging (MRI), radio-interferometry, mobile-robot sampling, ... In this paper, we propose to generate continuous sampling trajectories by drawing a small set of measurements independently and joining them using a travelling salesman problem solver. Our contribution lies in the theoretical derivation of the appropriate probability density of the initial drawings. Preliminary simulation results show that this strategy is as efficient as independent drawings while being implementable on real acquisition systems.**

I. Introduction

Compressed sensing theory provides guarantees on the reconstruction quality of sparse and compressible signals $x \in \mathbb{R}^n$ from a limited number of linear measurements $(\langle a_k, x\rangle)_{k\in K}$. In most applications, the measurement or acquisition basis $A = (a_k)_{k\in\{1,\cdots,n\}}$ is fixed (e.g. Fourier or Wavelet basis). In order to reduce the acquisition time, one then needs to find a set K of minimal cardinality that provides satisfactory reconstuction results. It is proved in [1], [2] that a good way to proceed consists of drawing the indices of K independently at random according to a distribution $\tilde{\pi}$ that depends on the sensing basis A. This result motivated a lot of authors to propose variable density random sampling strategies (see e.g. [3]–[7]). Fig. 1(a) illustrates a typical sampling pattern used in the MRI context. Simulations confirm that such schemes are efficient in practice. Unfortunately, they can hardly be implemented on real hardware where the physics of the acquisition processes imposes *at least* continuity of the sampling trajectory and sometimes a higher level of smoothness. Hence, actual CS-MRI solutions relie on adhoc solutions such as random radial or randomly perturbed spiral trajectories to impose gradient continuity. Nevertheless these strategies strongly deviate from the theoretical setting and experiments confirm their practical suboptimality.

In this work, we propose an alternative to the independent sampling scheme. It consists of picking a few samples independently at random according to a distribution π and joining them using a travelling salesman problem (TSP) solver in order to design continuous trajectories. The main theoretical result of this paper states that π should be proportional to $\tilde{\pi}^{d/(d-1)}$ where d denotes the space dimension (e.g. $d = 2$ for 2D images, $d = 3$ for 3D images) in order to emulate an independent drawing from distribution $\tilde{\pi}$. Similar ideas were previously proposed in the literature [8], but it seems that no author made this central observation.

The rest of this paper is organized as follows. The notation and definitions are introduced in Section II. Section III contains the main result of the paper along with its proof. Section IV shows how the proposed theory can be implemented in practice. Finally, Section V presents simulation results in the MRI context.

II. Notation and definitions

We shall work on the hypercube $\Omega = [0,1]^d$ with $d \geq 2$. Let $m \in \mathbb{N}$. The set Ω will be partitionned in m^d congruent hypercubes $(\omega_i)_{i\in I}$ of edge length $1/m$. In what follows, $\{x_i\}_{i\in\mathbb{N}^*}$ denotes a sequence of points in the hypercube Ω, independently drawn from a density $\pi : \Omega \mapsto \mathbb{R}_+$. The set of the first N points is denoted $X_N = \{x_i\}_{i\leqslant N}$. For a set of points F, we consider the solution to the TSP, that is the shortest Hamiltonian path between those points. We denote $T(F)$ its length. For any set $R \subseteq \Omega$ we define $T(F, R) = T(F \cap R)$.

We also introduce $C(X_N, \Omega)$ for the optimal curve itself, and $\gamma_N : [0,1] \to \Omega$ the function that parameterizes $C(X_N, \Omega)$ by moving along it at constant speed $T(X_N, \Omega)$.

The Lebesgue measure on an interval $[0,1]$ is denoted $\lambda_{[0,1]}$. We define the *distribution of the TSP solution as follows*.

Definition II.1 *The distribution of the TSP solution is denoted* $\tilde{\Pi}_N$ *and defined, for any Borelian* B *in* Ω *by:*

$$\tilde{\Pi}_N(B) = \lambda_{[0,1]}\left(\gamma_N^{-1}(B)\right).$$

Remark The distribution $\tilde{\Pi}_N$ *is defined for fixed* X_N*. It makes no reference to the stochastic component of* X_N*.*

In order to prove the main result, we need to introduce other tools. For a subset $\omega_i \subseteq \Omega$, we denote the length of $C(X_N, \Omega) \cap \omega_i$ as $T_{|\omega_i}(X_N, \Omega) = T(X_N, \Omega)\tilde{\Pi}_N(\omega_i)$. Using this definition, it follows that:

$$\tilde{\Pi}_N(B) = \frac{T_{|B}(X_N, \Omega)}{T(X_N, \Omega)}, \ \forall B. \tag{1}$$

Let $T_B(F,R)$ be the length of the boundary TSP on the set $F \cap R$. The boundary TSP is defined as the shortest Hamiltonian tour on $F \cap R$ for the metric obtained from the Euclidean metric by the quotient of the boundary of R, that is $d(a,b) = 0$ if $a, b \in \partial R$. Informally, it matches the original TSP while being allowed to travel along the boundary for free. We refer to [9] for a complete description of this concept.

III. Main theorem

Our main theoretical result reads as follows.

Theorem III.1 *Define the density $\tilde{\pi} = \frac{\pi^{(d-1)/d}}{\int_\Omega \pi^{(d-1)/d}(x)dx}$. Then almost surely with respect to the law $\pi^{\otimes\mathbb{N}}$ of the sequence $\{x_i\}_{i\in\mathbb{N}^*}$ of random points in the hypercube, the distribution $\tilde{\Pi}_N$ converges in distribution to $\tilde{\pi}$:*

$$\tilde{\Pi}_N \overset{(d)}{\to} \tilde{\pi} \qquad \pi^{\otimes\mathbb{N}}\text{-a.s.} \tag{2}$$

Intuition: Let us first provide a rough intuition of the result since the exact proof is technical. The distribution $\tilde{\Pi}_N$ in a small cube is the relative length of the TSP in this cube. The number of points N_c in the cube is proportional to π. Approximately, the TSP connects the points with other points in the cube, typically their neighbours, since they are close. Now, the typical distance between two neighbours in the cube is proportional to $N_c^{-1/d}$ or $\pi^{-1/d}$. So that the total length of the TSP in the small cube is proportional to $\pi\pi^{-1/d} = \pi^{1-1/d} \propto \tilde{\pi}$.

The remainder of this section is dedicated to proving this result. The following proposition is central to obtain the proof:

Proposition III.2 *Almost surely, for all ω_i in $\{\omega_i\}_{1\le i\le m^d}$:*

$$\lim_{N\to\infty} \tilde{\Pi}_N(\omega_i) = \tilde{\pi}(\omega_i) \tag{3}$$

$$= \frac{\int_{\omega_i} \pi^{(d-1)/d}(x)\mathrm{d}x}{\int_\Omega \pi^{(d-1)/d}(x)\mathrm{d}x} \qquad \pi^{\otimes\mathbb{N}}\text{-a.s.} \tag{4}$$

The strategy consists in proving that $T_{|\omega_i}(X_N,\Omega)$ tends asymptotically to $T(X_N,\omega_i)$. The estimation of each term can then be obtained by applying the asymptotic result of Beardwood, Halton and Hammersley [10]:

Theorem III.3 *If R is a Lebesgue-measurable set in $\mathbb{R}^d$ such that the boundary ∂R has zero measure, and $\{y_i\}_{i\in\mathbb{N}^*}$, with $Y_N = \{y_i\}_{i\leqslant N}$ is a sequence of i.i.d. points from a density p supported on R, then, almost surely,*

$$\lim_{N\to\infty} \frac{T(Y_N,R)}{N^{(d-1)/d}} = \beta(d)\int_R p^{(d-1)/d}(x)\mathrm{d}x, \tag{5}$$

where $\beta(d)$ depends on the dimension d only.

We shall use a set of classical results on TSP and boundary TSP, that may be found in the survey books [9] and [11].

Useful lemmas. *Let F denote a set of n points in Ω.*

1) *The boundary TSP is superadditive, that is, if R_1 and R_2 have disjoint interiors.*

$$T_B(F, R_1 \cup R_2) \geqslant T_B(F,R_1) + T_B(F,R_2). \tag{6}$$

2) *The boundary TSP is a lower bound on the TSP, both globally and on subsets. If $R_2 \subset R_1$:*

$$T(F,R) \geqslant T_B(F,R) \tag{7}$$

$$T_{|R_2}(F,R_1) \geqslant T_B(F,R_2) \tag{8}$$

3) *The boundary TSP approximates well the TSP [11, Lemma 3.7]):*

$$|T(F,\Omega) - T_B(F,\Omega)| = O(n^{(d-2)/(d-1)}). \tag{9}$$

4) *The TSP in Ω is well-approximated by the sum of TSPs in a grid of m^d congruent hypercubes [9, Eq. (33)].*

$$|T(F,\Omega) - \sum_{i=1}^{m^d} T(F,\omega_i)| = O(n^{(d-2)/(d-1)}). \tag{10}$$

We now have all the ingredients to prove the main results.

Proof of Proposition III.2:

$$\begin{aligned}\sum_{i\in I} T_B(X_N,\omega_i) &\overset{(6)}{\leqslant} T_B(X_N,\Omega)\\ &\overset{(7)}{\leqslant} T(X_N,\Omega) = \sum_{i\in I} T|_{\omega_i}(X_N,\Omega)\\ &\overset{(10)}{\leqslant} \sum_{i\in I} T(X_N,\omega_i) + O(N^{(d-1)/(d-2)})\end{aligned}$$

Let N_i be the number of points of X_N in ω_i.

Since $N_i \leqslant N$, we may use the bound (9) to get:

$$\lim_{N\to\infty} \frac{T(X_N,\omega_i)}{N^{(d-1)/d}} = \lim_{N\to\infty} \frac{T_B(X_N,\omega_i)}{N^{(d-1)/d}}. \tag{11}$$

Using the fact that there are only finitely many ω_i, the following equalities hold almost surely:

$$\begin{aligned}\lim_{N\to\infty} \frac{\sum_{i\in I} T_B(X_N,\omega_i)}{N^{(d-1)/d}} &= \lim_{N\to\infty} \frac{\sum_{i\in I} T(X_N,\omega_i)}{N^{(d-1)/d}}\\ &\overset{(10)}{=} \lim_{N\to\infty} \frac{\sum_{i\in I} T_{|\omega_i}(X_N,\Omega)}{N^{(d-1)/d}}.\end{aligned}$$

Since the boundary TSP is a lower bound (cf. Eqs. (8)-(7)) to both local and global TSPs, the above equality ensures that:

$$\begin{aligned}\lim_{N\to\infty} \frac{T_B(X_N,\omega_i)}{N^{(d-1)/d}} &= \lim_{N\to\infty} \frac{T(X_N,\omega_i)}{N^{(d-1)/d}}\\ &= \lim_{N\to\infty} \frac{T_{|\omega_i}(X_N,\Omega)}{N^{(d-1)/d}} \quad \pi^{\otimes\mathbb{N}}\text{-a.s, } \forall i.\end{aligned} \tag{12}$$

Finally, by the law of large numbers, almost surely $N_i/N \to \pi(\omega_i) = \int_{\omega_i} \pi(x)dx$. The law of any point x_j conditioned on being in ω_i has density $\pi/\pi(\omega_i)$. By applying Theorem III.3 to the hypercubes ω_i and Ω we thus get:

$$\lim_{N\to+\infty} \frac{T(X_N,\omega_i)}{N^{(d-1)/d}} = \beta(d)\int_{\omega_i} \pi(x)^{(d-1)/d}dx \quad \pi^{\otimes\mathbb{N}}\text{-a.s, } \forall i.$$

and

$$\lim_{N\to+\infty} \frac{T(X_N,\Omega)}{N^{(d-1)/d}} = \beta(d)\int_{\Omega} \pi(x)^{(d-1)/d}dx \quad \pi^{\otimes\mathbb{N}}\text{-a.s, } \forall i.$$

Combining this result with Eqs. (12) and (1) yields Proposition III.2. ∎

Proof of Theorem III.1: Let $\varepsilon > 0$ and m be an integer such that $\sqrt{d}m^{-d} < \varepsilon$. Then any two points in ω_i are at distance less than ε.

Using Theorem III.2 and the fact that there is a finite number of ω_i, almost surely, we get: $\lim_{N\to+\infty} \sum_{i\in I} \left|\tilde{\Pi}_N(\omega_i) - \tilde{\pi}(\omega_i)\right| = 0$. Hence, for any N large enough, there is a coupling K of $\tilde{\Pi}_N$ and $\tilde{\pi}$ such that both corresponding random variables are in the same ω_i with probability $1-\varepsilon$. Let $A \subseteq \Omega$ be a Borelian. The coupling satisfies $\tilde{\Pi}_N(A) = K(A \otimes \Omega)$ and $\tilde{\pi}(A) = K(\Omega \otimes A)$. Define the ε-neighborhood by $A^\varepsilon = \{X \in \Omega \,|\, \exists Y \in A,\ \|X - Y\| < \varepsilon\}$. Then, we have: $\tilde{\Pi}_N(A) = K(A \otimes \Omega) = K(\{A \otimes \Omega\} \cap \{|X-Y| < \varepsilon\}) + K(\{A \otimes \Omega\} \cap \{|X-Y| \geqslant \varepsilon\})$. It follows that:

$$\begin{aligned}\tilde{\Pi}_N(A) &\leqslant K(A \otimes A^\epsilon) + K(|X-Y| \geqslant \varepsilon) \\ &\leqslant K(\Omega \otimes A^\varepsilon) + \varepsilon = \tilde{\pi}(A^\varepsilon) + \varepsilon.\end{aligned}$$

This exactly matches the definition of convergence in the Prokhorov metric, which implies convergence in distribution. ∎

IV. Algorithm

The results presented in the previous section can be used to design a continuous sampling pattern with a target density $\tilde{\pi}$. The following algorithm summarizes this idea.

Algorithm 1: An algorithm to generate a continuous sampling pattern according to a target density.

Input: $\tilde{\pi} : \Omega \mapsto \mathbb{R}_+$: a target sampling density.
N: an initial number of drawings.
Output: A continuous sampling curve C.
begin
Define $\pi = \frac{\tilde{\pi}^{d/(d-1)}}{\int_\Omega \tilde{\pi}^{d/(d-1)}(x)dx}$.
Draw N points independently at random according to density π.
Link these points with a travelling salesman solver to generate the curve C.

Applying this algorithm raises various questions: how to choose the target density $\tilde{\pi}$? How to set the initial number of points N? Can the travelling salesman problem be solved for millions of points? We give various hints to the previous questions below.

a) Choosing a density $\tilde{\pi}$: We believe that this question is still treated superficially in the literature and deserves attention. Various strategies can be considered. A common empirical method consists in learning a density on image databases [4]. In the cases of Fourier measurements, this leads to the use of polynomially decreasing densities from low to high frequencies. The same strategy was proposed in [3] with no theoretical justification. The compressed sensing results allow to derive mathematically founded densities [2], [5]. However, as outlined in [7], an important ingredient is missing for these theories to provide good reconstruction results. The standard CS theory relies on the hypothesis that the signal is sparse, with no assumption on the sparsity structure. This makes the current theoretically founded sampling strategies highly suboptimal. Recent works partially address this problem (see e.g. the review paper [12]). However, to the best of our knowledge, the recent focus is on modifying the reconstruction algorithm, rather than deriving optimal sampling patterns.

b) Choosing an initial number of points N: In applications, one usually wishes to sample $\tilde{N}$ points out of the n possible ones. One should thus choose N so that the discretized TSP trajectory contains $\tilde{N}$ points. This problem is well studied in the TSP literature [10], [13]. Theorem III.3 ensures that the length of the TSP trajectory obtained by drawing N points should be close to $L(N) = N^{(d-1)/d}\beta(d)\int_R p^{(d-1)/d}(x)\mathrm{d}x$ where $\beta(d)$ can be evaluated numerically. Concentration results by Talagrand [13] show that this approximation is very accurate for moderate to large values of N. In order to obtain a discrete set of measurements from the continuous trajectory generated by Algorithm 1, we may discretize it with a stepsize Δt. The total number of points sampled is thus $N_s \simeq \lfloor \frac{L(N)}{\Delta t} \rfloor$ if an arclength parameterization is used. A possible way of obtaining approximately $\tilde{N}$ samples is thus to set:

$$N = \lfloor \Delta t L^{-1}(\tilde{N}) \rfloor. \tag{13}$$

c) Solving the TSP: Designing algorithms to solve the TSP is a widely studied problem. The book [9] provides a comprehensive review of exact and approximate algorithms. The TSP is known to be NP-hard and we cannot expect to solve it exactly for a large number of points N. From a theoretical point of view, Arora [14] shows that the TSP solution can be approximated to a factor $(1+\epsilon)$ with a complexity $O(N\log(N)^{1/\epsilon})$. From a practical point of view, there exist many heuristic algorithms that perform well in practice. The heuristics range from those that get within a few percent of optimum for 100,000-city instances in seconds to those that get within fractions of a percent of optimum for instances of this size in a few hours. In our experiments, we used a genetic algorithm [15].

V. Simulation results in MRI

The proposed sampling algorithm was assessed in a 2D MRI acquisition setup where images are sampled in the 2D Fourier domain and compressible in the wavelet domain. Hence, $A = \mathcal{F}^*\Psi$ where $\mathcal{F}^*$ and Ψ denote the discrete Fourier and inverse discrete wavelet transform, respectively. Following [7], it can be shown that a near optimal sampling strategy consists of probing m independent samples of the 2D Fourier plane (k_x, k_y) drawn independently from a target density $\tilde{\pi}$. The image is then reconstructed by solving the following l^1 problem using a Douglas-Rachford algorithm:

$$x^* = \underset{A_m x = y}{\operatorname{argmin}} \|x\|_1$$

where $A_m \in \mathbb{C}^{m\times n}$ is the sensing matrix, $x^* \in \mathbb{C}^n$ is the reconstructed image and $y \in \mathbb{C}^n$ is the collected data. A

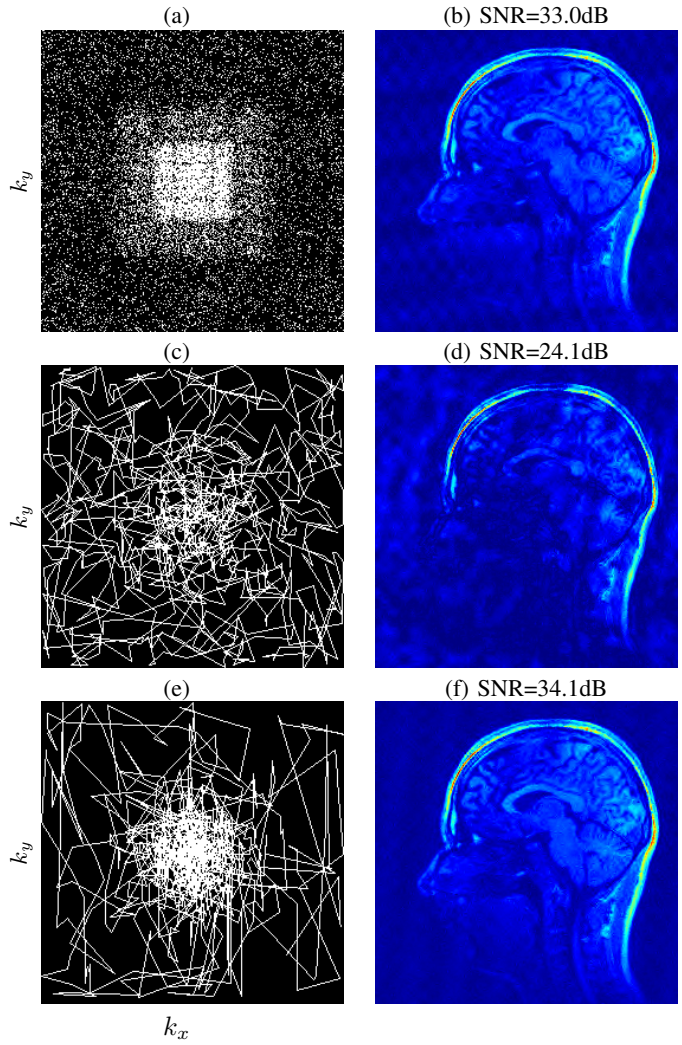


Fig. 1: **Left:** different sampling patterns (with an acceleration factor $r = 5$). **Right:** reconstruction results. From top to bottom: independent drawing from distribution $\tilde{\pi}$ (a), the same followed by a TSP solver (c) and finally independent drawing from distribution $\tilde{\pi}^2$ followed by a TSP solver.

typical realization is illustrated in Fig. 1(a) which in practice cannot be implemented since MRI requires probing samples along continuous curves. To circumvent such difficulties, a TSP solver was applied to such realization in order to join all samples through a countinuous trajectory, as illustrated in Fig. 1(c). Finally, Fig. 1(e) shows a curve generated by a TSP solver after drawing the same amount of Fourier samples from the density $\tilde{\pi}^2$ as underlied by Theorem III.1. In all sampling schemes the number of probed Fourier coefficients was equal to one fifth of the total number (acceleration factor $r = 5$).

Figs. 1(b,d,f) show the corresponding reconstruction results. It is readily seen that an independent random drawing from $\tilde{\pi}^2$ followed by a TSP-based solver yields promising results. Moreover, a dramatic improvement of 10dB was obtained compared to the initial drawing from $\tilde{\pi}$.

VI. Conclusion

Designing sampling patterns lying on continuous curves is central for practical applications such as MRI. In this paper, we proposed and justified an original two-step approach based on a TSP solver to produce such continuous trajectories. It allows to emulate any variable density sampling strategy and could thus be used in a large variety of applications. In the above mentioned MRI example, this method improves the signal-to-noise ratio by 10dB compared to more naive approaches and provides results similar to those obtained using unconstrained sampling schemes. From a theoretical point of view, we plan to assess the convergence rate of the empirical law of the travelling salesman trajectory to the target distribution $\pi^{(d-1)/d}$. From a practical point of view, we plan to develop algorithms that integrate stronger constraints into account such as the maximal curvature of the sampling trajectory, which plays a key role in many applications.

Acknowledgment

The authors would like to thank the mission pour l'interdisciplinarité from CNRS and the ANR SPHIM3D for partial support of Jonas Kahn's visit to Toulouse and the CIMI Excellence Laboratory for inviting Philippe Ciuciu on an excellence researcher position during winter 2013.

References

[1] E.J. Candes and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?," *Information Theory, IEEE Transactions on*, vol. 52, no. 12, pp. 5406–5425, 2006.

[2] H. Rauhut, "Compressive sensing and structured random matrices," *Theoretical foundations and numerical methods for sparse recovery*, vol. 9, pp. 1–92, 2010.

[3] M. Lustig, D. Donoho, and J.M. Pauly, "Sparse MRI: The application of compressed sensing for rapid mr imaging," *Magnetic Resonance in Medicine*, vol. 58, no. 6, pp. 1182–1195, 2007.

[4] F. Knoll, C. Clason, C. Diwoky, and R. Stollberger, "Adapted random sampling patterns for accelerated MRI," *Magnetic Resonance Materials in Physics, Biology and Medicine*, vol. 24, no. 1, pp. 43–50, 2011.

[5] Gilles Puy, Pierre Vandergheynst, and Yves Wiaux, "On variable density compressive sampling," *Signal Processing Letters, IEEE*, vol. 18, no. 10, pp. 595–598, 2011.

[6] F. Krahmer and R. Ward, "Beyond incoherence: stable and robust sampling strategies for compressive imaging," preprint, 2012.

[7] N. Chauffert, P. Ciuciu, and P. Weiss, "Variable density compressed sensing in MRI. Theoretical VS heuristic sampling strategies.," in *proceedings of IEEE ISBI*, 2013.

[8] H. Wang, X. Wang, Y. Zhou, Y. Chang, and Y. Wang, "Smoothed random-like trajectory for compressed sensing MRI," in *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*, 2012, pp. 404–407.

[9] A. M. Frieze and J. E. Yukich, "Probabilistic analysis of the tsp," in *The traveling salesman problem and its variations*, G. Gutin and A. P. Punnen, Eds., vol. 12 of *Combinatorial optimization*, pp. 257–308. Springer, 2002.

[10] J. Beardwood, J.H. Halton, and J.M. Hammersley, *The shortest path through many points*, vol. 55, 1959.

[11] J.E. Yukich, *Probability theory of classical Euclidean optimization problems*, Springer, 1998.

[12] Marco F Duarte and Yonina C Eldar, "Structured compressed sensing: From theory to applications," *Signal Processing, IEEE Transactions on*, vol. 59, no. 9, pp. 4053–4085, 2011.

[13] WanSoo T Rhee and Michel Talagrand, "A sharp deviation inequality for the stochastic traveling salesman problem," *The Annals of Probability*, vol. 17, no. 1, pp. 1–8, 1989.

[14] Sanjeev Arora, "Polynomial time approximation schemes for euclidean traveling salesman and other geometric problems," *Journal of the ACM (JACM)*, vol. 45, no. 5, pp. 753–782, 1998.

[15] P. Merz and B. Freisleben, "Genetic local search for the TSP: New results," in *IEEE International Conference on Evolutionary Computation*, 1997, pp. 159–164.

Incremental Sparse Bayesian Learning for Parameter Estimation of Superimposed Signals

Dmitriy Shutin, Wei Wang, Thomas Jost

***Abstract*—This work discuses a novel algorithm for joint sparse estimation of superimposed signals and their parameters. The proposed method is based on two concepts: a variational Bayesian version of the incremental sparse Bayesian learning (SBL)– fast variational SBL – and a variational Bayesian approach for parameter estimation of superimposed signal models. Both schemes estimate the unknown parameters by minimizing the variational lower bound on model evidence; also, these optimizations are performed incrementally with respect to the parameters of a single component. It is demonstrated that these estimations can be naturally unified under the framework of variational Bayesian inference. It allows, on the one hand, for an adaptive dictionary design for FV-SBL schemes, and, on the other hand, for a fast superresolution approach for parameter estimation of superimposed signals. The experimental evidence collected with synthetic data as well as with estimation results for measured multipath channels demonstrate the effectiveness of the proposed algorithm.**

I. Introduction

In this paper our goal is to estimate the parameters of the following model

$$\boldsymbol{y} = \sum_{l=1}^{L} \boldsymbol{s}(\boldsymbol{\theta}_l) w_l + \boldsymbol{\xi} = \boldsymbol{S}(\boldsymbol{\Theta})\boldsymbol{w} + \boldsymbol{\xi}, \tag{1}$$

where $\boldsymbol{y}$ is an N-dimensional signal vector, $\boldsymbol{s}(\boldsymbol{\theta}_l)$, $l = 1, \ldots, L$, is a set $\boldsymbol{S}(\boldsymbol{\Theta}) = [\boldsymbol{s}(\boldsymbol{\theta}_1), \ldots, \boldsymbol{s}(\boldsymbol{\theta}_L)]$ of known basis functions that are nonlinearly parameterized by $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_L]$; $\boldsymbol{w} = [w_1, \ldots, w_L]^T$ is a vector of basis weights, and $\boldsymbol{\xi}$ is a random perturbation vector, which is often assumed to follow a circular symmetric normal distribution with zero mean and covariance $\boldsymbol{\Sigma} = \lambda^{-1}\boldsymbol{I}$. Such model is almost ubiquitous in signal processing literature, and appears under different disguises in almost all fields of signal processing, e.g., in array processing, channel estimation, radar, to name just a few.

The estimation of signal parameters $\boldsymbol{\Theta}$ and $\boldsymbol{w}$ has often been solved using Expectation-Maximization (EM) type of algorithms [1]–[3], mainly due to the nonlinearity of (1) with respect to the parameter set $\boldsymbol{\Theta}$. Yet these methods are applicable only when the order L of the model is known and fixed – a requirement that is rarely satisfied in practice. However, introducing sparsity constraints into the parameter estimation step might eliminate this drawback of the EM-based estimation.

Sparse signal processing methods have become a very active area of research in recent years due to their rich theoretical nature and their usefulness in a wide range of applications (see e.g., [4]–[6]). With a few minor variations, the general goal of sparse reconstruction is to optimally estimate the parameters $\boldsymbol{w}$ of the model (1) with fixed design matrix $\boldsymbol{S}(\boldsymbol{\Theta}) \equiv [\boldsymbol{s}_1, \ldots, \boldsymbol{s}_L]$. The sparse solution is obtained by imposing specific sparsity constraints on the signal parameter $\boldsymbol{w}$ [4], [6].

Sparse Bayesian learning (SBL) [5], [7], [8] is a family of empirical Bayes techniques that finds a sparse estimate of $\boldsymbol{w}$ by modeling the weights using a hierarchical prior $p(\boldsymbol{w}|\boldsymbol{\alpha})p(\boldsymbol{\alpha}) = \prod_{l=1}^{L} p(w_l|\alpha_l)p(\alpha_l)$, where $p(w_l|\alpha_l)$ is a Gaussian probability density function (pdf) with zero mean and precision parameter α_l, also called the sparsity parameter; larger values of α_l drive the corresponding weight toward zero, thus encouraging a sparse solution. One particular method for SBL recently proposed in the literature is a fast variational SBL (FV-SBL) [8]. The FV-SBL algorithm optimizes the corresponding objective function – the variational lower bound on the model evidence $\log p(\boldsymbol{y})$ – incrementally, i.e., with respect to one basis function at a time. This allows for a very efficient and adaptive implementation of FV-SBL [9] — a feature that is very useful for estimating superimposed signals. Yet due to the nonlinear dependence of (1) on the parameter set $\boldsymbol{\Theta}$, the classical sparse estimation techniques are inapplicable. Obviously, an appropriate sampling or gridding of the parameters $\boldsymbol{\Theta}$ [10], [11] circumvents the nonlinearity problem. This approach, however, does not provide high resolution estimates of the parameters; alternatively, heuristics have to be used to make the gridding adaptive.

Our goal in this paper is to show how SBL technique can be applied to (1) to enable joint sparse signal extraction and superresolution parameter estimation. The proposed technique builds upon two key concepts: variational Bayesian estimation of signal parameters $\boldsymbol{\Theta}$, and an incremental FV-SBL algorithm [8]. Through the use of variational Bayesian techniques both schemes can be jointly realized within the same optimization framework. The first attempts to do so have been proposed in [12], where the authors make a typical assumption on the independence of individual components in (1). Our empirical evidence suggest that this assumption is overly optimistic. The new algorithm is based on the FV-SBL scheme. This allows taking correlations between the linear parameters of the superimposed signals into account. Additionally, the FV-SBL algorithm allows for an adaptive implementation [9], which further accelerates the inference.

Throughout the paper we make use of the following notation. Vectors and matrices are represented as, respectively, boldface lowercase letters, e.g., $\boldsymbol{x}$, and boldface uppercase letters, e.g., $\boldsymbol{X}$. The expression $[\boldsymbol{B}]_{\overline{lk}}$ denotes a matrix ob-

tained by deleting the lth row and kth column from the matrix $\boldsymbol{B}$; similarly, $[\boldsymbol{b}]_{\bar{l}}$ denotes a vector obtained by deleting the lth element from the vector $\boldsymbol{b}$. With a slight abuse of notation we will sometimes refer to a matrix as a set of column vectors; for instance we write $\boldsymbol{a} \in \boldsymbol{X}$ to imply that $\boldsymbol{a}$ is a column in $\boldsymbol{X}$, and $\boldsymbol{X} \setminus \boldsymbol{a}$ to denote a matrix obtained by deleting the column vector $\boldsymbol{a} \in \boldsymbol{X}$. We use $\boldsymbol{e}_l = [0_1, \ldots, 0_{l-1}, 1_l, 0_{l+1}, \ldots, 0_L]^T$ to denote a canonical vector of appropriate dimension. Finally, for a random vector $\boldsymbol{x}$, $\mathrm{CN}(\boldsymbol{x}|\boldsymbol{a}, \boldsymbol{B})$ denotes a circular symmetric normal distribution pdf with mean $\boldsymbol{a}$ and covariance matrix $\boldsymbol{B}$; similarly, for a random variable x, $\mathrm{Ga}(x|a,b) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp(-bx)$ denotes a gamma pdf with parameters a and b.

II. Signal model and adaptive fast sparse Bayesian Learning

In Fig. 1 we show the graphical model that captures the dependencies between the parameters of (1). According to the

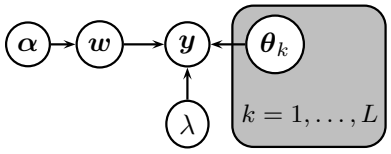


Fig. 1. Graphical model representing (1) with L components.

graph structure, the joint pdf of the graph variables can be factored as

$$p(\boldsymbol{w}, \lambda, \boldsymbol{\alpha}, \boldsymbol{\Theta}, \boldsymbol{y}) = p(\boldsymbol{y}|\boldsymbol{w}, \lambda, \boldsymbol{\theta}) p(\boldsymbol{w}|\boldsymbol{\alpha}) p(\boldsymbol{\alpha}) p(\lambda) p(\boldsymbol{\Theta}), \quad (2)$$

where $p(\boldsymbol{y}|\boldsymbol{w}, \lambda, \boldsymbol{\theta}) = \mathrm{CN}(\boldsymbol{y}|\boldsymbol{S}(\boldsymbol{\Theta})\boldsymbol{w}, \lambda^{-1}\boldsymbol{I})$, $p(\boldsymbol{w}|\boldsymbol{\alpha}) = \prod_{l=1}^{L} \mathrm{CN}(w_l|0, \alpha_l^{-1})$, $p(\boldsymbol{\alpha}) \propto \prod_{l=1}^{L} \alpha_l^{-1}$, and $p(\lambda) \propto \lambda^{-1}$, following the standard SBL model assumption [8], [9].[1] The choice of the prior $p(\boldsymbol{\Theta})$ is arbitrary in the context of this work and is generally application specific. The variational inference on this graph aims at estimating a "proxy" pdf $q(\boldsymbol{w}, \boldsymbol{\alpha}, \lambda, \boldsymbol{\Theta})$ that maximizes the lower bound on the log-evidence $\log p(\boldsymbol{y})$ [13]:

$$\log p(\boldsymbol{y}) \geq \mathop{\mathbb{E}}_{q(\boldsymbol{w}, \boldsymbol{\alpha}, \lambda, \boldsymbol{\Theta})} \log \frac{p(\boldsymbol{w}, \lambda, \boldsymbol{\alpha}, \boldsymbol{\Theta}, \boldsymbol{y})}{q(\boldsymbol{w}, \boldsymbol{\alpha}, \lambda, \boldsymbol{\Theta})} \quad (3)$$

We will assume that $q(\boldsymbol{w}, \boldsymbol{\alpha}, \lambda, \boldsymbol{\Theta})$ factors as follows

$$q(\boldsymbol{w}, \boldsymbol{\alpha}, \lambda, \boldsymbol{\Theta}) = q(\boldsymbol{w}) q(\lambda) \prod_{l=1}^{L} q(\alpha_l) q(\boldsymbol{\theta}_l), \quad (4)$$

with the variational factors in (4) constrained as: $q(\boldsymbol{w}) = \mathrm{CN}(\boldsymbol{w}|\widehat{\boldsymbol{w}}, \widehat{\boldsymbol{\Phi}})$, $q(\alpha_l) = \mathrm{Ga}(\alpha_l|1, \widehat{\alpha}_l^{-1})$, and $q(\lambda) = \mathrm{Ga}(\lambda|N/2, N\widehat{\lambda}^{-1}/2)$. In case of parameters $\boldsymbol{\Theta}$ we assume $q(\boldsymbol{\theta}_l) = \delta(\boldsymbol{\theta}_l - \widehat{\boldsymbol{\theta}}_l)$.[2] By doing so we restrict ourselves to point estimates[3] of these parameters. The optimal $q(\boldsymbol{w}, \boldsymbol{\alpha}, \lambda, \boldsymbol{\Theta})$ is then found by maximizing (3) with respect to the parameters $\{\widehat{\boldsymbol{w}}, \widehat{\boldsymbol{\Phi}}, \widehat{\lambda}, \widehat{\alpha}_1, \widehat{\boldsymbol{\theta}}_1, \ldots, \widehat{\alpha}_L, \widehat{\boldsymbol{\theta}}_l\}$ by cycling through all factors in a "round-robin" fashion [13].

Should the parameters $\boldsymbol{\Theta}$ be assumed as known and fixed, i.e., $\widehat{\boldsymbol{S}} \equiv \boldsymbol{S}(\widehat{\boldsymbol{\Theta}})$, update expressions for the variational parameters can be easily found [14]:

$$\widehat{\boldsymbol{\Phi}} = \left(\widehat{\lambda}\boldsymbol{S}(\widehat{\boldsymbol{\Theta}})^H \boldsymbol{S}(\widehat{\boldsymbol{\Theta}}) + \mathrm{diag}(\widehat{\boldsymbol{\alpha}})\right)^{-1}, \widehat{\boldsymbol{w}} = \widehat{\lambda}\widehat{\boldsymbol{\Phi}}\boldsymbol{S}(\widehat{\boldsymbol{\Theta}})^H \boldsymbol{y}, \quad (5)$$

$$\widehat{\alpha}_l = \frac{1}{|\widehat{w}_l|^2 + \widehat{\Phi}_{ll}}, \widehat{\lambda} = \frac{N}{\|\boldsymbol{t} - \widehat{\boldsymbol{S}}\widehat{\boldsymbol{w}}\|^2 + \mathrm{Trace}(\widehat{\boldsymbol{\Phi}}\widehat{\boldsymbol{S}}^H\widehat{\boldsymbol{S}})}, \quad (6)$$

where $\widehat{w}_l$ is the lth element of the vector $\widehat{\boldsymbol{w}}$, and $\widehat{\Phi}_{ll}$ is the lth element on the main diagonal of the matrix $\widehat{\boldsymbol{\Phi}}$.

The FV-SBL algorithm is a computationally efficient method to accelerate the convergence of the inference expressions (5) and (6). Essentially, it maximizes the bound (3) incrementally: the variational updates of $q(\alpha_l)$ and $q(\boldsymbol{w})$ for a fixed l are performed successively *ad infinitum* while keeping the other variational factors fixed. The convergence of $q(\alpha_l)$ can then be established analytically, which allows for a significant speed-up [8]; moreover, FV-SBL allows for an adaptive implementation, where basis functions can also be easily added to the model (for more information on the adaptive FV-SBL algorithm the reader is referred to [9]).

One of the key features of variational methods is that the factors in (4) can be updated in any order.[4] This allows incorporating the estimation of $q(\boldsymbol{\Theta})$ in the FV-SBL scheme, as explained in the following.

III. Estimation of signal parameters $\boldsymbol{\Theta}$

Let us begin by considering a variational inference of $q(\boldsymbol{\Theta})$. To this end we define $\boldsymbol{\Theta}_{\bar{l}} = \boldsymbol{\Theta} \setminus \boldsymbol{\theta}_l$. Following the standard variation inference steps (see [13]), it can be shown that the bound on $\log p(\boldsymbol{y})$ with respect to $q(\boldsymbol{\theta}_l)$ can be expressed as $\log p(\boldsymbol{y}) \geq \mathbb{E}_{q(\boldsymbol{\theta}_l)} \log \frac{\tilde{p}(\boldsymbol{\theta}_l)}{q(\boldsymbol{\theta}_l)}$, where $\tilde{p}(\boldsymbol{\theta}_l) \propto \exp\left(\mathbb{E}_{q(\boldsymbol{w}, \lambda, \boldsymbol{\Theta}_{\bar{l}})} \log p(\boldsymbol{y}|\boldsymbol{w}, \boldsymbol{\Theta}, \lambda) p(\boldsymbol{\Theta})\right)$. This bound is maximized when the Kullback-Leibler divergence between $q(\boldsymbol{\theta}_l)$ and $\tilde{p}(\boldsymbol{\theta}_l)$ is minimal. Since $q(\boldsymbol{\theta}_l)$ is constrained to be a Dirac distribution, the minimum divergence is achieved when $q(\boldsymbol{\theta}_l)$ is aligned with the mode of $\tilde{p}(\boldsymbol{\theta}_l)$. By evaluating $\tilde{p}(\boldsymbol{\theta}_l)$ we find $\widehat{\boldsymbol{\theta}}_l$ as

$$\widehat{\boldsymbol{\theta}}_l = \mathop{\mathrm{argmax}}_{\boldsymbol{\theta}_l} \Big\{ \log p(\boldsymbol{\theta}_l) - \widehat{\lambda}\|\boldsymbol{r}_l - \widehat{w}_l \boldsymbol{s}(\boldsymbol{\theta}_l)\|^2 - \widehat{\lambda} \sum_{k \neq l} 2\Re\left\{\Phi_{kl} \boldsymbol{s}(\widehat{\boldsymbol{\theta}}_k)^H \boldsymbol{s}(\boldsymbol{\theta}_l)\right\} - \widehat{\lambda}\Phi_{ll}\|\boldsymbol{s}(\boldsymbol{\theta}_l)\|^2 \Big\}, \quad (7)$$

[1] In the following we will consider complex measurement data; extensions for real case are trivial. Also, we will use non-informative form of prior $p(\lambda)$ and $p(\alpha_l)$, $\forall l$. This is known as SBL with automatic relevance determination [7].

[2] More complex forms of $q(\boldsymbol{\theta}_l)$ are outside the scope of this paper.

[3] As a point estimate we understand maximum likelihood or maximum *a posteriori* estimation; the latter case is automatically obtained when a prior $p(\boldsymbol{\theta}_l) \neq \mathrm{const}$.

[4] Note, however, that the order in which the factors are updated is important since different update orderings might lead to different local optima of the variational lower bound.

where

$$\boldsymbol{r}_l = \boldsymbol{y} - \sum_{k=1,k\neq l}^{L} \widehat{w}_k \boldsymbol{s}(\widehat{\boldsymbol{\theta}}_k), \tag{8}$$

and $\Re\{\cdot\}$ denotes the real part operator. Finding $\widehat{\boldsymbol{\theta}}_l$ from (7), which requires nonlinear optimization, readily gives the optimal pdf $q(\boldsymbol{\theta}_l)$. Note that the last two terms in (7) account for the correlations between the weights $\boldsymbol{w}$ of the components, effectively penalizing the estimator for $\boldsymbol{\theta}_l$. We are now ready to bring all the pieces of the proposed sparse estimation scheme together.

The proposed algorithm updates the factors in (4) in groups, where an lth group contains factors $\{q(\boldsymbol{\theta}_l), q(\alpha_l), q(\boldsymbol{w})\}$: starting with $q(\boldsymbol{\theta}_l)$, we then update $q(\alpha_l)$ and $q(\boldsymbol{w})$ using the FV-SBL scheme. If the estimate of $\widehat{\alpha}_l$ diverges, the corresponding component is removed from the model; otherwise, its parameters are updated, and the next component is considered. The realization of the algorithm includes two steps: the initialization and update which are sequentially carried out and summarized in Algorithms 1 and 2, respectively. Note that

Algorithm 1 Initialization

$L \leftarrow 0$, $\boldsymbol{S}(\boldsymbol{\Theta}) \leftarrow [\,]$, $\widehat{\boldsymbol{\Phi}} \leftarrow [\,]$, $\widehat{\boldsymbol{\alpha}} \leftarrow [\,]$, Continue $\leftarrow$ true
while Continue **do**
  Compute $\boldsymbol{r}_{L+1}$ from (8) and $q(\boldsymbol{\theta}_{L+1})$ from (7)
  $\overline{\boldsymbol{s}} \leftarrow \boldsymbol{s}(\widehat{\boldsymbol{\theta}}_{L+1})$
  $\varsigma \leftarrow (\widehat{\lambda}\overline{\boldsymbol{s}}^H\overline{\boldsymbol{s}} - \widehat{\lambda}^2\overline{\boldsymbol{s}}^H\boldsymbol{S}(\boldsymbol{\Theta})\widehat{\boldsymbol{\Phi}}\boldsymbol{S}(\boldsymbol{\Theta})^H\overline{\boldsymbol{s}})^{-1}$
  $\omega^2 \leftarrow \varsigma^2(\widehat{\lambda}\overline{\boldsymbol{s}}^H\boldsymbol{y} - \widehat{\lambda}^2\overline{\boldsymbol{s}}^H\boldsymbol{S}(\boldsymbol{\Theta})\widehat{\boldsymbol{\Phi}}(\boldsymbol{\Theta})^H\boldsymbol{y})^2$
  if $\omega^2 > \varsigma$ **then**
    Add a new component $\boldsymbol{s}(\widehat{\boldsymbol{\theta}}_{L+1})$
    Update $q(\alpha_{L+1})$: $\widehat{\alpha}_{L+1} \leftarrow (\omega^2 - \varsigma)^{-1}$
    Update $q(\boldsymbol{w})$ using a new basis $\overline{\boldsymbol{s}}$:

$$\boldsymbol{X}_{L+1} = \widehat{\boldsymbol{\Phi}}^{-1} - \frac{\widehat{\lambda}\boldsymbol{S}(\boldsymbol{\Theta})^H\overline{\boldsymbol{s}}\overline{\boldsymbol{s}}^H\boldsymbol{S}(\boldsymbol{\Theta})}{\widehat{\alpha}_{L+1} + \overline{\boldsymbol{s}}^H\overline{\boldsymbol{s}}}$$
$$\widehat{\boldsymbol{\Phi}}_{L+1} = \begin{pmatrix} \boldsymbol{X}_{L+1}^{-1} & -\widehat{\lambda}\frac{\widehat{\boldsymbol{\Phi}}\boldsymbol{S}(\boldsymbol{\Theta})^H\overline{\boldsymbol{s}}}{\widehat{\alpha}_{L+1}+\varsigma^{-1}} \\ -\widehat{\lambda}\frac{\overline{\boldsymbol{s}}^H\boldsymbol{S}(\boldsymbol{\Theta})\widehat{\boldsymbol{\Phi}}}{\widehat{\alpha}_{L+1}+\varsigma^{-1}} & (\widehat{\alpha}_{L+1}+\varsigma^{-1})^{-1} \end{pmatrix} \tag{9}$$

    $\boldsymbol{S}(\boldsymbol{\Theta}) \leftarrow [\boldsymbol{S}(\boldsymbol{\Theta}), \boldsymbol{s}(\widehat{\boldsymbol{\theta}}_{L+1})]$,
    $\widehat{\boldsymbol{w}}_{L+1} \leftarrow \widehat{\lambda}\widehat{\boldsymbol{\Phi}}_{L+1}\boldsymbol{S}(\boldsymbol{\Theta})^H\boldsymbol{y}$,
    $L \leftarrow L+1$
  else
    Reject $\boldsymbol{s}(\widehat{\boldsymbol{\theta}}_{L+1})$; Continue = False
  end if
end while

the inverse of a Schur complement $\boldsymbol{X}_{L+1}$ in the Algorithm 1 can be computed efficiently using a rank-one update [15]. The variables ω and ς and the test $\omega^2 > \varsigma$ are explained in detail in [8], [9]. Let us point out that the sparsity inducing property of the whole scheme is "encoded" in the test $\omega^2 > \varsigma$ that determines the convergence of $q(\alpha_l)$ update: if the mean of $q(\alpha_l)$ diverges, the component is removed from the model.

Algorithm 2 Update

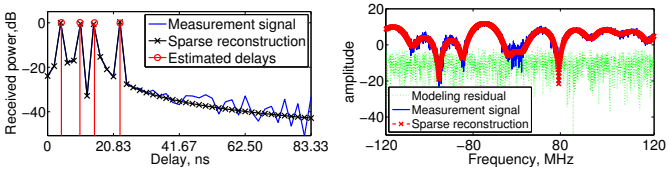
while Continue **do**
  Compute $\boldsymbol{r}_l$ from (8) and $q(\boldsymbol{\theta}_l)$ from (7)
  $\overline{\boldsymbol{s}} \leftarrow \boldsymbol{s}(\widehat{\boldsymbol{\theta}}_l)$
  $\overline{\boldsymbol{S}}_{\overline{l}} \leftarrow \boldsymbol{S}(\widehat{\boldsymbol{\Theta}}) \setminus \overline{\boldsymbol{s}}_l$, $\widehat{\boldsymbol{\Phi}}_{\overline{l}} = \left[\widehat{\boldsymbol{\Phi}} - \frac{\widehat{\boldsymbol{\Phi}}\boldsymbol{e}_l\boldsymbol{e}_l^H\widehat{\boldsymbol{\Phi}}}{\boldsymbol{e}_l^H\widehat{\boldsymbol{\Phi}}\boldsymbol{e}_l}\right]_{\overline{ll}}$
  $\varsigma \leftarrow (\widehat{\lambda}\overline{\boldsymbol{s}}_l^H\overline{\boldsymbol{s}}_l - \widehat{\lambda}^2\overline{\boldsymbol{s}}_l^H\overline{\boldsymbol{S}}_{\overline{l}}\widehat{\boldsymbol{\Phi}}_{\overline{l}}\overline{\boldsymbol{S}}_{\overline{l}}^H\overline{\boldsymbol{s}}_l)^{-1}$
  $\omega^2 \leftarrow (\widehat{\lambda}\varsigma_l\overline{\boldsymbol{s}}_l^H\boldsymbol{y} - \widehat{\lambda}^2\varsigma_l\overline{\boldsymbol{s}}_l^H\overline{\boldsymbol{S}}_{\overline{l}}\widehat{\boldsymbol{\Phi}}_{\overline{l}}\overline{\boldsymbol{S}}_{\overline{l}}^H\boldsymbol{y})^2$
  if $\omega^2 > \varsigma$ **then**
    $\boldsymbol{S}(\boldsymbol{\Theta}) \leftarrow [\boldsymbol{S}_{\overline{l}}, \overline{\boldsymbol{s}}]$
    Update $q(\alpha_l)$: $\widehat{\alpha}_l \leftarrow (\omega^2 - \varsigma)^{-1}$;
    Update $q(\boldsymbol{w})$ using $\overline{\boldsymbol{s}}$ and $\widehat{\alpha}_l$
    Compute $\widehat{\boldsymbol{\Phi}}$ as in (9), $\widehat{\boldsymbol{w}} \leftarrow \widehat{\lambda}\widehat{\boldsymbol{\Phi}}\boldsymbol{S}(\boldsymbol{\Theta})^H\boldsymbol{y}$
  else
    Remove the component $\overline{\boldsymbol{s}}_l$
    $\boldsymbol{S}(\boldsymbol{\Theta}) \leftarrow \overline{\boldsymbol{S}}_{\overline{l}}$; $L \leftarrow L-1$
    Update $q(\boldsymbol{\alpha})$: $\widehat{\boldsymbol{\alpha}}_{\overline{l}} \leftarrow [\widehat{\boldsymbol{\alpha}}]_{\overline{l}}$
    Update $q(\boldsymbol{w})$:$\widehat{\boldsymbol{\Phi}} \leftarrow \widehat{\boldsymbol{\Phi}}_{\overline{l}}$, $\widehat{\boldsymbol{w}} \leftarrow \widehat{\lambda}\widehat{\boldsymbol{\Phi}}\boldsymbol{S}(\boldsymbol{\Theta})^H\boldsymbol{y}$
  end if
end while

IV. Simulation results

Here we study the performance of the proposed estimation scheme using synthetic data generated according to model (1) as well measured data.

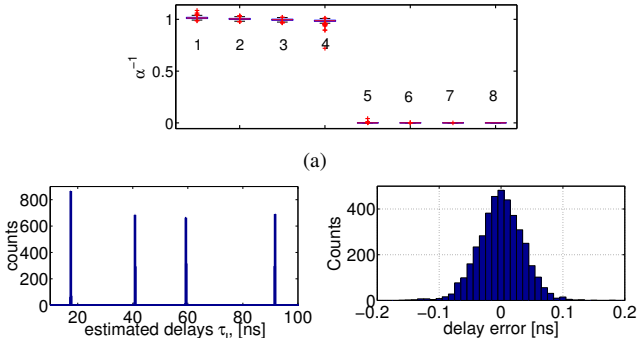
For simplicity we consider a Single-Input-Single-Output channel with zero Doppler shift; thus, each component is characterized by a delay $\boldsymbol{\theta}_l = \{\tau_l\}$ and a complex gain w_l, i.e., $\boldsymbol{y} = \sum_{l=1}^{L} w_l\boldsymbol{s}(\tau_l) + \boldsymbol{\xi}$. The channel is synthesized in frequency domain with the following parameters: $L = 4$, $N = 1537$, signal bandwidth is $f_B = 120\,\mathrm{MHz}$; the signal was sampled at the Nyquist rate and the carrier frequency is assumed to be $5.2\,\mathrm{GHz}$. The delays of synthetic multipath components are set to $17.5\,\mathrm{ns}$, $40.83\,\mathrm{ns}$, $59.33\,\mathrm{ns}$, and $91.67\,\mathrm{ns}$; corresponding complex amplitudes are selected as $w_l = e^{j\varphi_l}$, $l = 1,\ldots,4$, where φ_l is a random variable drawn from a uniform distribution. As as replica of the transmitted signal $s(t)$ we use the actual measured calibration data of the Medav RUSK-DLR channel sounder [16]. The calibration data is obtained by directly connecting the transmitter to the receiver and recording the received signal. Its sampled version is then used to construct a vector $\boldsymbol{s}(\cdot)$, whose shifted versions are used in synthesizing the channel, as well as in the estimation step.

In Fig.2 we show the estimated impulse response and transfer function for 15dB SNR. Observe that the estimated responses closely follow the measured data with only four components. Let us stress that depending on the actual noise realization, the algorithm tends to overestimate the number of components. In Fig. 3(a) we plot distributions of estimated sparsity parameters for all detected components collected over 1000 Monte Carlo runs with different noise realizations. Note that in the worst case the algorithm identifies up to 8 components, all of which are very close to the true ones. This



(a) (b)

Fig. 2. Estimated synthetic channel in a) time domain and b) frequency domain for 15dB SNR.



(a) (b) (c)

Fig. 3. a)Estimated sparsity parameters $\widehat{\alpha}$ and b) delays for synthetic scenarios. c) Error distribution for estimated τ_l when 4 components are found.

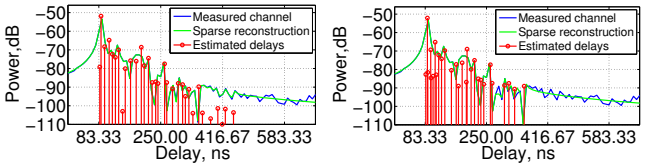
can also be seen in Fig. 3(b), where we plot the distribution of all estimated delays. The inverse sparsity parameters of these artifact components are also quite small, which means they do not contribute to the model. In the case when the algorithm identifies exactly 4 components we can compute the error between the true and the estimated delay. In Fig 3(c) we plot the histogram of estimated delay errors. Note that the estimation error is smaller than 1% of the used sampling period (≈ 8.3ns).

A. Estimation results for measured multipath channels.

Here we consider the estimation of the actual measured multipath channels using the proposed algorithm. The data was collected during a recent measurement campaign [16] performed at German Aerospace Center in Oberpfaffenhofen, Germany. The measurements parameters coincide with those used in simulations. As the actual channel parameters cannot be known for a measured channel, we qualitatively compare the performance of the proposed scheme to that of the SAGE algorithm [3]. As the latter scheme requires knowing the number of components L, we first estimate it using the proposed method, and then use SAGE with same model order. The estimation results are summarized in Fig. In total $L = 31$ path has been identified. Despite some similarities, the SAGE algorithm tends to miss weak components. Also, it tends to cluster multipaths around areas of high power, which often indicates estimation artifacts [12].

V. Conclusion

In this work an adaptive fast variational Sparse Bayesian Learning (FV-SBL) algorithm has been used for parameter estimation of superimposed signals. Using variational framework both superresolution parameter estimation and sparse signal extraction can be done jointly by minimizing the common objective function. Thus, the proposed scheme "frees" the classical EM-based parameter estimation from specifying a model order. Simulation results obtained with synthetic and measured data demonstrate the effectiveness of the proposed estimation scheme. However, more detailed analysis of experimental data is needed.



(a) (b)

Fig. 4. Estimated channel response using a) proposed algorithm and b) using SAGE algorithm.

References

[1] H. Krim and M. Viberg, "Two decades of array signal processing research: the parametric approach," *IEEE Signal Processing Mag.*, pp. 67–94, July 1996.

[2] M. Feder and E. Weinstein, "Parameter Estimation of Superimposed Signals Using the EM Algorithm," *IEEE Trans. on Acoustics, Speech, and Sig. Proc.*, vol. 36, no. 4, pp. 477–489, April 1988.

[3] B. Fleury, M. Tschudin, R. Heddergott, D. Dahlhaus, and K. I. Pedersen, "Channel parameter estimation in mobile radio environments using the SAGE algorithm," *IEEE Journal on Sel. Areas in Comm.*, vol. 17, no. 3, pp. 434–450, March 1999.

[4] E. Candes and M. Wakin, "An introduction to compressive sampling," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 21–30, Mar. 2008.

[5] D. G. Tzikas, A. C. Likas, and N. P. Galatsanos, "The variational approximation for Bayesian inference," *IEEE Signal Process. Mag.*, vol. 25, no. 6, pp. 131–146, November 2008.

[6] D. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289 –1306, april 2006.

[7] M. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Machine Learning Res.*, vol. 1, pp. 211–244, June 2001.

[8] D. Shutin, T. Buchgraber, S. R. Kulkarni, and H. V. Poor, "Fast variational sparse Bayesian learning with automatic relevance determination for superimposed signals," *IEEE Trans. on Sig. Proc.*, vol. 59, no. 12, pp. 6257 –6261, Dec. 2011.

[9] ——, "Fast adaptive variational sparse Bayesian learning with automatic relevance determination," in *IEEE Int. Conf. on Acoustics Speech and Sig. Proc.*, Prague, Czech Republic, May 2011, pp. 2180–2183.

[10] W. Bajwa, J. Haupt, A. Sayeed, and R. Nowak, "Compressed channel sensing: A new approach to estimating sparse multipath channels," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 1058–1076, Jun. 2010.

[11] D. Malioutov, M. Cetin, and A. Willsky, "A sparse signal reconstruction perspective for source localization with sensor arrays," *IEEE Trans. on Sign. Proc.*, vol. 53, no. 8, pp. 3010–3022, 2005.

[12] D. Shutin and B. H. Fleury, "Sparse variational Bayesian SAGE algorithm with application to the estimation of multipath wireless channels," *IEEE Trans. on Sig. Proces.*, vol. 59, no. 8, pp. 3609 – 3623, Aug. 2011.

[13] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer, 2006.

[14] C. M. Bishop and M. E. Tipping, "Variational relevance vector machines," in *Proc. 16th Conf. Uncer. in Artif. Intell.* San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2000, pp. 46–53.

[15] W. W. Hager, "Updating the inverse of a matrix," *SIAM Review*, vol. 31, no. 2, pp. pp. 221–239, 1989.

[16] W. Wang and T. Jost, "A low-cost platform for time-variant wireless channel measurements with application to positioning," *IEEE Trans. on Instrumentation and Measurement*, vol. 61, no. 6, pp. 1597 –1604, june 2012.

Sparse MIMO Radar with Random Sensor Arrays and Kerdock Codes

Thomas Strohmer, Haichao Wang
Department of Mathematics
UC Davis
Davis, California 95616

***Abstract*—We derive a theoretical framework for the recoverability of targets in the azimuth-range-Doppler domain using random sensor arrays and tools developed in the area of compressive sensing. In one manifestation of our theory we use Kerdock codes as transmission waveforms and exploit some of their peculiar properties in our analysis. Not only is our result the first rigorous mathematical theory for the detection of moving targets using random sensor arrays, but also the transmitted waveforms satisfy a variety of properties that are very desirable and important from a practical viewpoint.**

I. Introduction

In recent years, radar systems employing multiple antennas at the transmitter and the receiver (also referred to as MIMO radar, where MIMO stands for multiple-input multiple-output) have attracted enormous attention in the engineering and signal processing community. Existing theory focuses mainly on the detection of a single target. Only very recently, in the footsteps of compressive sensing, do we see the emergence of a rigorous mathematical theory for MIMO radar that addresses the more realistic and more interesting case of multiple targets [13]. However, for the widely popular case of randomly spaced antennas, the mathematical theory is still in its infancy.

On the other hand, mathematicians and engineers have devoted substantial efforts to the design of radar transmission waveforms that satisfy a variety of desirable properties. The vast majority of this research has focused on single antenna radar systems, and it is a priori not clear whether and how these waveforms can be utilized for MIMO radar. In this paper we bring together these two independent areas of research, MIMO radar with random antenna arrays and radar waveform design, by developing a rigorous mathematical framework for accurate target detection via random arrays, which at the same time utilizes some of the most attractive radar waveforms, such as Kerdock codes.

In radar processing we are interested in a given area, which is usually called the radar scene. We would like to detect the location and the strength of the objects of interest, as well as the velocity if there is relative motion between the radar and the objects. Usually the radar scene is divided into a grid of range-azimuth-Doppler (distance, direction and speed) resolution cells. In many practical cases the radar scene is sparse in the sense that only a small fraction of the grid points is occupied by the targets of interest.

While the conventional radar processing techniques do not take advantage of the fact that the radar scene is often sparse, the recent development of compressive sensing (CS) provides us the possibility to utilize this structure. In fact recent works (such as [8], [12], [13] and the reference therein) created important linkage between radar processing and CS. As in CS, we also have to solve the following inverse problem in radar processing:

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{w}, \tag{1}$$

where $\mathbf{y}$ is a vector of measurements collected by the receiver antennas over an observation interval, $\mathbf{A}$ is a measurement matrix whose columns correspond to the signal received from a single unit-strength scatterer at a particular range-azimuth-Doppler grid point, $\mathbf{x}$ is a vector whose elements represent the complex amplitudes of the scatterers, and $\mathbf{w}$ is the unknown noise vector. Note that this is an under-determined equation (if $\dim(\mathbf{y}) < \dim(\mathbf{x})$) and in general it has infinitely many solutions. But given that $\mathbf{x}$ is sparse from our assumption, this problem can have a satisfactory solution.

One of the algorithms that can be used to solve (1) is as follows:

$$\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2 + \lambda\|\mathbf{x}\|_1, \tag{2}$$

which is also known as lasso. Here $\lambda > 0$ is a regularization parameter that trades off goodness of fit with sparsity. [3] showed that if we assume $\mathbf{x}$ is drawn from a generic S-sparse target model (i.e. the support of $\mathbf{x}$ is selected uniformly at random and the phases of the non-zero entries of $\mathbf{x}$ are random and uniformly distributed in $[0, 2\pi)$) then with a particular choice of λ, (2) will recover the support of $\mathbf{x}$ correctly with high probability given that the coherence and the operator norm of $\mathbf{A}$ can be well controlled.

Our paper provides two main contributions: (i) We derive the first rigorous mathematical theory for the detection of moving targets in the azimuth-range-Doppler domain for random sensor arrays. (ii) The transmitted waveforms satisfy a variety of properties that are very desirable and important from a practical viewpoint. In particular, we show that Kerdock sequences, which would perform very poorly in single-antenna radar, are nearly ideally suited for MIMO radar with randomly spaced antennas. Thus, our framework does not just lead to useful theoretical insights, but also has a very strong practical appeal.

A. Connections with prior work and innovations

Random sensor arrays have been around for decades [11]. Recently, [4] made an explicit connection between random sensor arrays and the CS. The setup in [4] is quite different from ours, since the author is only concerned with angular resolution, while it is often crucial in practice to be able to estimate range and Doppler as well. Moreover, the theoretical analysis in [4] follows more an engineering style and places less emphasis on mathematical rigor.

On the other hand, [13] is closest to this paper. [13] considers a MIMO radar setting with a very specific (non-random) choice for the antenna locations, but random waveforms, while the current paper deals with randomly spaced antennas, but very specific, deterministic waveforms. In practice, the random waveforms are much harder to implement on a digital device and they exhibit a larger peak-to-average-power ratio compared to carefully designed deterministic waveforms. On the other hand it makes no difference from the viewpoint of physics or hardware, if we place the antennas at random or at deterministic locations.

B. Notation

For a matrix $\mathbf{A}$, we use $\mathbf{A}^*$ to denote its adjoint matrix, which is its conjugate transpose. The operator norm of $\mathbf{A}$ is the largest singular value of $\mathbf{A}$ and is denoted by $\|\mathbf{A}\|_{\text{op}}$.

For $\mathbf{x} \in \mathbb{C}^n$, let $\mathbf{T}_\tau$ denote the circulant translation operator, defined by $\mathbf{T}_\tau \mathbf{x}(l) = \mathbf{x}(l-\tau)$, for $\tau = 1, \dots, n$, where $l-\tau$ is understood modulo n, and let $\mathbf{M}_f$ be the modulation operator defined by $\mathbf{M}_f \mathbf{x}(l) = \mathbf{x}(l) e^{2\pi i f l/n}$.

II. Problem Setup

We consider a MIMO radar employing N_T antennas at the transmitter and N_R antennas at the receiver. We assume for convenience that transmitter and receiver are co-located. Furthermore, we assume a coherent propagation scenario, i.e., the element spacing is sufficiently small so that the radar return from a given scatterer is fully correlated across the array. The arrays and all the scatterers are assumed to be in the same 2-D plane. The extension to the 3-D case is straightforward.

The array manifolds $\mathbf{a}_T(\beta)$, $\mathbf{a}_R(\beta)$ with randomly spaced antennas are given by

$$\mathbf{a}_T(\beta) = \left[e^{2\pi i p_1 \beta}, e^{2\pi i p_2 \beta}, \dots, e^{2\pi i p_{N_T} \beta}\right]^T, \tag{3}$$

and

$$\mathbf{a}_R(\beta) = \left[e^{2\pi i q_1 \beta}, e^{2\pi i q_2 \beta}, \dots, e^{2\pi i q_{N_R} \beta}\right]^T, \tag{4}$$

where we assume that the relative antenna spacings p_j's and q_j's are i.i.d. uniformly on $[0, \frac{N_R N_T}{2}]$. The j-th transmit antenna repeatedly transmits the signal $s_j(t)$ and the receive antennas take N_s samples of the signal. Let $\mathbf{Z}(t; \beta, \tau, f)$ be the $N_R \times N_s$ noise-free received signal matrix from a unit strength target at direction β, delay τ, and Doppler f (corresponding to its radial velocity with respect to the radar). Then

$$\mathbf{Z}(t; \beta, \tau, f) = \mathbf{a}_R(\beta) \mathbf{a}_T^T(\beta) \mathbf{S}_{\tau,f}^T,$$

where $\mathbf{S}_{\tau,f}$ is a $N_s \times N_T$ matrix whose columns are the circularly delayed and Doppler shifted signals $s_j(t-\tau)e^{2\pi i f t}$.

We let $\mathbf{z}(t; \beta, \tau, f) = \text{vec}\{\mathbf{Z}\}(t; \beta, \tau, f)$ be the noise-free vectorized received signal. We set up a discrete azimuth-range-Doppler grid $\{\beta_l, \tau_j, f_k\}$ for $1 \le l \le N_\beta$, $1 \le j \le N_\tau$ and $1 \le k \le N_f$, where $\Delta_\beta, \Delta_\tau$ and Δ_f denote the corresponding discretization stepsizes. Using vectors $\mathbf{z}(t; \beta_l, \tau_j, f_k)$ for all grid points (β_l, τ_j, f_k) we construct a complete response matrix $\mathbf{A}$ whose columns are $\mathbf{z}(t; \beta_l, \tau_j, f_k)$ for $1 \le l \le N_\beta$ and $1 \le j \le N_\tau$, $1 \le k \le N_f$. In other words, $\mathbf{A}$ is a $N_R N_s \times N_\tau N_\beta N_f$ matrix with columns

$$\mathbf{A}_{\beta,\tau,f} = \mathbf{a}_R(\beta) \otimes \mathbf{S}_{\tau,f} \mathbf{a}_T(\beta). \tag{5}$$

Assume that the radar illuminates a scene consisting of S scatterers located on S points of the (β_l, τ_j, f_k) grid. Let $\mathbf{x}$ be a sparse vector whose non-zero elements are the complex amplitudes of the scatterers in the scene. The zero elements corresponds to grid points which are not occupied by scatterers. We can then define the radar signal $\mathbf{y}$ received from this scene by (1) where $\mathbf{y}$ is an $N_R N_s \times 1$ vector, $\mathbf{x}$ is an $N_\tau N_\beta N_f \times 1$ sparse vector and $\mathbf{w}$ is an $N_R N_s \times 1$ complex Gaussian noise vector. Our goal is to solve for $\mathbf{x}$, i.e., to locate the scatterers (and their reflection coefficients) in the azimuth-delay-Doppler domain.

As for the signal matrix $\mathbf{S}$, for our main results we choose the Kerdock waveforms, as described in Section III, as discrete transmission waveforms.

Remark: The assumption that the targets lie on the grid points, while common in compressive sensing, is certainly restrictive. A violation of this assumption will result in a model mismatch, sometimes dubbed *gridding error*, which can potentially be quite severe [9], [5]. Recently some interesting strategies have been proposed to overcome this gridding error [6], [15]. But these methods are not directly applicable to our setting. This model mismatch issue is beyond the scope of this paper and will be addressed in our future research.

III. Kerdock Codes

We briefly review the construction of Kerdock codes and some of their fundamental properties. A simple way to construct these Kerdock codes is the following, in which they arise as eigenvectors of time-frequency shift operators. Let p be an odd prime number and consider the translation operator $\mathbf{T}$ and the modulation operator $\mathbf{M}$ on $\mathbb{C}^p$. For each $k = 0, \dots, p-1$ we compute the eigenvector decomposition of $\mathbf{T}\mathbf{M}_k$ (which always exists, since $\mathbf{T}\mathbf{M}_k$ is a unitary matrix)

$$U_{(k)} \Sigma_{(k)} U_{(k)}^* = \mathbf{T}\mathbf{M}_k, \tag{6}$$

where the unitary matrix $U_{(k)}$ contains the eigenvectors of $\mathbf{T}\mathbf{M}_k$ and the diagonal matrix $\Sigma_{(k)}$ the associated eigenvalues[1]. Furthermore, we define $U_{(p)} := \mathbf{I}_p$. Now, let $u_{k,j}$ be the j-th column of $U_{(k)}$. The set consisting of the $p^2 + p$ vectors $\{u_{k,j}, k = 0, \dots, p; j = 0, \dots, p-1\}$ forms a $\mathbb{Z}_p$-Kerdock code. There are numerous equivalent ways to derive this Kerdock code, but, as pointed out earlier, not *all* Kerdock codes over $\mathbb{Z}_p$ are equivalent (see also the comment following

[1]The attentive reader will have noticed that $U_{(0)}$ is just the $p \times p$ DFT matrix.

Corollary 11.6 in [2]). But we will be a bit sloppy, and simply refer to the Kerdock code constructed above as *the* Kerdock code.

In the following theorem we collect those key properties of Kerdock codes that are most relevant for radar. These properties are either explicitly proved in [2], [10] or can be derived easily from properties stated in those papers.

Theorem 3.1: Kerdock codes over $\mathbb{Z}_p$, where p is an odd prime, satisfy the following properties:

(i) Mutually unbiased bases: For all $k = 0, \dots, p$ and all $j = 0, \dots, p-1$, there holds:

$$|\langle u_{k,j}, u_{k',j'}\rangle| = \begin{cases} 1 & \text{if } k = k', j = j', \\ 0 & \text{if } k = k', j \neq j', \\ \frac{1}{\sqrt{p}} & \text{if } k \neq k'. \end{cases}$$

(ii) Time-frequency "autocorrelation":
(a) For any fixed $(f, l) \neq (0, 0)$ there exists a unique k_0 such that

$$|\langle \mathbf{M}_f \mathbf{T}_l u_{k_0,j}, u_{k_0,j}\rangle| = 1 \qquad \text{for } j = 0, \dots, p-1, \quad (7)$$

$$|\langle \mathbf{M}_f \mathbf{T}_l u_{k,j}, u_{k,j}\rangle| = 0 \qquad \text{for } k \neq k_0. \quad (8)$$

(b) For any fixed $0 \leq k \leq p-1$, there exist distinct (f_r, l_r), $r = 1, \dots, p$ such that

$$|\langle \mathbf{M}_{f_r} \mathbf{T}_{l_r} u_{k,j}, u_{k,j}\rangle| = 1 \qquad \text{for } j = 0, \dots, p-1, \quad (9)$$

(iii) Time-frequency crosscorrelation: For all $k \neq k'$ and all f and l there holds:

$$|\langle \mathbf{M}_f \mathbf{T}_l u_{k,j}, u_{k',j}\rangle| \leq \frac{1}{\sqrt{p}} \qquad \text{for } j = 0, \dots, p-1. \quad (10)$$

We emphasize though that Kerdock codes would not be very effective for a radar system with a single transmit antenna (SISO or SIMO radar). This can be easily seen as follows: Suppose we only have one antenna that transmits one waveform $\vec{s}$. Because of (9), $\vec{s}$ is equal to (up to a phase factor) $\mathbf{M}_f \mathbf{T}_l \vec{s}$ for some f, l. In practice, this prevents us from determining the distance and the speed of the object.

As a consequence of the aforementioned ambiguity we will not use *all* of the Kerdock codes as transmission signals for our MIMO radar, instead we will choose one code for each index k. The reason is that we need the waveforms to have low time-frequency crosscorrelation, while (10) only holds when k and k' are different.

*Definition 3.2 (***Kerdock waveforms):** Let $\{\mathbf{u}_{k,j}, k = 0, \dots, p, j = 0, \dots, p-1\}$ be a Kerdock code over $\mathbb{Z}_p$. The *Kerdock waveforms* $\mathbf{k}_0, \dots, \mathbf{k}_r$, where $r < p$, are given by $\mathbf{k}_k = \mathbf{u}_{k,j}$ for some arbitrary j. In other words, for each $k = 0, \dots, r-1$ we pick an arbitrary vector from the orthonormal basis $\{\mathbf{u}_{k,j}\}_{j=0}^{p-1}$.

Note Kerdock waveforms do not include any canonical vectors, since only the first r unitary matrices $\mathbf{U}_{(0)}, \dots, \mathbf{U}_{(r-1)}$ are considered and r is strictly less than p (recall $\mathbf{U}_{(p)} = \mathbf{I}_p$).

IV. The main theorem

As mentioned in the introduction, a standard approach to solve (1) when $\mathbf{x}$ is sparse, is given in (2). But instead of (2), we will use the *debiased lasso*. That means first we compute an approximation $\tilde{I}$ for the support of $\mathbf{x}$ by solving (2). This is the detection step. Then, in the estimation step, we "debias" the solution by computing the amplitudes of $\mathbf{x}$ via solving the reduced-size least squares problem $\min \|\mathbf{A}_{\tilde{I}} \mathbf{x}_{\tilde{I}} - \mathbf{y}\|_2$, where $\mathbf{A}_{\tilde{I}}$ is the submatrix of $\mathbf{A}$ consisting of the columns corresponding to the index set $\tilde{I}$, and similarly for $\mathbf{x}_{\tilde{I}}$.

We assume x is drawn from a generic S-sparse target model. We are now ready to state our main result (more details of this theorem can be found in [14]).

Theorem 4.1: Consider $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{w}$, where $\mathbf{A}$ is defined as in (5) and $\mathbf{w}_j \in \mathcal{C}N(0, \sigma^2)$. Assume that the positions of the transmit and receive antennas p_j's and q_j's are chosen i.i.d. uniformly on $[0, \frac{N_R N_T}{2}]$ at random. Suppose further that each transmit antenna sends a different Kerdock waveform, i.e. the columns of the signal matrix $\mathbf{S}$ are different Kerdock waveforms. Suppose that

$$\max\left(N_R N_T, 32 N_T^3 \log N_\tau N_f N_\beta\right) \leq N_s = N_\tau, \quad (11)$$

and also

$$\log^2 N_\tau N_f N_\beta \leq N_T \leq N_R. \quad (12)$$

If $\mathbf{x}$ is drawn from the generic S-sparse scatterer model with

$$S \leq \frac{c_0 N_\tau}{\log N_\tau N_f N_\beta} \quad (13)$$

for some constant $c_0 > 0$, and if

$$\min_{k \in I} |\mathbf{x}_k| > \frac{8\sqrt{3}\sigma}{\sqrt{N_R N_T}} \sqrt{2 \log N_\tau N_f N_\beta}, \quad (14)$$

then the solution $\tilde{\mathbf{x}}$ of the debiased lasso computed with $\lambda = 2\sigma\sqrt{2 \log N_\tau N_f N_\beta}$ satisfies with high probability

$$\operatorname{supp}(\tilde{\mathbf{x}}) = \operatorname{supp}(\mathbf{x}), \quad (15)$$

and

$$\frac{\|\tilde{\mathbf{x}} - \mathbf{x}\|_2}{\|\mathbf{x}\|_2} \leq \frac{5\sigma\sqrt{3 N_R N_s}}{\|\mathbf{y}\|_2}. \quad (16)$$

Remarks:

1) The condition $N_T \leq N_R$ in (12) is by no means necessary, but rather to make our computation a little cleaner. We could change it into $N_T \leq 2N_R$, then the theorem would be true with a slightly different probability of success.
2) It may seem that the conditions in (11) and (12) are a bit restrictive. But, in practice, our method works with a broad range of parameters.

The proof of the above theorem is rather involved and too long to be included in this brief paper. The full proof of this theorem, as well as other results presented in this paper can be found in the journal version of this paper [14]. Here, we can only sketch the key steps. To prove Theorem 4.1, we use a theorem by Candès and Plan (Theorem 1.3 in [3]) which requires to estimate the operator norm of $\mathbf{A}$ and the coherence of $\mathbf{A}$. The original theorem only treats the real-valued case, it can be extended to complex-valued case after some straightforward modifications (see Appendix B in [13]).

V. Extension of the main result

In this section, we present a modified version of Theorem 4.1 that applies to waveforms that satisfy slightly more restrictive incoherence conditions. As such, Theorem 5.1 below does not hold for Kerdock waveforms, but the advantage compared to Theorem 4.1 is that the result also applies to radar systems with only one transmit antenna.

Theorem 5.1: Consider $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{w}$, where $\mathbf{A}$ is defined as in (5) and $\mathbf{w}_j \in \mathcal{C}N(0, \sigma^2)$. Suppose the transmission waveforms satisfy the following conditions

$$|\langle \vec{s}_j, \mathbf{M}_f \mathbf{T}_\tau \vec{s}_j \rangle| \leq \frac{\gamma}{\sqrt{p}} \quad \text{for} \quad (f,\tau) \neq (0,0), \tag{17}$$

$$|\langle \vec{s}_k, \mathbf{M}_f \mathbf{T}_\tau \vec{s}_j \rangle| \leq \frac{\gamma}{\sqrt{p}} \quad \text{for} \quad k \neq j, \tag{18}$$

where $\gamma > 0$ is a fixed constant. Assume that the positions of the transmit and receive antennas p_j's and q_j's are chosen i.i.d. uniformly on $[0, \frac{N_R N_T}{2}]$. Choose the same discretization stepsizes to be $\Delta_\beta = \frac{2}{N_R N_T}$,$\Delta_\tau = \frac{1}{2B}$, $\Delta_f = \frac{1}{T}$ and suppose that

$$\max\left(\gamma^2 N_R N_T, 16\gamma^2 N_T \log^3 N_\tau N_f N_\beta\right) \leq N_s = N_\tau$$

and also

$$\gamma^2 N_T \log^4 N_\tau N_f N_\beta \leq N_s N_R, \quad \log^2 N_\tau N_f N_\beta \leq N_T \leq N_R.$$

Then if the rest of the conditions of Theorem 4.1 hold, we have the same conclusion as in Theorem 4.1.

There are several examples of signal sets that satisfy the above conditions. Perhaps the most intriguing example is the finite harmonic oscillator system (FHOS) constructed in [7]. This signal set in $\mathbb{C}^p$ (where p is a prime number) of cardinality $\mathcal{O}(p^3)$ satisfies (17) and (18) with $\gamma = 4$. An elementary construction of the FHOS for prime number $p \geq 5$ can be found in [16].

VI. Simulations

In this section we will demonstrate the performance of our algorithms via numerical simulations. We use the Matlab Toolbox TFOCS ([1]).

We choose Kerdock codes as transmission waveforms along with the parameters: $N_T = 6, N_R = 6, N_s = 37, N_f = 37$. The number of the scatters $S = 10, 20, 40$ while the SNR is chosen to be 20dB.

The values of the estimated vector $\hat{\mathbf{x}}$ corresponding to the true scatterer locations are compared to a threshold. Detection is declared whenever a value exceeds the threshold. The probability of detection P_d is defined as the number of detections divided by S. Next the values of the estimated vector $\hat{\mathbf{x}}$ corresponding to locations not containing scatterers are compared to the same threshold. A false alarm is declared whenever one of these values exceeds the threshold. The probability of false alarm P_{fa} is defined as the number of false alarms divided by $n - S$, where n is the signal dimension. The results are averaged over the 50 repetitions of the experiment. The probabilities are computed for a range of values of the threshold to produce the so-called Receiver Operating Characteristics (ROC)- the graph of P_d vs. P_{fa}.

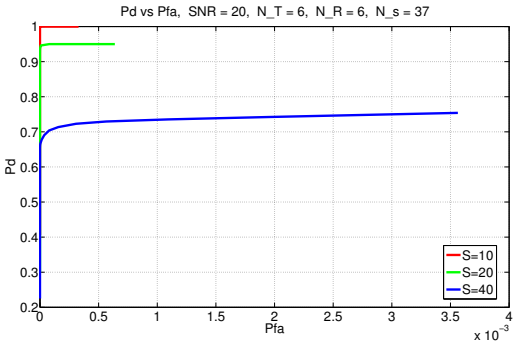


Fig. 1. MIMO, Kerdock codes, SNR=20

Acknowledgment

The authors acknowledge generous support by the National Science Foundation under grant DTRA-DMS 1042939 and by DARPA under grant N66001-11-1-4090.

References

[1] S. Becker, E. Candes, and M. Grant. Templates for convex cone problems with applications to sparse signal recovery. *Mathematical Programming Computation*, 3(3):165–218, 2011.

[2] A. R. Calderbank, P. J. Cameron, W. M. Kantor, and J. J. Seidel. Z_4-Kerdock codes, orthogonal spreads, and extremal Euclidean line-sets. *Proc. London Math. Soc. (3)*, 75(2):436–480, 1997.

[3] E.J. Candès and Y. Plan. Near-ideal model selection by ℓ_1 minimization. *Annals of Statistics*, 37(5A):2145–2177, 2009.

[4] L. Carin. On the relationship between compressive sensing and random sensor arrays. *IEEE Antennas and Propagation Magazine*, 51(5):72–81, 2009.

[5] Y. Chi, L.L. Scharf, A. Pezeshki, and A.R. Calderbank. Sensitivity to basis mismatch in compressed sensing. *IEEE Trans. Signal Processing*, 59(5):2182–2195, 2011.

[6] A. Fannjiang and W. Liao. Coherence pattern-guided compressive sensing with unresolved grids. *SIAM J. Imaging Sci.*, 5:179–202, 2012.

[7] S. Gurevich, R. Hadani, and N. Sochen. The finite harmonic oscillator and its applications to sequences, communication and radar. *IEEE Trans. Inf. Theory*, 54(9):4239–4253, 2008.

[8] M. Herman and T. Strohmer. High-resolution radar via compressed sensing. *IEEE Trans. on Signal Processing*, 57(6):2275–2284, 2009.

[9] M. Herman and T. Strohmer. General deviants: an analysis of perturbations in compressed sensing. *IEEE Journal of Selected Topics in Signal Processing: Special Issue on Compressive Sensing*, 4(2):342–349, 2010.

[10] S. D. Howard, A. R. Calderbank, and W. Moran. The finite Heisenberg-Weyl groups in radar and communications. *EURASIP J. Appl. Signal Process.*, 2006:1–12, 2006.

[11] Y. Lo. A mathematical theory of antenna arrays with randomly spaced element. *IEEE Trans. Antennas and Propagation*, 12(3):257–268, 1964.

[12] L. C. Potter, E. Ertin, J. T. Parker, and M. Cetin. Sparsity and compressed sensing in radar imaging. *Proceedings of the IEEE*, 98(6):1006–1020, 2010.

[13] T. Strohmer and B. Friedlander. Analysis of sparse MIMO radar. 2012. Preprint.

[14] T. Strohmer and H. Wang. Accurate detection of moving targets via random sensor arrays and kerdock codes. 2013. Preprint.

[15] G. Tang, B.N. Bhaskar, P. Shah, and B. Recht. Compressed sensing off the grid. *Preprint, [arvix:1207.6053]*, 2012.

[16] Zilong Wang and Guang Gong. New sequences design from weil representation with low two-dimensional correlation in both time and phase shifts. *IEEE Trans. Inf. Theory*, 57(7):4600–4611, 2011.

Phase retrieval using time and Fourier magnitude measurements

Martin Ehler
Helmholtz Zentrum München
Institute of Computational Biology
85764 Neuherberg
martin.ehler@helmholtz-muenchen.de

Stefan Kunis
University Osnabrück
Institute of Mathematics
49069 Osnabrück
and Helmholtz Zentrum München
Institute of Computational Biology
85764 Neuherberg
stefan.kunis@math.uos.de

***Abstract*—We discuss the reconstruction of a finite-dimensional signal from the absolute values of its Fourier coefficients. In many optical experiments the signal magnitude in time is also available. We combine time and frequency magnitude measurements to obtain closed reconstruction formulas. Random measurements are discussed to reduce the number of measurements.**

I. Introduction

Phase retrieval, within a discrete model, deals with the problem of reconstructing a signal $z \in \mathbb{C}^d$ from a collection of magnitudes $\{|\langle z, x_j\rangle|^2\}_{j=1}^n$, where $\{x_j\}_{j=1}^n \subset \mathbb{C}^d$ are measurement vectors. The signal z, of course, can be determined up to a global phase factor at best.

Standard algorithms are based on Gerchberg/Saxton [14] and Fienup [13] and usually involve some iterative alternate projection scheme. As phase retrieval is such a long-standing problem, it appears impossible to give a complete list of references, so let us simply refer to [7], [16], [17], [22] and references therein.

Algebraic conditions on measurement vectors have led to closed reconstruction formulas of zz^* [6], so that a singular value decomposition enables the extraction of z up to its global phase. However, such conditions can only be satisfied when the number of measurements n scales at least like d^2. Currently, reducing this number to scale linearly in d is an active field of research, see [1] for the use of graph theory. Random measurement vectors and signal recovery with high probability has been considered in [8], [9], [10]. There, the reconstruction formula is replaced with an optimization procedure based on semidefinite programming, and the number of random measurement vectors n then scales linearly in d.

Both approaches though suffer from limitations. The deterministic reconstruction formula in [6] does not apply to Fourier measurements, which arise in many optical measurement processes and appear to be the largest application field of phase retrieval. The underlying probability measure of the random measurement vectors in [8], [9], [10] has full support on the unit sphere. Thus, although only linearly many measurements have to be performed in physical experiments, any point on the sphere is a potential measurement vector and is not allowed to be excluded a-priori. It is desirable to decrease the set of potential measurement vectors to better reflect the physical constraints in actual experiments.

In this short note we shall discuss approaches to overcome the aforementioned problems and limitations. Many physical experiments additionally provide the signal power in time, i.e., $\{|z_k|\}_{k=1}^d$. By using a generalization of the algebraic condition in [6], developed in [3], we observe that certain Fourier measurements combined with the signal power in time lead to a closed reconstruction formula for zz^*. Building upon such results, we also propose specific Fourier type probability measures that may allow for signal reconstruction within the random setting. For the latter, we do not provide rigorous proofs here but collect some indications.

II. Unstructured measurements

Let $\mathbb{K}$ denote either $\mathbb{R}$ or $\mathbb{C}$. The aim of the present note is to discuss some ideas about the reconstruction of an unknown vector $x \in \mathbb{K}^d$ from a collection of magnitude measurement $\{|\langle z, x_j\rangle|^2\}_{j=1}^n$, where $\{x_j\}_{j=1}^n \subset \mathbb{K}^d$ are some measuring vectors chosen a-priori.

A. Lower bounds on the number of measurements for $\mathbb{K} = \mathbb{R}$

Although we shall concentrate on $\mathbb{K} = \mathbb{C}$ later, let us consider $\mathbb{K} = \mathbb{R}$ for a moment. If we assume that the first entry of x is nonzero, then $\{e_k\}_{k=1}^d \cup \{e_1 + e_k\}_{k=2}^d \subset \mathbb{R}^d$ is a collection of $n = 2d - 1$ measurement vectors that allow to recover x. Indeed, the first d measurements yield the absolute values of the entries of x and the following measurements enable us to check if the signs change from one coordinate to the other. The reconstruction algorithm is simple but note that we assumed the first entry of x to be nonzero. Similarly, a stable algorithm requiring $\mathcal{O}(d \log d)$ measurements, starting from $\{e_k\}_{k=1}^d$, and then determining relative phases between entries of z has been proposed in [24], [1]. If we can perform adaptive measurements, then the application of $\{e_k\}_{k=1}^d$ would tell us the location of the nonzero entries of z, say $k_1, \dots, k_\ell$. So the additional measurement vectors $\{e_{k_1} + e_{k_j}\}_{j=2}^\ell$ would enable us to recover $\pm z$ from a total of $d+\ell-1$ measurements. Without any adaptivity and knowledge on z, we shall see next that $2d - 1$ measurements are sufficient to reconstruct z up

to its sign, but there may not be any efficient reconstruction algorithms. Let us deal with the collection of matrices $\mathcal{M} := \{zz^* : z \in \mathbb{R}^d\}$ and define the map $\mathcal{F}_n : \mathcal{M} \to \mathbb{R}^n$ as

$$\mathcal{F}_n(zz^*) := \big(\operatorname{trace}(zz^* x_j x_j^*)\big)_{j=1}^n = \big(|\langle z, x_j\rangle|^2\big)_{j=1}^n. \quad (1)$$

There are $n = 2d-1$ measuring vectors $\{x_j\}_{j=1}^n$ necessary (and generically sufficient) to ensure injectivity of $\mathcal{F}_n$, cf. [6]. If we are willing to remove a set of measure zero, then the lower bound can be relaxed: There is a set $\Omega \subset \mathbb{R}^d$ of measure zero, such that any $z \in \mathbb{R}^d \setminus \Omega$ is uniquely determined up to its sign by measuring with the $d+1$ vectors $\{e_k\}_{k=1}^d \cup \{e_1 + \ldots + e_d\}$.

B. Deterministic measurements

From here on we suppose $\mathbb{K} = \mathbb{C}$ and denote $S^{d-1} = \{x \in \mathbb{K}^d : \|x\| = 1\}$. A collection $\{x_j\}_{j=1}^n \subset S^{d-1}$ with weights $\{\omega_j\}_{j=1}^n \subset \mathbb{R}_+$ is called a projective cubature of strength 2 if $\sum_{j=1}^n \omega_j = 1$ and

$$\sum_{j=1}^n \omega_j |\langle z, x_j\rangle|^4 = \frac{2}{d(d+1)}\|z\|^4, \quad \text{for all } z \in \mathbb{C}^d. \quad (2)$$

Given such a projective cubature of strength 2, the results in [3] yield that

$$zz^* = d\sum_{j=1}^n \omega_j |\langle z, x_j\rangle|^2 \big((d+1)x_j x_j^* - I\big). \quad (3)$$

Equation (3) was derived in [5] for constant weights. Therefore, any matrix zz^*, for $z \in \mathbb{C}^d$, can be reconstructed from its measurements $\{|\langle z, x_j\rangle|^2\}_{j=1}^n$.

C. Random measurements

It is well-known that any projective cubature of strength 2 must have cardinality $n \geq d^2$, cf. [2], [21]. To reduce the number of measurements, semidefinite programming and random measuring vectors were used in [9], [10], [11] to reconstruct zz^* with high probability. Indeed, let $\mathscr{H}$ denote the collection of hermitian matrices in $\mathbb{C}^{d\times d}$. For $\{x_j\}_{j=1}^n \subset \mathbb{C}^d$, we extend the operator in (1) and define

$$\mathcal{F}_n : \mathscr{H} \to \mathbb{R}^n, \quad H \mapsto \big(\operatorname{trace}(Hx_j x_j^*)\big)_{j=1}^n. \quad (4)$$

Given $b := \mathcal{F}_n(zz^*)$ and excluding the pathological case $b = 0$, we see that zz^* is a solution to

$$\min_{H\in\mathscr{H}} (\operatorname{rank}(H)), \quad \text{subject to} \quad \mathcal{F}_n(H) = b,\ H \succeq 0, \quad (5)$$

where $H \succeq 0$ stands for H being positive semidefinite. The general affine rank minimization problem is NP-hard, see for instance [19], [20], and commonly replaced by

$$\min_{H\in\mathscr{H}} (\operatorname{trace}(H)), \quad \text{subject to} \quad \mathcal{F}_n(H) = b,\ H \succeq 0, \quad (6)$$

a semidefinite program, for which efficient solvers such as interior point methods are available.Let us assume that $\{x_j\}_{j=1}^n \subset S^{d-1}$ are an independent sample from the uniform distribution on S^{d-1}. According to [9], [10], there are two constants $c, \gamma > 0$, such that, for all $n \geq cd$, the minimizer of (6) is unique and given by zz^* with probability at least $1 - e^{-\gamma n}$. The same statement holds if the entries of $\{x_j\}_{j=1}^n \subset \mathbb{C}^d$ are chosen independently from the standard Gaussian distribution.

The proof in [9], [10], see also [3], for the uniform distribution on the sphere is based on the probabilistic reconstruction formula

$$zz^* = d\,\mathbb{E}|\langle z, X\rangle|^2\big((d+1)XX^* - I\big), \quad \text{(PRF-1)}$$

for all $z \in \mathbb{C}^d$, where $X \in \mathbb{C}^d$ denotes a random vector uniformly distributed on S^{d-1}. In view of (2), we observe that (PRF-1) is equivalent to

$$\mathbb{E}|\langle z, X\rangle|^4 = \frac{2}{d(d+1)}\|z\|^4, \quad \text{for all } z \in \mathbb{C}^d, \quad (7)$$

which implies

$$d\,\mathbb{E}|\langle z, X\rangle|^2 = \|z\|^2, \quad \text{for all } z \in \mathbb{C}^d, \quad (8)$$

cf. [4]. The condition (8) is equivalent to

$$d\,\mathbb{E}XX^* = I, \quad (9)$$

so that (PRF-1) implies (9). The proof of the equivalence between (5) and (6) is based on (PRF-1) and (9), and, besides some technical ingredients, then turns the expectation in both conditions into suitable statements on the sample mean by using tail bound estimates, cf. [3], [10].

III. Time-frequency structured measurements

A. Fourier measurements

The measuring vectors in the previous section were either unstructured or chosen from the uniform distribution on the sphere. In optical experiments, Fourier type measurements are performed. Naturally, we consider the random Fourier vector

$$X = \tfrac{1}{\sqrt{d}}(e^{2\pi i\lambda_1 t}, \ldots, e^{2\pi i\lambda_d t})^\top, \quad (10)$$

where $\{\lambda_i\}_{i=1}^d$ are real numbers and t is a random variable uniformly distributed on $[0, 1)$. Of course, the measurements

$$\langle z, x_j\rangle = \tfrac{1}{\sqrt{d}}\sum_{k=1}^d z_k e^{-2\pi i\lambda_k t_j}$$

consist of a randomly sampled trigonometric polynomial, which brings all sorts of nonequispaced fast Fourier transforms into play. Unfortunately and to no surprise, the vector X is not uniformly distributed on S^{d-1}. Nevertheless, we could check if (7) holds, and if so, then there might be a good chance that the trace minimization works out numerically although not stringently proven mathematically yet. Unfortunately, Fourier magnitude measurements alone are not sufficient to resolve time translates. For instance, the canonical basis vectors $e_1, \ldots, e_d$ cannot be distinguished by the absolute values of their Fourier coefficients. Thus, (PRF-1) is violated:

Proposition III.1. *Let $\{\lambda_k\}_{k=1}^d$ be a sequence of real numbers. If $t : \Omega \to [0, 1)$ is a random variable, then the Fourier random vector* (10) *does not satisfy* (PRF-1).

The same, of course, holds for the deterministic setting:

Proposition III.2. *If $\{\lambda_k\}_{k=1}^d$ and $\{t_j\}_{j=1}^n$ are sequences of real numbers, then there are no weights $\{\omega_j\}_{j=1}^n$ such that the Fourier vectors*

$$\{x_j\}_{j=1}^n = \{\tfrac{1}{\sqrt{d}}(e^{2\pi i\lambda_1 t_j},\dots,e^{2\pi i\lambda_d t_j})^\top\}_{j=1}^n$$

satisfy (3).

B. Additional time measurements

To resolve time translates we must perform additional measurements beyond the Fourier spectrum. In optical experiments the magnitudes in time are often available as well. The latter results in additional measurement vectors $\{e_k\}_{k=1}^d$. We shall discuss three scenarios:

1) Deterministic time-frequency measurements: First, we combine special Fourier vectors with time measurements, inspired by ideas in [15, Proposition 4] and [23, Section 2.1.2]. Let q be a prime and let $d = q^r + 1$ for some $r \in \mathbb{N}$. For $m = d^2 - d + 1$, there exist integers $0 \le \lambda_1 < \cdots < \lambda_d < m$ such that all numbers $1,\dots,m-1$ occur as residues $\bmod\ m$ of the $d(d-1)$ differences $(\lambda_k - \lambda_\ell)$, for $k \ne \ell$, cf. [15]. For $j = 1,\dots,m$ we define the Fourier vectors

$$x_j = \frac{1}{\sqrt{d}}(e^{2\pi i\lambda_1 j/m},\dots,e^{2\pi i\lambda_d j/m})^\top \in \mathbb{C}^d. \qquad (11)$$

To add time measurements, we form the set $\mathcal{X} = \{x_j\}_{j=1}^m \cup \{e_k\}_{k=1}^d$ with weights $\mathcal{W} = \{\frac{d}{d^3+1}\}_{j=1}^m \cup \{\frac{1}{d(d+1)}\}_{i=1}^d$, respectively. Note that $\mathcal{X}$ is a projective cubature of strength 2, cf. [15], and, therefore, satisfies (3). Its cardinality is $n = d^2 + 1$, the weights split into two groups of constants, and they become almost equal for large ambient dimensions d. The set $\mathcal{X}$ models special Fourier and time measurements, hence, forms a highly structured collection of measurement vectors. In contrast to a naive evaluation of the reconstruction formula (3) in $\mathcal{O}(d^4)$, this allows for a computation in only $\mathcal{O}(d^3 \log d)$ arithmetic operations.

Example III.3. For $d = 4 = 3^1 + 1$, we can select $(\lambda_1,\lambda_2,\lambda_3,\lambda_4) = (0,3,5,12)$ for the above scheme which yields a projective cubature of strength 2 whose cardinality is $n = 17$.

2) Random time-frequency measurements: Let μ_1 denote the discrete probability measure with support $\mathcal{X}$ and mass distribution according to the weights $\mathcal{W}$. Any random vector $X_1 \sim \mu_1$ satisfies (PRF-1). Therefore, (9) is also satisfied. As a first step for the proof about equivalence of the optimization problems (5) and (6), we can turn (9) into a suitable statement on the sample mean. Indeed, let $\{x_j\}_{j=1}^n$ be independent and identical distributed according to μ_1 and $0 < s < 1$ arbitrary. The Chernoff's matrix inequalities yield that there exist constants $c, C > 0$ such that, for all $n \ge cd\log(d)$,

$$\|\frac{d}{n}\sum_{j=1}^n x_j x_j^* - I\| \le s \qquad (12)$$

holds with probability at least $1 - e^{-Cn/d}$, cf. [12]. The estimate (12) turns the identity about the population mean (9) into an estimate on the deviation of the sample mean measured by the operator norm. It is just a first step, and to derive a complete mathematical proof of the equivalence between (5) and (6), we additionally need a suitable sample mean version of (PRF-1) and few more technical ingredients that would go beyond the present note. Here, we understand the above observations as an indication that the proof can be completed.

3) Deterministic time and random frequency measurements: Switching from the deterministic to the random setting avoids the requirement of d^2 many measurements. We are still consistent with this objective when choosing d measurements in a deterministic fashion and on the order of d many additional random measurements. Matching experimental setups, we propose to keep the time measurements $\{e_k\}_{k=1}^d$ as deterministic information and randomly select samples from the random vector

$$X = \tfrac{1}{\sqrt{d}}(e^{2\pi i\lambda_1 t},\dots,e^{2\pi i\lambda_d t})^\top,$$

where t is uniformly distributed on $[0,1)$ and $\{\lambda_j\}_{j=1}^d$ is a Golomb ruler, i.e, a set of integers whose pairwise differences $\lambda_k - \lambda_\ell$, $k \ne \ell$ are all distinct. Then one can verify that, for all $z \in \mathbb{C}^d$,

$$zz^* = d^2\mathbb{E}|\langle z, X\rangle|^2 XX^* + \sum_{k=1}^d |\langle z, e_k\rangle|^2(e_k e_k^* - I) \quad \text{(PRF-2)}$$

holds. Note that (PRF-2) is the analogue of (PRF-1). The requirements on $\{\lambda_k\}_{k=1}^d$ can be satisfied for special values d as above and for any d, for instance, by choosing $\lambda_k := d(k-1)^2 + k - 1$, $k = 1,\dots,d$, cf. [18] and references therein. Thus, the maximal frequency (the length of the Golomb ruler) can be chosen smaller than d^3. A simple counting argument yields that it must be bigger than $\frac{1}{2}d(d-1)$, and it is conjectured that, for any $d > 0$, one can find a Golomb ruler with length less than d^2.

Using the semidefinite program (6) for the last two scenarios in particular asks for the iterated evaluation of $\mathcal{F}_n(H)$. Assuming moreover that only $n = \mathcal{O}(d\log d)$ measurement vectors $x_j \in \mathbb{C}^d$ suffice for reconstruction with high probability, a naive evaluation of $\mathcal{F}_n(H) = (\mathrm{trace}(Hx_j x_j^*))_{j=1}^n$ requires $\mathcal{O}(d^3\log d)$ floating point operations. Applying fast Fourier transforms, tailored to the indices $\lambda_k \in \mathbb{Z}$, $k = 1,\dots,d$, we expect a reduction to preferably $\mathcal{O}(d^2\log^2 d)$ floating point operations for one application of the map $\mathcal{F}_n$.

IV. Discussion and conclusion

The deterministic time-frequency measurements yield a closed reconstruction formula. However, this formula is only available in certain dimensions and the number of measurements is $d^2 + 1$. This number can be reduced by switching to the random setting, in which we proposed to select time-frequency measurements through a discrete probability measure with mass distributed according to the proposed deterministic measurement process. There is still the restriction to certain special dimensions. To overcome such limitations, we propose a hybrid model in which Fourier measurements

are performed randomly and time measurements are added in a deterministic fashion. The latter may also better match the experimental measurement setting. When the associated Fourier vectors are based on Golomb rulers, then the key ingredient (PRF-2) for a proof that the semidefinite program recovers the correct signal is satisfied. Therefore, we have strong indication that the rigorous mathematical proof can be derived.

ACKNOWLEDGMENT

M. E. is supported by the NIH/DFG Research Career Transition Awards Program (EH 405/1-1/575910). S. K. is supported by the DFG project KU 2557/1-2 and by the Helmholtz young investigator group VH-NG-526.

REFERENCES

[1] B. Alexeev, A. S. Bandeira, M. Fickus, and D. G. Mixon, "Phase retrieval with polarization," *arXiv:1210.7752*.

[2] C. Bachoc, E. Bannai, and R. Coulangeon, "Codes and designs in Grassmannian spaces," *Discrete Mathematics*, vol. 277, pp. 15–28, 2004.

[3] C. Bachoc and M. Ehler, "Signal reconstruction from the magnitude of subspace components," *arXiv:1209.5986*, 2012.

[4] ——, "Tight p-fusion frames," *Appl. Comput. Harmon. Anal.*, vol. 35, no. 1, pp. 1–15, 2013.

[5] R. Balan, B. G. Bodmann, P. G. Casazza, and D. Edidin, "Painless reconstruction from magnitudes of frame coefficients," *J. Fourier Anal. Appl.*, vol. 15, no. 4, pp. 488–501, 2009.

[6] R. Balan, P. Casazza, and D. Edidin, "On signal reconstruction without phase," *Appl. Comput. Harmon. Anal*, vol. 20, pp. 345–356, 2006.

[7] H. H. Bauschke, P. L. Combettes, and D. R. Luke, "Phase retrieval, error reduction algorithm, and Fienup variants: A view from convex optimization," *J. Opt. Soc. Amer. A*, vol. 19, pp. 1334–1345, 2002.

[8] E. J. Candès, Y. Eldar, T. Strohmer, and V. Voroninski, "Phase retrieval via matrix completion," *SIAM J. Imaging Sci.*, vol. 6, no. 1, pp. 199–225, 2013.

[9] E. J. Candès and X. Li., "Solving quadratic equations via PhaseLift when there are about as many equations as unknowns," *arXiv:1208.6247*, 2012.

[10] E. J. Candès, T. Strohmer, and V. Voroninski, "PhaseLift: Exact and stable signal recovery from magnitude measurements via convex programming," *Comm. Pure Appl. Math., DOI:10.1002/cpa.21432*, 2012.

[11] L. Demanet and P. Hand, "Stable optimizationless recovery from phaseless linear measurements," *arXiv:1208.1803*, 2012.

[12] M. Ehler, "On tight generalized frames," *arXiv:1305.0716*, 2013.

[13] J. R. Fienup, "Phase retrieval algorithms: a comparison," *Appl. Optics*, vol. 21, no. 15, pp. 2758–2769, 1982.

[14] R. W. Gerchberg and W. O. Saxton, "A practical algorithm for the determination of the phase from image and diffraction plane pictures," *Optik*, vol. 35, no. 2, pp. 237–246, 1972.

[15] H. König, "Cubature formulas on spheres," *Adv. Multivar. Approx. Math. Res.*, vol. 107, pp. 201–211, 1999.

[16] D. Langemann and M. Tasche, "Phase reconstruction by a multilevel iteratively regularized Gauss–Newton method," *Inverse Problems*, vol. 24, no. 3, 2008.

[17] ——, "Multilevel phase reconstruction for a rapidly decreasing interpolating function," *Results Math.*, vol. 53, no. 3-4, pp. 333–340, 2009.

[18] C. Meyera and P. A. Papakonstantinou, "On the complexity of constructing Golomb rulers," *Discrete Appl. Math.*, vol. 157, no. 4, pp. 738–748, 2009.

[19] B. Natarajan, "Sparse Approximate Solutions to Linear Systems," *SIAM J. Comput.*, vol. 24, no. 2, pp. 227–234, 1995.

[20] B. Recht, M. Fazel, and P. Parillo, "Guaranteed minimum rank solutions to linear matrix equations via nuclear norm minimization," *SIAM Rev.*, vol. 52, no. 3, pp. 471–501, 2010.

[21] A. Roy, "Bounds for codes and designs in complex subspaces," *J. Algebr. Comb.*, vol. 31, no. 1, pp. 1–32, 2010.

[22] B. Seifert, H. Stolz, M. Donatelli, D. Langemann, and M. Tasche, "Multilevel Gauss-Newton methods for phase retrieval problems," *J. Phys. A*, vol. 39, pp. 4191–4206, 2006.

[23] T. Strohmer and R. W. Heath, "Grassmannian frames with applications to coding and communication," *Appl. Comput. Harmon. Anal.*, vol. 14, no. 3, pp. 257–275, May 2003.

[24] Y. Wang, "A frame construction for fast phase retrieval," (joint work with M. Fickus), talk at ESI, Modern Methods of Time-Frequency Analysis II, Phase Retrieval.

Fast Ewald summation under 2d- and 1d-periodic boundary conditions based on NFFTs

Franziska Nestler
Chemnitz University of Technology
Faculty of Mathematics
09107 Chemnitz, Germany
Email: franziska.nestler@mathematik.tu-chemnitz.de

Daniel Potts
Chemnitz University of Technology
Faculty of Mathematics
09107 Chemnitz, Germany
Email: potts@mathematik.tu-chemnitz.de

***Abstract*—Ewald summation has established as basic element of fast algorithms evaluating the Coulomb interaction energy of charged particle systems in three dimensions subject to periodic boundary conditions. In this context particle mesh routines, as the P3M method, and the P2NFFT, which is based on nonequispaced fast Fourier transforms (NFFT), should be mentioned. These methods treat the problem efficiently in case that periodic boundary conditions in all three dimensions are assumed. In this paper we present a new approach for the efficient calculation of the Coulomb interaction energy subject to mixed boundary conditions based on NFFTs.**

I. Introduction

Let a set of N charges $q_j \in \mathbb{R}$ at positions $\boldsymbol{x}_j \in \mathbb{R}^3$, $j = 1, \dots, N$, be given. Throughout this paper we assume that the system is electrical neutral, i.e., $\sum_{j=1}^{N} q_j = 0$. The electrostatic energy of the particle system is basically a sum of the form

$$E(\mathcal{S}) := \frac{1}{2} \sum_{i,j=1}^{N} {\sum_{\boldsymbol{n} \in S}}' \frac{q_i q_j}{\|\boldsymbol{x}_i - \boldsymbol{x}_j + B\boldsymbol{n}\|}, \tag{1}$$

where $\mathcal{S} \subseteq \mathbb{Z}^3$ is set according to the given boundary conditions and $B \in \mathbb{R}$ is the edge length of the periodically duplicated simulation box. The prime on the second sum indicates that in the case $\boldsymbol{n} = \boldsymbol{0}$ the terms for $i = j$ are omitted.

If periodic boundary conditions are applied in all three dimensions, the particle positions $\boldsymbol{x}_j$ are commonly assumed to be distributed in a cubic box, i.e., $\boldsymbol{x}_j \in B\mathbb{T}^3$ for some $B > 0$, and $\mathcal{S} := \mathbb{Z}^3$. We thereby define the torus $\mathbb{T} := \mathbb{R}/\mathbb{Z} \simeq [-1/2, 1/2)$. In some applications periodic boundary conditions are assumed in two or one dimension only, where we choose $\mathcal{S} := \mathbb{Z}^2 \times \{0\}$ with $\boldsymbol{x}_j \in B\mathbb{T}^2 \times \mathbb{R}$ and $\mathcal{S} := \mathbb{Z} \times \{0\}^2$ with $\boldsymbol{x}_j \in B\mathbb{T} \times \mathbb{R}^2$, respectively.

It is important to note that in all three cases the infinite sum (1) is only conditionally convergent, i.e., the value of the energy is not well defined unless a precise order of summation is specified.

The well known Ewald summation formulas, which have at first been derived for the fully periodic case, cf. [1], are the principle behind many fast algorithms evaluating the energy (1). The Ewald method is based on the trivial identity

$$\frac{1}{r} = \frac{\operatorname{erf}(\alpha r)}{r} + \frac{\operatorname{erfc}(\alpha r)}{r}, \tag{2}$$

where $\alpha > 0$, $\operatorname{erf}(x) := \frac{2}{\sqrt{\pi}} \int_0^x \mathrm{e}^{-t^2} \mathrm{d}t$ is the well known error function and $\operatorname{erfc}(x) := 1 - \operatorname{erf}(x)$ is the complementary error function. If (2) is applied to (1), the poorly converging sum is split into two exponentially converging parts. The first infinite sum, including the erfc-terms, is short ranged and absolutely convergent in spatial domain. Taking a specific summation order into account and exploiting the charge neutrality, the second sum, which is still long ranged, can be transformed into a rapidly converging sum in frequency domain. Usually, the energy (1) is defined over a spherical order of summation, see [4] for a detailed derivation for the fully periodic case. The Ewald summation formulas for 2d- and 1d-periodic boundary conditions are derived in [2] and [3], respectively.

In the fully periodic case, the Ewald method has the complexity $\mathcal{O}(N^{3/2})$ if the splitting parameter α is chosen appropriately. However, the computational effort can be reduced to $\mathcal{O}(N \log N)$ arithmetic operations by evaluating the long range part efficiently using Fast Fourier transforms (FFT). For this purpose, the problem has to be modified in a way that the FFT as a grid transformation can be used. This discretization is performed by replacing the charges q_j by a grid based charge density. This is the basic idea behind Particle Mesh approaches such as the P3M method, see [5] to get an overview over some of these techniques. The same principle is used in the P2NFFT method [6], which is based on nonequispaced fast Fourier transforms (NFFT). Here the discretization process is part of the NFFT algorithms.

For open boundary conditions, i.e., $\mathcal{S} := \{0\}^3$ in (1), fast summation methods [8], [9] based on NFFTs where suggested, too. In this note we aim to close the gap and propose FFT based algorithms also for 2d- and 1d-periodic boundary conditions.

We remark that the fast multipole method can also handle all boundary conditions very efficiently, see [10].

The outline of this paper is as follows. We start with a short introduction to the NFFT. Thereafter we consecutively consider the problem of evaluating (1) subject to 2d- and 1d-periodic boundary conditions. In each case we consider at first the according Ewald summation formula and then present a new approach for the efficient calculation of the Coulomb interaction energy (1) based on NFFTs.

To keep the notation short we define the difference vectors

$\boldsymbol{x}_{ij} := \boldsymbol{x}_i - \boldsymbol{x}_j$. For some $\boldsymbol{M} \in 2\mathbb{N}^d$ we refer to $\mathcal{I}_{\boldsymbol{M}}$ as the index set given by

$$\mathcal{I}_{\boldsymbol{M}} := \left\{ \left[-\tfrac{M_1}{2}, \tfrac{M_1}{2}\right) \times \cdots \times \left[-\tfrac{M_d}{2}, \tfrac{M_d}{2}\right) \right\} \cap \mathbb{Z}^d.$$

Throughout this paper we do not distinguish between row and column vectors and denote by $\boldsymbol{x} \cdot \boldsymbol{y} := x_1 y_1 + \ldots + x_d y_d$ the scalar product and by $\boldsymbol{x} \odot \boldsymbol{y} := (x_1 y_1, \ldots, x_d y_d) \in \mathbb{R}^d$ the component wise product of two vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$. For some $\boldsymbol{x} \in \mathbb{R}^d$ with non-vanishing components we further define the vector $\boldsymbol{x}^{-1} \in \mathbb{R}^d$ by $\boldsymbol{x}^{-1} := (x_1^{-1}, \ldots, x_d^{-1})$.

II. Nonequispaced Fast Fourier Transforms

Let $\boldsymbol{M} \in 2\mathbb{N}^d$, the index set $\mathcal{I}_{\boldsymbol{M}}$ and the coefficients $\hat{f}_{\boldsymbol{k}} \in \mathbb{C}$ for $\boldsymbol{k} \in \mathcal{I}_{\boldsymbol{M}}$ be given. The fast evaluation of a trigonometric polynomial

$$f(\boldsymbol{x}) := \sum_{\boldsymbol{k} \in \mathcal{I}_{\boldsymbol{M}}} \hat{f}_{\boldsymbol{k}} \mathrm{e}^{-2\pi \mathrm{i} \boldsymbol{k} \cdot \boldsymbol{x}}$$

at $N \in \mathbb{N}$ given nodes $\boldsymbol{x}_j \in \mathbb{T}^d$, i.e., the fast computation of $f_j := f(\boldsymbol{x}_j)$, $j = 1, \ldots, N$, is known as d-dimensional NFFT. The algorithm uses an approximation of f in the form

$$f(\boldsymbol{x}) \approx \sum_{\boldsymbol{l} \in \mathcal{I}_{\boldsymbol{m}}} g_{\boldsymbol{l}} \tilde{\varphi}(\boldsymbol{x} - \boldsymbol{l} \odot \boldsymbol{m}^{-1}), \tag{3}$$

where $\tilde{\varphi}$ is a multivariate 1-periodic function, which is well localized in spatial and frequency domain, and $\boldsymbol{m} \in 2\mathbb{N}^d$ with $\boldsymbol{m} > \boldsymbol{M}$ (component wise). It can be shown that it is reasonable to set, cf. [7],

$$g_{\boldsymbol{l}} := \frac{1}{|\mathcal{I}_{\boldsymbol{m}}|} \sum_{\boldsymbol{k} \in \mathcal{I}_{\boldsymbol{M}}} \frac{\hat{f}_{\boldsymbol{k}}}{c_{\boldsymbol{k}}(\tilde{\varphi})} \mathrm{e}^{2\pi \mathrm{i} \boldsymbol{k} \cdot (\boldsymbol{l} \odot \boldsymbol{m}^{-1})},$$

where $c_{\boldsymbol{k}}(\tilde{\varphi})$ denotes the Fourier coefficient with index $\boldsymbol{k}$ of $\tilde{\varphi}$. Obviously, the coefficients $g_{\boldsymbol{l}} \in \mathbb{C}$ can be calculated using the FFT. The function values $f(\boldsymbol{x}_j)$ are then computed via (3), where the sums can be truncated due to the good localization of $\tilde{\varphi}$ in spatial domain.

Correspondingly, the adjoint NFFT is an algorithm for the efficient calculation of

$$\hat{h}_{\boldsymbol{k}} := \sum_{j=1}^{N} f_j \mathrm{e}^{2\pi \mathrm{i} \boldsymbol{k} \cdot \boldsymbol{x}_j}, \quad \boldsymbol{k} \in \mathcal{I}_{\boldsymbol{M}},$$

for N given nodes $\boldsymbol{x}_j \in \mathbb{T}^d$ and coefficients $f_j \in \mathbb{C}$, $j = 1, \ldots, N$. The resulting algorithm has a very similar structure and the same arithmetic complexity of $\mathcal{O}(|\mathcal{I}_{\boldsymbol{M}}| \log |\mathcal{I}_{\boldsymbol{M}}| + N)$, see [7] for instance. In this reference different choices for the window function $\tilde{\varphi}$ are discussed, too.

III. 2d-periodic systems

For N charges q_j at positions $\boldsymbol{x}_j = (x_{j,1}, x_{j,2}, x_{j,3}) \in B\mathbb{T}^2 \times \mathbb{R}$, $j = 1, \ldots, N$, we define the electrostatic energy subject to periodic boundary conditions in the first two dimensions by $E^{\mathrm{p2}} := E(\mathbb{Z}^2 \times \{0\})$.

At first we review the corresponding Ewald formula, as derived in [2], and then present an NFFT approach for the fast calculation of the energy E^{p2}. In this section we refer to $\tilde{\boldsymbol{x}} := (x_1, x_2) \in B\mathbb{T}^2$ as the vector of the first two components of some $\boldsymbol{x} \in B\mathbb{T}^2 \times \mathbb{R}$.

A. Ewald Formula

If a spherical order of summation is applied, the electrostatic energy E^{p2} can be written in the form, cf. [2],

$$E^{\mathrm{p2}} = E^{\mathrm{p2,S}} + E^{\mathrm{p2,L}} + E^{\mathrm{p2},\boldsymbol{0}} + E^{\mathrm{p2,self}}, \tag{4}$$

where for some $\alpha > 0$

$$E^{\mathrm{p2,S}} := \frac{1}{2} \sum_{\boldsymbol{n} \in \mathbb{Z}^2 \times \{0\}} {\sum_{i,j=1}^{N}}' q_i q_j \frac{\operatorname{erfc}(\alpha \|\boldsymbol{x}_{ij} + B\boldsymbol{n}\|)}{\|\boldsymbol{x}_{ij} + B\boldsymbol{n}\|}$$

$$E^{\mathrm{p2,L}} := \frac{1}{4B} \sum_{\boldsymbol{k} \in \mathbb{Z}^2 \setminus \{\boldsymbol{0}\}} \sum_{i,j=1}^{N} q_i q_j \, \mathrm{e}^{2\pi \mathrm{i} \boldsymbol{k} \cdot \tilde{\boldsymbol{x}}_{ij}/B} \Theta^{\mathrm{p2}}(\|\boldsymbol{k}\|, x_{ij,3})$$

$$E^{\mathrm{p2},\boldsymbol{0}} := -\frac{\sqrt{\pi}}{B^2} \sum_{i,j=1}^{N} q_i q_j \Theta_0^{\mathrm{p2}}(x_{ij,3})$$

$$E^{\mathrm{p2,self}} := -\frac{\alpha}{\sqrt{\pi}} \sum_{j=1}^{N} q_j^2.$$

We thereby define the functions Θ_0^{p2} and Θ^{p2} by

$$\Theta_0^{\mathrm{p2}}(r) := \frac{\mathrm{e}^{-\alpha^2 r^2}}{\alpha} + \sqrt{\pi} r \operatorname{erf}(\alpha r)$$

and

$$\Theta^{\mathrm{p2}}(k, r) := \frac{\Psi(k, r) + \Psi(k, -r)}{k},$$

where we set

$$\Psi(k, r) := \mathrm{e}^{2\pi k r / B} \operatorname{erfc}\left(\tfrac{\pi k}{\alpha B} + \alpha r\right).$$

We immediately see that $\Theta_0^{\mathrm{p2}} \in \mathrm{C}^\infty(\mathbb{R})$ as well as $\Theta^{\mathrm{p2}}(k, \cdot) \in \mathrm{C}^\infty(\mathbb{R})$ for each $k \neq 0$.

Lemma 1. *For arbitrary $r \in \mathbb{R}$ we have $\Theta^{\mathrm{p2}}(k, r) \to 0$ with $\Theta^{\mathrm{p2}}(k, r) \sim k^{-2} \mathrm{e}^{-k^2}$ for $k \to \infty$.*

Proof: The function Θ^{p2} has the integral representation

$$\Theta^{\mathrm{p2}}(k, r) = \frac{4\sqrt{\pi}}{B} \int_0^\alpha \frac{1}{t^2} \exp\left(-\frac{\pi^2 k^2}{B^2 t^2} - r^2 t^2\right) \mathrm{d}t,$$

cf. [11, number 7.4.33]. We now easily see

$$\Theta^{\mathrm{p2}}(k, r) \leq \Theta^{\mathrm{p2}}(k, 0) = \frac{2}{k} \operatorname{erfc}\left(\frac{\pi k}{\alpha B}\right) \approx \frac{2\alpha B}{k^2 \pi^{3/2}} \mathrm{e}^{-\frac{\pi^2 k^2}{\alpha^2 B^2}},$$

which is valid for large k, cf. [11, number 7.1.23]. □

B. An NFFT approach

The infinite sum in $E^{\mathrm{p2,S}}$ is short ranged and can be computed by direct evaluation. Due to Lemma 1 the infinite sum in $E^{\mathrm{p2,L}}$ can be truncated, i.e., we can replace $\mathbb{Z}^2$ by $\mathcal{I}_{\boldsymbol{M}}$ for some appropriate $\boldsymbol{M} \in 2\mathbb{N}^2$.

In the following we choose $h > 0$ and $\varepsilon > 0$ such that $|x_{ij,3}| \leq h(1/2 - \varepsilon)$ for all $i, j = 1, \ldots, N$. In order to compute the far field $E^{\mathrm{p2,L}} + E^{\mathrm{p2},\boldsymbol{0}}$ efficiently we employ the idea of NFFT based fast summation methods [8] and consider the regularization

$$K_{\mathrm{R}}(k, r) := \begin{cases} \frac{1}{4B} \Theta^{\mathrm{p2}}(k, r) & : k \neq 0, |h^{-1} r| \leq 1/2 - \varepsilon \\ \frac{-\sqrt{\pi}}{B^2} \Theta_0^{\mathrm{p2}}(r) & : k = 0, |h^{-1} r| \leq 1/2 - \varepsilon \\ K_{\mathrm{B}}(k, r) & : |h^{-1} r| \in (1/2 - \varepsilon, 1/2] \end{cases},$$

where for each $k \in \{\|\boldsymbol{k}\| : \boldsymbol{k} \in \mathcal{I}_{\boldsymbol{M}}\}$ the function $K_{\mathrm{B}}(k,\cdot)$ is defined such that $K_{\mathrm{R}}(k,\cdot)$ is in the space $\mathrm{C}^p(h\mathbb{T})$ for some $p \in \mathbb{N}$ large enough, i.e., $K_{\mathrm{B}}(k,\cdot)$ fulfills the conditions

$$\begin{aligned} K_{\mathrm{B}}^{(n)}(k, {}^{h}/_{2} - h\varepsilon) &= K_{\mathrm{R}}^{(n)}(k, {}^{h}/_{2} - h\varepsilon) \\ K_{\mathrm{B}}^{(n)}(k, -{}^{h}/_{2} + h\varepsilon) &= K_{\mathrm{R}}^{(n)}(k, -{}^{h}/_{2} + h\varepsilon) \\ &= (-1)^n K_{\mathrm{R}}^{(n)}(k, {}^{h}/_{2} - h\varepsilon) \end{aligned}$$

for all $n = 0, \dots, p$ and is chosen such that

$$K_{\mathrm{R}}^{(n)}(k, {}^{h}/_{2}) = K_{\mathrm{R}}^{(n)}(k, -{}^{h}/_{2}) \quad \forall\, n = 0, \dots, p$$

is satisfied, too. The order p can be chosen arbitrarily large as the functions Θ_0^{p2} and $\Theta^{\mathrm{p2}}(k,\cdot)$ are differentiable for all degrees of differentiation. The resulting functions $K_{\mathrm{R}}(k,\cdot)$ then are h-periodic and smooth. Thus we can find good approximations of the form

$$K_{\mathrm{R}}(k,r) \approx \sum_{l \in \mathcal{I}_{M_3}} b_{k,l} \mathrm{e}^{2\pi \mathrm{i} l r/h}$$

with $M_3 \in 2\mathbb{N}$ large enough and the Fourier coefficients

$$b_{k,l} := \frac{1}{M_3} \sum_{j \in \mathcal{I}_{M_3}} K_{\mathrm{R}}\left(k, {}^{jh}/_{M_3}\right) \mathrm{e}^{-2\pi \mathrm{i} j l/M_3}.$$

With $\boldsymbol{M}^* := (\boldsymbol{M}, M_3) \in 2\mathbb{N}^3$ we obtain

$$\begin{aligned} E^{\mathrm{p2,L}} + E^{\mathrm{p2},\mathbf{0}} &\approx \sum_{\boldsymbol{k} \in \mathcal{I}_{\boldsymbol{M}}} \sum_{l \in \mathcal{I}_{M_3}} b_{\|\boldsymbol{k}\|,l} \sum_{i,j=1}^{N} q_i q_j \mathrm{e}^{2\pi \mathrm{i} \boldsymbol{v}_{\boldsymbol{k},l} \cdot \boldsymbol{x}_{ij}} \\ &= \sum_{(\boldsymbol{k},l) \in \mathcal{I}_{\boldsymbol{M}^*}} b_{\|\boldsymbol{k}\|,l} |S(\boldsymbol{k},l)|^2, \end{aligned} \tag{5}$$

where we define

$$\boldsymbol{v}_{\boldsymbol{k},l} := \begin{pmatrix} \boldsymbol{k}/B \\ l/h \end{pmatrix} \text{ as well as } S(\boldsymbol{k},l) := \sum_{j=1}^{N} q_j \mathrm{e}^{2\pi \mathrm{i} \boldsymbol{v}_{\boldsymbol{k},l} \cdot \boldsymbol{x}_j}.$$

Obviously, the sums $S(\boldsymbol{k},l)$, $(\boldsymbol{k},l) \in \mathcal{I}_{\boldsymbol{M}^*}$, can efficiently be computed by a trivariate adjoint NFFT.

Remark 1. *The energy E^{p2} can also be written in the form $E^{\mathrm{p2}} = \frac{1}{2}\sum_{j=1}^{N} q_j \phi^{\mathrm{p2}}(\boldsymbol{x}_j)$, where for each $\boldsymbol{x}_j$ the potential $\phi^{\mathrm{p2}}(\boldsymbol{x}_j)$ is defined by*

$$\phi^{\mathrm{p2}}(\boldsymbol{x}_j) := \sum_{\boldsymbol{n} \in \mathbb{Z}^2 \times \{0\}} {\sum_{i=1}^{N}}' \frac{q_i}{\|\boldsymbol{x}_{ij} + B\boldsymbol{n}\|}.$$

The term $q_j \phi^{\mathrm{p2}}(\boldsymbol{x}_j)$ then represents the energy of the single particle j. It is easy to see that we can write

$$\phi^{\mathrm{p2}}(\boldsymbol{x}_j) = \phi^{\mathrm{p2,S}}(\boldsymbol{x}_j) + \phi^{\mathrm{p2,L}}(\boldsymbol{x}_j) + \phi^{\mathrm{p2},\mathbf{0}}(\boldsymbol{x}_j) + \phi^{\mathrm{p2,self}}(\boldsymbol{x}_j),$$

according to (4)*. By* (5) *we find that the long range part $\phi^{\mathrm{p2,L}}(\boldsymbol{x}_j) + \phi^{\mathrm{p2},\mathbf{0}}(\boldsymbol{x}_j)$ can be approximated by*

$$2 \sum_{(\boldsymbol{k},l) \in \mathcal{I}_{\boldsymbol{M}^*}} b_{\|\boldsymbol{k}\|,l} S(\boldsymbol{k},l) \mathrm{e}^{-2\pi i \boldsymbol{v}_{\boldsymbol{k},l} \cdot \boldsymbol{x}_j}.$$

Having calculated the sums $S(\boldsymbol{k},l)$ the long range parts of the potentials $\phi^{\mathrm{p2}}(\boldsymbol{x}_j)$, $j = 1, \dots, N$, can be computed by a trivariate NFFT. Note that computing this additional NFFT is not necessary if only the total energy E^{p2} is of interest.

IV. 1D-PERIODIC SYSTEMS

For N charges q_j at positions $\boldsymbol{x}_j \in B\mathbb{T} \times \mathbb{R}^2$, $j = 1, \dots, N$, we denote by $E^{\mathrm{p1}} := E(\mathbb{Z} \times \{0\}^2)$ the electrostatic energy (1) subject to periodic boundary conditions in the first dimension.

In this section we refer to $\tilde{\boldsymbol{x}} := (x_2, x_3) \in \mathbb{R}^2$ as the vector of the last two components of some $\boldsymbol{x} \in B\mathbb{T} \times \mathbb{R}^2$. Furthermore we define by

$$\Gamma(s,x) := \int_x^\infty t^{s-1} \mathrm{e}^{-t} \mathrm{d}t$$

the upper incomplete gamma function and by γ the Euler constant.

A. Ewald formula

The Ewald summation formula for the electrostatic energy E^{p1} reads as, cf. [3],

$$E^{\mathrm{p1}} = E^{\mathrm{p1,S}} + E^{\mathrm{p1,L}} + E^{\mathrm{p1},0} + E^{\mathrm{p1,self}},$$

where

$$\begin{aligned} E^{\mathrm{p1,S}} &:= \frac{1}{2} \sum_{\boldsymbol{n} \in \mathbb{Z} \times \{0\}^2} {\sum_{i,j=1}^{N}}' q_i q_j \frac{\operatorname{erfc}(\alpha \|\boldsymbol{x}_{ij} + B\boldsymbol{n}\|)}{\|\boldsymbol{x}_{ij} + B\boldsymbol{n}\|} \\ E^{\mathrm{p1,L}} &:= \frac{1}{B} \sum_{k \in \mathbb{Z} \setminus \{0\}} \sum_{i,j=1}^{N} q_i q_j \mathrm{e}^{2\pi \mathrm{i} k x_{ij,1}/B} \Theta^{\mathrm{p1}}(k, \|\tilde{\boldsymbol{x}}_{ij}\|) \\ E^{\mathrm{p1},0} &:= -\frac{1}{2B} \sum_{\substack{i,j=1 \\ \tilde{\boldsymbol{x}}_{ij} \neq \mathbf{0}}}^{N} q_i q_j \Theta_0^{\mathrm{p1}}(\|\tilde{\boldsymbol{x}}_{ij}\|) \\ E^{\mathrm{p1,self}} &:= -\frac{\alpha}{\sqrt{\pi}} \sum_{j=1}^{N} q_j^2 \end{aligned}$$

for some $\alpha > 0$. Thereby the functions Θ^{p1} and Θ_0^{p1} are defined by

$$\Theta^{\mathrm{p1}}(k,r) := \int_0^\alpha \frac{1}{t} \exp\left(-\frac{\pi^2 k^2}{B^2 t^2} - r^2 t^2\right) \mathrm{d}t$$

and

$$\Theta_0^{\mathrm{p1}}(r) := \gamma + \Gamma(0, \alpha^2 r^2) + \ln(\alpha^2 r^2).$$

It can easily be seen that $\Theta^{\mathrm{p1}}(k,\cdot) \in \mathrm{C}^\infty(\mathbb{R})$ for any k.

Lemma 2. *For arbitrary $r \in \mathbb{R}$ we have $\Theta^{\mathrm{p1}}(k,r) \to 0$ with $\Theta^{\mathrm{p1}}(k,r) \sim k^{-2}\mathrm{e}^{-k^2}$ for $k \to \infty$.*

Proof: We immediately see

$$\Theta^{\mathrm{p1}}(k,r) \leq \Theta^{\mathrm{p1}}(k,0) = \frac{1}{2}\Gamma\left(0, \frac{\pi^2 k^2}{\alpha^2 B^2}\right).$$

The claim follows by applying the asymptotic expansion

$$\Gamma(0,x) \approx \frac{\mathrm{e}^{-x}}{x},$$

cf. [11, number 6.5.32], which holds for large x. □

Lemma 3. *For the univariate function*

$$\vartheta(x) := \begin{cases} 0 & : x = 0 \\ \gamma + \Gamma(0, x^2) + \ln(x^2) & : \textit{else} \end{cases}$$

we have $\vartheta \in \mathrm{C}^\infty(\mathbb{R})$.

Proof: From the identity, cf. [11, number 5.1.11],

$$\gamma + \Gamma(0,t) + \ln(t) = \sum_{k=1}^{\infty} \frac{(-1)^{k+1} t^k}{k!k}, \tag{6}$$

which is fulfilled for all positive t, it can be seen that $\lim_{t\to 0} \Gamma(0,t) + \ln t + \gamma = 0$. Thus, the function ϑ is continuous. Since (6) holds for $t > 0$ we obtain

$$\gamma + \Gamma(0,x^2) + \ln(x^2) = \sum_{k=1}^{\infty} \frac{(-1)^{k+1} x^{2k}}{k!k}$$

for all $x \neq 0$ and conclude

$$\begin{aligned} &\lim_{x\to+0} \frac{\mathrm{d}^n}{\mathrm{d}x^n} \left(\Gamma(0,x^2) + \ln(x^2)\right) \\ &= \lim_{x\to-0} \frac{\mathrm{d}^n}{\mathrm{d}x^n} \left(\Gamma(0,x^2) + \ln(x^2)\right) \neq \pm\infty \end{aligned}$$

for all $n \in \mathbb{N}$. □

B. An NFFT approach

Due to Lemma 2 the infinite sum in $E^{\mathrm{p1,L}}$ can be truncated, i.e., we can replace $\mathbb{Z}$ by $\mathcal{I}_{M_0}$ for some appropriate $M_0 \in 2\mathbb{N}$.

In the following we choose $h > 0$ and $\varepsilon > 0$ such that $\|\tilde{\boldsymbol{x}}_{ij}\| \leq h(1/2 - \varepsilon)$ for all $i,j = 1,\dots,N$. In order to compute the far field $E^{\mathrm{p1,L}} + E^{\mathrm{p1,0}}$ efficiently we define the regularization K_{R} by

$$K_{\mathrm{R}}(k,r) := \begin{cases} \frac{1}{B}\Theta^{\mathrm{p1}}(k,r) & : k \neq 0, |h^{-1}r| \leq 1/2 - \varepsilon \\ -\frac{1}{2B}\Theta_0^{\mathrm{p1}}(r) & : k = 0, |h^{-1}r| \leq 1/2 - \varepsilon \\ K_{\mathrm{B}}(k,r) & : |h^{-1}r| \in (1/2 - \varepsilon, 1/2] \end{cases},$$

where for each $k \in \mathbb{N}_0 \cap \mathcal{I}_{M_0}$ the function $K_{\mathrm{B}}(k,\cdot)$ is chosen such that the bivariate function $K_{\mathrm{R}}(k,\|\cdot\|) : h\mathbb{T}^2 \to \mathbb{R}$ is in the space $\mathrm{C}^p(h\mathbb{T}^2)$ for $p \in \mathbb{N}$ sufficiently large, i.e., $K_{\mathrm{B}}(k,\cdot)$ fulfills the conditions

$$\begin{aligned} K_{\mathrm{B}}^{(n)}(k, h/2 - h\varepsilon) &= K_{\mathrm{R}}^{(n)}(k, h/2 - h\varepsilon) \\ K_{\mathrm{B}}^{(n)}(k, -h/2 + h\varepsilon) &= (-1)^n K_{\mathrm{R}}^{(n)}(k, h/2 - h\varepsilon) \end{aligned}$$

for all $n = 0,\dots,p$ and is chosen such that

$$\begin{aligned} K_{\mathrm{R}}(k, h/2) &= K_{\mathrm{R}}(k, -h/2) \\ K_{\mathrm{R}}^{(n)}(k, h/2) &= K_{\mathrm{R}}^{(n)}(k, -h/2) = 0, \quad n = 1,\dots,p. \end{aligned}$$

We further set $K_{\mathrm{R}}(k,\|\boldsymbol{y}\|) := K_{\mathrm{R}}(k, h/2)$ for all $\boldsymbol{y} \in h\mathbb{T}^2$ with $\|\boldsymbol{y}\| > h/2$.

The smooth and periodic functions $K_{\mathrm{R}}(k,\|\cdot\|)$ can then be approximated by a bivariate trigonometric polynomial. To this end, we set $r := \|\boldsymbol{y}\|$, $\boldsymbol{y} \in h\mathbb{T}^2$, and obtain for each $k \in \mathbb{N} \cap \mathcal{I}_{M_0}$ with an appropriate $\boldsymbol{M} \in 2\mathbb{N}^2$

$$K_{\mathrm{R}}(k,\|\boldsymbol{y}\|) \approx \sum_{\boldsymbol{l} \in \mathcal{I}_{\boldsymbol{M}}} b_{k,\boldsymbol{l}} \mathrm{e}^{2\pi \mathrm{i} \boldsymbol{l}\cdot\boldsymbol{y}/h}$$

with the Fourier coefficients

$$b_{k,\boldsymbol{l}} := \frac{1}{|\mathcal{I}_{\boldsymbol{M}}|} \sum_{\boldsymbol{j} \in \mathcal{I}_{\boldsymbol{M}}} K_{\mathrm{R}}\left(k, \|\boldsymbol{j} \odot \boldsymbol{M}^{-1}\| h\right) \mathrm{e}^{-2\pi\mathrm{i}\boldsymbol{j}\cdot(\boldsymbol{l}\odot\boldsymbol{M}^{-1})}.$$

With $\boldsymbol{M}^* := (M_0, \boldsymbol{M}) \in 2\mathbb{N}^3$ we obtain, analogously to (5),

$$\begin{aligned} E^{\mathrm{p1,L}} + E^{\mathrm{p1,0}} &\approx \sum_{(k,\boldsymbol{l}) \in \mathcal{I}_{\boldsymbol{M}^*}} b_{|k|,\boldsymbol{l}} \sum_{i,j=1}^{N} q_i q_j \mathrm{e}^{2\pi\mathrm{i}\boldsymbol{v}_{k,\boldsymbol{l}}\cdot\boldsymbol{x}_{ij}} \\ &= \sum_{(k,\boldsymbol{l}) \in \mathcal{I}_{\boldsymbol{M}^*}} b_{|k|,\boldsymbol{l}} |S(k,\boldsymbol{l})|^2, \end{aligned}$$

where we set

$$S(k,\boldsymbol{l}) := \sum_{j=1}^{N} q_j \mathrm{e}^{2\pi\mathrm{i}\boldsymbol{v}_{k,\boldsymbol{l}}\cdot\boldsymbol{x}_j} \text{ with } \boldsymbol{v}_{k,\boldsymbol{l}} := \begin{pmatrix} k/B \\ \boldsymbol{l}/h \end{pmatrix}.$$

The sums $S(k,\boldsymbol{l})$, $(k,\boldsymbol{l}) \in \mathcal{I}_{\boldsymbol{M}^*}$, can efficiently be evaluated by a trivariate adjoint NFFT. For the 1d-periodic case a similar statement to that in Remark 1 can be given.

V. Conclusion

In this paper we proposed a new approach for the efficient calculation of the Coulomb interaction energy under 2d- and 1d- periodic boundary conditions. The presented methods are based on the corresponding Ewald summation formulas and nonequispaced fast Fourier transforms, where the ansatz is very much related to those of NFFT based fast summation methods. Numerical results of these algorithms will be reported in a further paper, where we aim to set the main focus on the derivation of error estimates as well as concluding statements about the optimal choice of the cutoff parameters and the regularization variables h, ε and p.

References

[1] P. P. Ewald, "Die Berechnung optischer und elektrostatischer Gitterpotentiale," *Annalen der Physik*, vol. 369, no. 3, pp. 253–287, 1921.

[2] A. Grzybowski, E. Gwóźdź, and A. Bródka, "Ewald summation of electrostatic interactions in molecular dynamics of a three-dimensional system with periodicity in two directions," *Phys. Rev. B*, vol. 61, pp. 6706–6712, Mar 2000.

[3] M. Porto, "Ewald summation of electrostatic interactions of systems with finite extent in two of three dimensions," *J. Phys. A*, vol. 33, pp. 6211 – 6218, 2000.

[4] S. W. de Leeuw, J. W. Perram, and E. R. Smith, "Simulation of electrostatic systems in periodic boundary conditions. I. Lattice sums and dielectric constants," *Proc. Roy. Soc. London Ser. A*, vol. 373, no. 1752, pp. 27 – 56, 1980.

[5] M. Deserno and C. Holm, "How to mesh up Ewald sums. I. A theoretical and numerical comparison of various particle mesh routines," *J. Chem. Phys.*, vol. 109, pp. 7678 – 7693, 1998.

[6] M. Pippig and D. Potts, "Particle simulation based on nonequispaced fast Fourier transforms," in *Fast Methods for Long-Range Interactions in Complex Systems*, ser. IAS-Series, G. Sutmann, P. Gibbon, and T. Lippert, Eds. Jülich: Forschungszentrum Jülich, 2011, pp. 131 – 158.

[7] J. Keiner, S. Kunis, and D. Potts, "Using NFFT3 - a software library for various nonequispaced fast Fourier transforms," *ACM Trans. Math. Software*, vol. 36, pp. Article 19, 1 – 30, 2009.

[8] D. Potts and G. Steidl, "Fast summation at nonequispaced knots by NFFTs," *SIAM J. Sci. Comput.*, vol. 24, pp. 2013 – 2037, 2003.

[9] D. Potts, G. Steidl, and A. Nieslony, "Fast convolution with radial kernels at nonequispaced knots," *Numer. Math.*, vol. 98, pp. 329 – 351, 2004.

[10] I. Kabadshow and H. Dachsel, "The Error-Controlled Fast Multipole Method for Open and Periodic Boundary Conditions," in *Fast Methods for Long-Range Interactions in Complex Systems*, ser. IAS-Series, G. Sutmann, P. Gibbon, and T. Lippert, Eds. Jülich: Forschungszentrum Jülich, 2011, pp. 85 – 113.

[11] M. Abramowitz and I. A. Stegun, Eds., *Handbook of Mathematical Functions*. Washington, DC, USA: National Bureau of Standards, 1972.

A sparse Prony FFT

Sabine Heider, Stefan Kunis*
University Osnabrück
Institute of Mathematics
49069 Osnabrück, Germany
* and Helmholtz Zentrum München
Institute of Computational Biology
85764 Neuherberg, Germany
Email: stefan.kunis@math.uos.de

Daniel Potts, Michael Veit
Chemnitz University of Technology
Faculty of Mathematics
09107 Chemnitz, Germany
Email: potts@mathematik.tu-chemnitz.de

***Abstract*—We describe the application of Prony-like reconstruction methods to the problem of the sparse Fast Fourier transform (sFFT) [6]. In particular, we adapt both important parts of the sFFT, quasi random sampling and filtering techniques, to Prony-like methods.**

Key words and phrases: **sparse Fast Fourier Transform, sFFT, Prony-like methods**
2000 AMS Mathematics Subject Classification **: 65T50**

I. Introduction

Computing the discrete Fourier transform of a vector of size N requires $\mathcal{O}(N \log N)$ arithmetical operations. The problem of a sparse Fourier transform (sFFT) now reads as follows: For a vector $\boldsymbol{x} = (x_l)_{l=0}^{N-1} \in \mathbb{C}^N$, assume that its Fourier representation

$$x_l = \frac{1}{N} \sum_{j=0}^{N-1} \hat{x}_j \, \mathrm{e}^{2\pi \mathrm{i} l j/N}, \quad l = 0, \dots, N-1,$$

has only $K \ll N$ non-vanishing Fourier coefficients $\hat{x}_{j_k} \in \mathbb{C}$, $j_k \in \{0, \dots, N-1\}$, $k = 1, \dots, K$. Now given part of the vector of samples $\boldsymbol{x} \in \mathbb{C}^N$, determine the non-vanishing Fourier coefficients $\hat{x}_{j_k} \in \mathbb{C}$ and their support $j_k \in \{0, \dots, N-1\}$, $k = 1, \dots, K$.

This problem has recently attracted much attention in the field of compressed sensing [2], [4], where one generally aims to reconstruct a vector with few non-vanishing coefficients from a relatively small number of linear measurements. Besides measurement matrices with independent random entries, structured matrices generated by a smaller number of random variables have been studied over the last years. Here, the most prominent example is given by a random selection of K rows of the N-th Fourier matrix, see e.g. [15], [11]. For this particular setting, the class of sublinear-time Fourier algorithms [5], [10] with a runtime that is polynomial in $\log N$ and K received much attention. The key idea, as outlined recently in [8], [6], [7] is the use of quasi random sampling and a band pass filter. Recently, these methods have been generalised for off-grid frequencies as well [1].

On the other hand, Prony-like methods are known for a long time in parameter estimation, in particular for exponential sums, see e.g. [13], [14] and references therein. In this note, we combine Prony-like methods with the above quasi random sampling and band pass filtering techniques.

II. Prony method

Let $K \geq 1$ be an integer, $f_k \in (-\infty, 0] + \mathrm{i}\,[-\pi, \pi)$, $k = 1, \dots, K$, be distinct complex numbers and $c_k \in \mathbb{C} \setminus \{0\}$, $k = 1, \dots, K$. We assume that $|c_k| > \varepsilon$ for a convenient bound $0 < \varepsilon \ll 1$ and consider the exponential sum of order K,

$$h(x) := \sum_{k=1}^{K} c_k \, \mathrm{e}^{f_k x}, \quad x \geq 0, \tag{II.1}$$

where the nodes $z_k := \mathrm{e}^{f_k}$, $k = 1, \dots, K$ are distinct values in the unit disk $\mathbb{D} := \{z \in \mathbb{C} : 0 < |z| \leq 1\}$ without zero. The well known Prony method recovers all parameters of the exponential sum (II.1), if sampled data

$$h(m) = \sum_{k=1}^{K} c_k \, \mathrm{e}^{f_k m} = \sum_{k=1}^{K} c_k \, z_k^m \in \mathbb{C}, \quad m = 0, \dots, M-1, \tag{II.2}$$

with $M \geq 2K$ are given. This problem is known as frequency analysis problem, which is important within many disciplines in sciences and engineering, see [13]. For a survey of the most successful methods for the data fitting problem with linear combinations of complex exponentials, we refer to [12]. We follow the lines in [14] and consider the case of an unknown order K for the exponential sum (II.1) and given noiseless sampled data $h(m)$, $m = 0, \dots, M-1$. Let $K_0 \in \mathbb{N}$ be a convenient upper bound of K, i.e. $K \leq K_0 \leq M/2$. With the M sampled data $h(m) \in \mathbb{C}$, $m = 0, \dots, M-1$, we form the rectangular Hankel matrix

$$\boldsymbol{H}_{M-K_0,K_0+1} \quad := \quad \big(h(m+k)\big)_{m,k=0}^{M-K_0-1,K_0} \tag{II.3}$$

and compute the singular value decomposition (SVD)

$$\boldsymbol{H}_{M-K_0,K_0+1} = \boldsymbol{U}_{M-K_0} \, \boldsymbol{D}_{M-K_0,K_0+1} \, \boldsymbol{W}_{K_0+1}, \tag{II.4}$$

where $\boldsymbol{U}_{M-K_0}$ and $\boldsymbol{W}_{K_0+1}$ are unitary matrices and where $\boldsymbol{D}_{M-K_0,K_0+1}$ is a rectangular diagonal matrix. The diagonal entries of $\boldsymbol{D}_{M-K_0,K_0+1}$ are the singular

values of $\boldsymbol{H}_{M-K_0,K_0+1}$ arranged in nonincreasing order $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_K > \sigma_{K+1} = \ldots = \sigma_{K_0+1} = 0$. Thus we can determine the rank K of the Hankel matrix (II.3) which coincides with the order of the exponential sum (II.1). Introducing the matrices

$$\begin{aligned} \boldsymbol{D}_{M-K_0,K} &:= \boldsymbol{D}_{M-K_0,K_0+1}(1:M-K_0,\,1:K) \\ &= \begin{pmatrix} \operatorname{diag}(\sigma_k)_{k=1}^{K} \\ \boldsymbol{O}_{M-K_0-K,K} \end{pmatrix}, \\ \boldsymbol{W}_{K,K_0+1} &:= \boldsymbol{W}_{K_0+1}(1:K,\,1:K_0+1), \end{aligned}$$

we can simplify the SVD of the Hankel matrix (II.3) as $\boldsymbol{H}_{M-K_0,K_0+1} = \boldsymbol{U}_{M-K_0}\,\boldsymbol{D}_{M-K_0,K}\,\boldsymbol{W}_{K,K_0+1}$. Setting

$$\boldsymbol{W}_{K,K_0}(s) = \boldsymbol{W}_{K,K_0+1}(1:K,\,1+s:K_0+s) \quad , s = 0,\,1, \tag{II.5}$$

we determine the nodes $z_k \in \mathbb{D}$, $k = 1,\ldots,K$, as eigenvalues of the matrix

$$\boldsymbol{F}_K^{\mathrm{SVD}} := \left(\boldsymbol{W}_{K,K_0}(0)^{\mathrm{T}}\right)^{\dagger} \boldsymbol{W}_{K,K_0}(1)^{\mathrm{T}}, \tag{II.6}$$

where † denotes the Moore-Penrose-Inverse. Thus the ESPRIT [16] algorithm reads as follows:

Algorithm II.1 (ESPRIT method)

Input: K_0, $M \in \mathbb{N}$ ($M \gg 2$, $3 \leq K_0 \leq M/2$, K_0 is upper bound of the order K of (II.1)), $h(m) \in \mathbb{C}$, $m = 0,\ldots,M-1$, $0 < \varepsilon \ll 1$.

1. Compute the SVD of the rectangular Hankel matrix (II.4). Determine the rank K of $\boldsymbol{H}_{M-K_0,K_0+1}$ such that $\sigma_{K+1} < \varepsilon\sigma_1$ and form the matrices (II.5).

2. Compute all eigenvalues $z_k \in \mathbb{D}$, $k = 1,\ldots,K$, of the square matrix $\boldsymbol{F}_K^{\mathrm{SVD}}$. Set $f_k := \log z_k$, $k = 1,\ldots,K$.

3. Compute the coefficients $c_k \in \mathbb{C}$, $k = 1,\ldots,K$, as least squares solution of the overdetermined linear Vandermonde–type system

$$(z_k^m)_{m=0,k=1}^{M-1,K}\,\boldsymbol{c} = \big(h(m)\big)_{m=0}^{M-1} \tag{II.7}$$

with $\boldsymbol{z} := (z_k)_{k=1}^{K}$ and $\boldsymbol{c} := (c_k)_{k=1}^{K}$

Output: $K \in \mathbb{N}$, $f_k \in (-\infty,\,0] + \mathrm{i}\,[-\pi,\pi)$, $c_k \in \mathbb{C}$, $k = 1,\ldots,K$.

Remark II.2 For noiseless sampled data, the authors in [14] describe the close connections between the classical Prony method, the matrix pencil method based on a QR decomposition, and the ESPRIT method. □

III. Random sampling and integer frequencies

We consider the sparse Fourier approximation problem. For a vector $\boldsymbol{x} \in \mathbb{C}^N$, we assume that its Fourier representation

$$x_l = \frac{1}{N}\sum_{j=0}^{N-1} \hat{x}_j\,\mathrm{e}^{2\pi\mathrm{i}lj/N}, \quad l = 0,\ldots,N-1,$$

has only $K \ll N$ non-vanishing Fourier coefficients $\hat{x}_{j_k}$, $j_k \in \{0,\ldots,N-1\}$, $k = 1,\ldots,K$. That is,

$$x_l = \frac{1}{N}\sum_{k=1}^{K} \hat{x}_{j_k}\,\mathrm{e}^{2\pi\mathrm{i}lj_k/N} = \sum_{k=1}^{K} c_k\,\mathrm{e}^{\tilde{f}_k l}, \quad l = 0,\ldots,N-1,$$

with $c_k = \frac{1}{N}\hat{x}_{j_k} \in \mathbb{C}\backslash\{0\}$ and $\tilde{f}_k = 2\pi\mathrm{i}j_k/N \in \mathbb{C}$ satisfying $\operatorname{Re}\tilde{f}_k = 0$ and $\operatorname{Im}\tilde{f}_k \in [0,2\pi)$, $k = 1,\ldots,K$. Applying e.g. the ESPRIT method to the first M entries $x_0,\ldots,x_{M-1}$ of $\boldsymbol{x}$ would yield coefficients $c_k \in \mathbb{C}\backslash\{0\}$ and frequencies $f_k \in \mathbb{C}$ with $\operatorname{Re} f_k = 0$ and $\operatorname{Im} f_k \in [-\pi,\pi)$, $k = 1,\ldots,K$. By

$$\tilde{f}_k = \begin{cases} f_k, & \operatorname{Im} f_k \geq 0, \\ f_k + 2\pi\mathrm{i}, & \operatorname{Im} f_k < 0, \end{cases}$$

and

$$\begin{aligned} j_k &= \operatorname{round}\Big(\frac{N}{2\pi}\operatorname{Im}\tilde{f}_k\Big), \\ \hat{x}_{j_k} &= N c_k, \end{aligned}$$

$k = 1,\ldots,K$, we could accomplish the computation of the K-sparse Fourier transform $\hat{\boldsymbol{x}} \in \mathbb{C}^N$ that way.

However, we do not intend to take necessarily the first M entries $x_0,\ldots,x_{M-1}$ as input for the Prony-like method but (to a certain extent) random M entries of the vector $\boldsymbol{x}$. We use a random parameter $\sigma \in \{1,\ldots,N-1\}$ being invertible modulo N and a random shift parameter $\tau \in \{0,\ldots,N-1\}$ similarly as used in [8]. The following theorem confirms the possibility to connect randomized signal samples and a Prony-like algorithm for computation as suggested in [17].

Theorem III.1 *Let the vector $\boldsymbol{x} = (x_l)_{l=0}^{N-1} \in \mathbb{C}^N$ with a K-sparse Fourier representation*

$$x_l = \frac{1}{N}\sum_{k=1}^{K} \hat{x}_{j_k}\,\mathrm{e}^{2\pi\mathrm{i}lj_k/N}, \quad l = 0,\ldots,N-1, \tag{III.1}$$

and two integers $\sigma,\,\tau \in \{0,\ldots,N-1\}$, σ being invertible modulo N, be given. Then we have

$$x_{\sigma l+\tau} = \sum_{k=1}^{K} c_k\,\mathrm{e}^{\tilde{f}_k l}, \quad l = 0,\ldots,N-1,$$

with coefficients

$$c_k = \frac{1}{N}\hat{x}_{j_k}\omega_N^{j_k\tau} \in \mathbb{C}\backslash\{0\}$$

and frequencies

$$\tilde{f}_k = \mathrm{i}\frac{2\pi}{N}((j_k\sigma) \bmod N) \in \mathbb{C}$$

such that $\operatorname{Re}\tilde{f}_k = 0$ and $\operatorname{Im}\tilde{f}_k \in [0,2\pi)$, $k = 1,\ldots,K$. Here, $\omega_N = \mathrm{e}^{2\pi\mathrm{i}/N}$ denotes the principal N-th primitive root of unity.

A simple consequence is the following: Let two integers $\sigma,\,\tau \in \{0,\ldots,N-1\}$, σ invertible modulo N, and a sufficiently big number of samples $M \geq 2K$ be given.

One can determine the K non-zero Fourier coefficients $\hat{x}_{j_k} \in \mathbb{C}$ and integer frequencies $j_k \in \{0, \dots, N-1\}$ of the vector $\boldsymbol{x} \in \mathbb{C}^N$ with entries (III.1) using the samples $x_{\sigma l+\tau},\ l = 0, \dots, M-1$, by

$$j_k = (\mathrm{round}(\frac{N}{2\pi} \mathrm{Im}\, f_k)\sigma^{-1}) \bmod N, \quad k = 1, \dots, K,$$

and

$$\hat{x}_{j_k} = N c_k \omega_N^{-j_k \tau}, \quad k = 1, \dots, K,$$

where σ^{-1} denotes the inverse of σ modulo N and $c_k,\ f_k,\ k = 1, \dots, K$, the output of a Prony-like reconstruction method. Hence, the Prony-like methods are well-suited for the computation of sparse Fourier transforms. After applying the Algorithm II.1 to the permuted signal samples, we obtain the integer frequencies by rounding. Further, we need to invert the random separation and take the modulo of the result in order to guarantee that $j_k \in \{0, \dots, N-1\}$. The random shift in the sampling index causes a modulation of the Fourier coefficients which can be easily corrected. Assigning such a quasi-random sign is intended to prevent cancellations of nearby coefficients which would look alike in time domain samples. A more detailed analysis follows for the expected separation of nearby frequencies [9] which leads to a stabilization of the Prony method, see [14].

Theorem III.2 *Let $N \in \mathbb{N}$ be prime, the vector $\hat{\boldsymbol{x}} \in \mathbb{C}^N$ contain K nonzeros and choose $\sigma \in \{1, \dots, N-1\}$ uniformly distributed at random. Then the separation distance of the vector $(\hat{x}_{\sigma j})_{j=0,\dots,N-1} \in \mathbb{C}^N$ fulfils*

$$\mathbb{P}\left(\min_{k \neq l} |\sigma j_k - \sigma j_l| \geq \frac{N-1}{2K(K-1)}\right) \geq \frac{1}{2}. \qquad \text{(III.2)}$$

Proof: The frequencies $\{j_k\}$ have $\binom{K}{2}$ pairwise differences. For each fixed difference $s \in \{1, \dots, N-1\}$, there exist at most $2\binom{K}{2}$ different values $\sigma \in \{1, \dots, N-1\}$ such that $\min_{k \neq l} |\sigma j_k - \sigma j_l| = s$. We estimate

$$\mathbb{P}\left(\min_{k \neq l} |\sigma j_k - \sigma j_l| = s\right) \leq \frac{K(K-1)}{N-1},$$

from which the assertion easily follows. ∎

This result becomes effective for $K \in o(\sqrt{N})$ and gives high probability of success for the Prony method by independent repetition.

IV. A splitting approach

The most time-consuming steps of Prony-like recovery methods are the factorization of the Hankel matrix (II.4) and the least squares solution of the system (II.7). Hence, the computational costs for the Prony-like methods are $\mathcal{O}(K^2 M)$ in general and $\mathcal{O}(K^3 \log(\frac{N}{K}))$ if we choose $M = \mathcal{O}(K \log(\frac{N}{K}))$ samples. This recovery method is sublinear in the problem size N but scales cubic in the number K of non-zeros, such that only very sparse Fourier transforms can be computed in an efficient way. We proceed by a modification of the Prony-like methods adapted to the sparse Fourier transform problem to further reduce computational costs.

Let a number $B \in \mathbb{N}$, $B \leq K$, of frequency bands be chosen. Instead of recovering all of the K non-vanishing Fourier coefficients at once, we split the frequency set $\{0, 1, \dots, N-1\}$ into the disjoint subsets $\{\frac{b-1}{B}N, \frac{b-1}{B}N + 1, \dots, \frac{b}{B}N - 1\}$, $b = 1, \dots, B$, and assume that $\frac{N}{B}$ is prime or that $\{1, \dots, \frac{N}{B}\}$ contains many invertible elements σ. We now determine only coefficients with frequencies in such a subset in each recovery step.

In order to do so, we use a filter that is concentrated both in time and frequency. Let $\varepsilon > \varepsilon' > 0$ be two parameters and set $N_1 = \lceil \varepsilon' N/2 \rceil$ and $N_2 = \lfloor \varepsilon N/2 \rfloor$. We define the auxiliary function $a : [N_1, N_2] \to \mathbb{R}$ using $a_1 : \mathbb{R} \to \mathbb{R}$ and $a_2 : [-1, 1] \to \mathbb{R}$ via

$$\begin{aligned} a_1(x) &= \mathrm{e}^{-1/x^2},\ x \in \mathbb{R}\backslash\{0\},\ a_1(0) = 0, \\ a_2(x) &= \frac{a_1(1-x)}{a_1(1-x) + a_1(1+x)}, \\ a(x) &= a_2(2/(N_2 - N_1)(x - N_1) - 1). \end{aligned} \qquad \text{(IV.1)}$$

Then, the function a is smooth in $[N_1, N_2]$, $a(N_1) = 1$, $a(N_2) = 0$, and all derivatives of a vanish at N_1 and N_2. We now set $\hat{\boldsymbol{g}} = (\hat{g}_j)_{j=0}^{N-1} \in \mathbb{R}^N$,

$$\hat{g}_j = \begin{cases} 1, & j < N_1 \text{ and } j > N - N_1, \\ 0, & N_2 < j < N - N_2, \\ a(j), & j \in [N_1, N_2], \\ a(N-j), & j \in [N - N_2, N - N_1]. \end{cases}$$

and define the final filter for a spatial cut-off $W \in \mathbb{N}$, $W < N$, typically chosen as $W = \mathcal{O}(B \log N)$, by $\boldsymbol{h} = (h_l)_{l=0}^{N-1} \in \mathbb{C}^N$,

$$h_l = \begin{cases} g_l, & l \in [-\frac{W}{2}, \frac{W}{2}], \\ 0, & \text{elsewise.} \end{cases}$$

Instead of the signal $\boldsymbol{x}$, we use the convolved vector

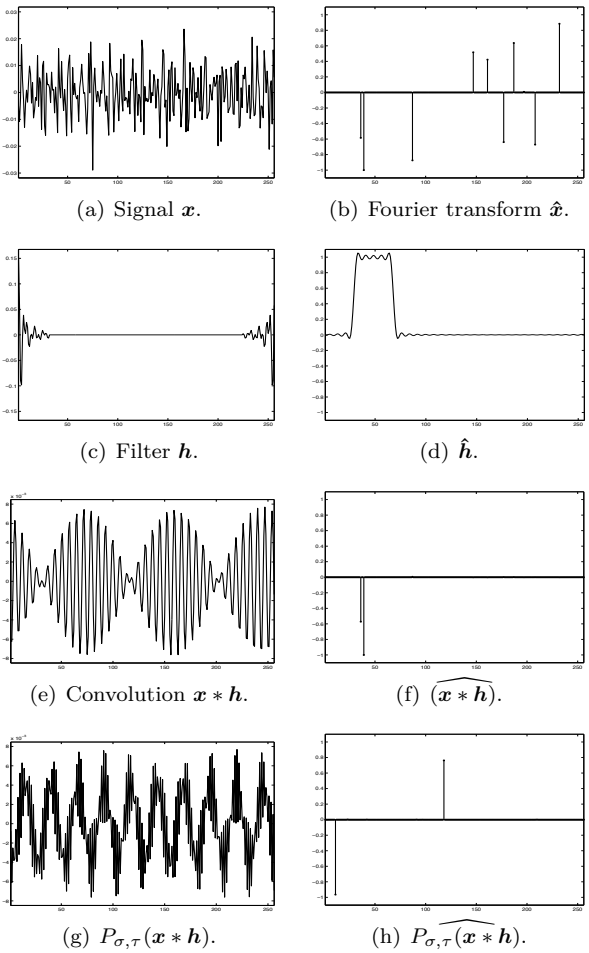
$$\boldsymbol{x} * \boldsymbol{h} = (\sum_{k=0}^{N-1} x_k h_{l-k})_{l=0}^{N-1} \in \mathbb{C}^N$$

in the computation of the sparse Fourier transform. We have

$$\widehat{(\boldsymbol{x} * \boldsymbol{h})} = \hat{\boldsymbol{x}} \cdot \hat{\boldsymbol{h}} = (\hat{x}_j \hat{h} x_j)_{j=0}^{N-1}.$$

Therefore, it is likely that the number of non-zero Fourier coefficients of $\boldsymbol{x} * \boldsymbol{h}$ is smaller than before since most of the coefficients $\hat{h}_j,\ j \notin \{\frac{b-1}{B}N, \dots, \frac{b}{B}N - 1\}$, are (almost) zero. In case the randomized Prony-like method outputs a frequency which is not in the currently considered subset $\{\frac{b-1}{B}N, \dots, \frac{b}{B}N - 1\}$, we discard the corresponding coefficient, such that the expected number of non-zero coefficients which we seek to identify in each of the B steps is $\frac{K}{B}$. We employ the randomization in each step and use $M_1 = \mathcal{O}(\frac{K}{B} \log(\frac{NB}{K}))$ samples

$$(P_{\sigma,\tau}(\boldsymbol{x} * \boldsymbol{h}))_l := (\boldsymbol{x} * \boldsymbol{h})_{\sigma l + \tau},$$



(a) Signal $\boldsymbol{x}$. (b) Fourier transform $\hat{\boldsymbol{x}}$.

(c) Filter $\boldsymbol{h}$. (d) $\hat{\boldsymbol{h}}$.

(e) Convolution $\boldsymbol{x} * \boldsymbol{h}$. (f) $\widehat{(\boldsymbol{x} * \boldsymbol{h})}$.

(g) $P_{\sigma,\tau}(\boldsymbol{x} * \boldsymbol{h})$. (h) $\widehat{P_{\sigma,\tau}(\boldsymbol{x} * \boldsymbol{h})}$.

Figure IV.1. **Example step of the split Prony-like method.** In step $b = 2$ of $B = 8$ computation steps, the signal $\boldsymbol{x}$ of length $N = 2^8 = 256$ (a) with $K = 10$ non-zeros in its Fourier transform $\hat{\boldsymbol{x}}$ (b) is convolved with the filter $\boldsymbol{h}$ (c). The resulting vector $\boldsymbol{x} * \boldsymbol{h}$ (e) is randomly permuted. In all the subfigures only the real parts of the complex vectors are plotted.

$l = 0, \dots, M_1 - 1$, in each recovery step. Figure IV.1, see also [17], serves as an illustration for this splitting approach. The effects of convolving the signal $\boldsymbol{x}$ with the filter $\boldsymbol{h}$ and applying the randomization afterwards are pictured both in the time domain and in the frequency domain for a particular example.

Finally, we shortly analyse the expected computational complexity of this splitted Prony method. As argued above, the expected number of frequencies is $\mathcal{O}(\frac{K}{B})$ and we thus choose $M_1 = \mathcal{O}(\frac{K}{B} \log(\frac{NB}{K}))$ samples per recovery step. The computationally most expensive parts of one step then is spatial filtering which takes $\mathcal{O}(M_1 W)$ arithmetic operations and the Prony-like method requiring $\mathcal{O}(M_1 \frac{K^2}{B^2})$ arithmetic operations. Moreover, we assume a spatial filter length $W = \mathcal{O}(B \log N)$, which is supported for a particular error measure in [8], [6], [7], and choose the optimal value $B = \mathcal{O}(K^{\frac{2}{3}})$ of recovery steps. In total, this leads to a complexity of $\mathcal{O}(K^{\frac{5}{3}} \log^2 N)$. While this is beyond the recently achieved bounds for sparse FFTs we nevertheless expect a wider applicability due to the fact that Prony-like methods seem to be more stable with respect to off-grid frequencies, cf. [3].

Acknowledgments

The authors gratefully acknowledge support by the German Research Foundation within the project KU 2557/1-2 and PO 711/10-2 and by the Helmholtz Association within the young investigator group VH-NG-526.

References

[1] P. Boufounos, V. Cevher, A.C. Gilbert, Y. Li, and M. Strauss: *What's the frequency, Kenneth?: Sublinear Fourier sampling off the grid.* In A. Gupta, K. Jansen, J. Rolim, and R. Servedio (eds.): *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, vol. 7408 of *Lecture Notes in Computer Science*, pp. 61–72. Springer, 2012.

[2] E. Candès, J. Romberg, and T. Tao: *Stable signal recovery from incomplete and inaccurate measurements.* Comm. Pure Appl. Math., 59:1207 – 1223, 2006.

[3] Y. Chi, L.L. Scharf, A. Pezeshki, and A.R. Calderbank: *Sensitivity to basis mismatch in compressed sensing.* IEEE Trans. Signal Process., 59:2182 – 2195, 2011.

[4] D.L. Donoho: *Compressed sensing.* IEEE Trans. Inform. Theory, 52:1289 – 1306, 2006.

[5] A. Gilbert, S. Muthukrishnan, and M. Strauss: *Improved time bounds for near-optimal sparse Fourier representations.* In *Proc.* SPIE, 2005.

[6] H. Hassanieh, P. Indyk, D. Katabi, and E. Price: *Nearly optimal sparse Fourier transform.* In *STOC*, 2012.

[7] H. Hassanieh, P. Indyk, D. Katabi, and E. Price: *sFFT: Sparse fast Fourier transform. http://groups.csail.mit.edu/netmit/sFFT/*, 2012.

[8] H. Hassanieh, P. Indyk, D. Katabi, and E. Price: *Simple and practical algorithm for sparse Fourier transform.* In *SODA*, pp. 1183–1194, 2012.

[9] S. Heider: *Separation benachbarter Frequenzen bei dünnbesetzten trigonometrischen Polynomen.* Studienarbeit, Universität Osnabrück, 2012.

[10] M.A. Iwen: *Combinatorial sublinear-time Fourier algorithms.* Found. Comput. Math., 10:303 – 338, 2010.

[11] S. Kunis and H. Rauhut: *Random sampling of sparse trigonometric polynomials II, Orthogonal matching pursuit versus basis pursuit.* Found. Comput. Math., 8:737 – 763, 2008.

[12] V. Pereyra and G. Scherer: *Exponential data fitting.* In V. Pereyra and G. Scherer (eds.): *Exponential Data Fitting and its Applications*, pp. 1 – 26, Sharjah, 2010. Bentham Sci. Publ., IEEE Comput. Soc.

[13] V. Pereyra and G. Scherer: *Exponential Data Fitting and its Applications.* Bentham Sci. Publ., Sharjah, 2010.

[14] D. Potts and M. Tasche: *Parameter estimation for nonincreasing exponential sums by Prony-like methods.* Linear Algebra Appl., p. accepted, 2012.

[15] H. Rauhut: *Random sampling of sparse trigonometric polynomials.* Appl. Comput. Harmon. Anal., 22:16 – 42, 2007.

[16] R. Roy and T. Kailath: *ESPRIT—estimation of signal parameters via rotational invariance techniques.* IEEE Trans. Acoustic speech and Signal Process., 37:984 – 994, 1989.

[17] M. Veit: *Parameter estimation based on Prony-like methods and sparse fast Fourier transforms.* Diplomarbeit, Technische Universität Chemnitz, 2012.

Taylor and rank-1 lattice based nonequispaced fast Fourier transform

Toni Volkmer
Chemnitz University of Technology
Department of Mathematics
09107 Chemnitz, Germany
Email: toni.volkmer@mathematik.tu-chemnitz.de

***Abstract*—The nonequispaced fast Fourier transform (NFFT) allows the fast approximate evaluation of trigonometric polynomials with frequencies supported on full box-shaped grids at arbitrary sampling nodes. Due to the curse of dimensionality, the total number of frequencies and thus, the total arithmetic complexity can already be very large for small refinements at medium dimensions. In this paper, we present an approach for the fast approximate evaluation of trigonometric polynomials with frequencies supported on symmetric hyperbolic cross index sets at arbitrary sampling nodes. This approach is based on Taylor expansion and rank-1 lattice methods. We prove error estimates for the approximation and present numerical results.**

I. Introduction

We consider the evaluation of trigonometric polynomials $f\colon \mathbb{T}^d := [0,1)^d \to \mathbb{C}$,

$$f(\boldsymbol{x}) = \sum_{\boldsymbol{l}\in\mathcal{I}_N} \hat{f}_{\boldsymbol{l}}\, \mathrm{e}^{-2\pi \mathrm{i}\boldsymbol{l}\boldsymbol{x}},\ \hat{f}_{\boldsymbol{l}} \in \mathbb{C},\ \mathcal{I}_N \subset \mathbb{Z}^d \cap [-N,N]^d, \quad (1)$$

at arbitrary sampling nodes $\boldsymbol{y}_\ell \in \mathbb{T}^d$, $\ell = 0,\ldots,L-1$. For given Fourier coefficients $\hat{f}_{\boldsymbol{l}}$, the direct evaluation of the trigonometric sums $f(\boldsymbol{y}_\ell)$, $\ell = 0,\ldots,L-1$, takes $\mathcal{O}(L\,|\mathcal{I}_N|)$ arithmetic operations. Various fast methods for the approximate evaluation of the trigonometric sums $f(\boldsymbol{y}_\ell)$ were developed.

In the case, when the frequency index set $\mathcal{I}_N$ is a full grid, $\mathcal{I}_N = G_N^d := \mathbb{Z}^d \cap [-N,N)^d$, the nonequispaced fast Fourier transform (NFFT, see [1] and references therein) allows the fast approximate evaluation of the trigonometric polynomial f at arbitrary sampling nodes $\boldsymbol{y}_\ell$, $\ell = 0,\ldots,L-1$, in $\mathcal{O}(|\log\epsilon|^d L + |G_N^d|\log|G_N^d|)$ arithmetic operations, where ϵ is the approximation error. Furthermore, there exist Taylor based versions (cf. [2], [3]) with an arithmetic complexity of $\mathcal{O}(|\log\epsilon|^d(L+|G_N^d|\log|G_N^d|))$, which use fast Fourier transforms (FFT) for evaluating the trigonometric polynomial f as well as its derivatives at equispaced nodes and approximate the trigonometric sum $f(\boldsymbol{y}_\ell)$ by a Taylor expansion at the closest equispaced node. However, since the cardinality of the full grid G_N^d is $|G_N^d| = (2N)^d$, the total number of arithmetic operations can already be very large for small refinements N at medium dimensionality (e.g. $d = 3,4,5$).

For dyadic hyperbolic crosses $\tilde{H}_n^d := \cup_{\boldsymbol{j}\in\mathbb{N}_0^d,\|\boldsymbol{j}\|_1=n} \tilde{G}_{\boldsymbol{j}}$, $\tilde{G}_{\boldsymbol{j}} := \mathbb{Z}^d \cap \times_{t=1}^d (-2^{j_t-1}, 2^{j_t-1}]$, $\|\boldsymbol{j}\|_1 = |j_1| + \ldots + |j_d|$, the nonequispaced hyperbolic cross fast Fourier transform (NHCFFT) [4] allows the fast approximate evaluation of trigonometric polynomials with frequencies supported on the index set $\mathcal{I}_N = \tilde{H}_n^d$ at arbitrary sampling nodes $\boldsymbol{y}_\ell$, $\ell = 0,\ldots,L-1$. The NHCFFT is based on the hyperbolic cross FFT (cf. [5], [6]) and has an arithmetic complexity of $\mathcal{O}(|\log\epsilon|^d\ L\ \log|\tilde{H}_n^d| + |\log\epsilon|\ |\tilde{H}_n^d| + |\tilde{H}_n^d|\log|\tilde{H}_n^d|)$, where $|\tilde{H}_n^d| \le C\, n^{d-1}\, 2^n$ with a constant $C > 0$ depending only on d. In [7], the stability of the hyperbolic cross discrete Fourier transform was studied.

For symmetric hyperbolic cross index sets $\mathcal{I}_N = H_N^d := \{\boldsymbol{j}\in\mathbb{Z}^d\colon r(\boldsymbol{j}) \le N\}$ in frequency domain with refinement $N \in \mathbb{N}$, $r(\boldsymbol{j}) := \prod_{t=1}^d \max(1,|j_t|)$, we present an approach for the fast approximate evaluation at arbitrary sampling nodes $\boldsymbol{y}_\ell$. This method uses one-dimensional FFTs for evaluating the trigonometric polynomial f and its derivatives at nodes of a rank-1 lattice. Then, for each sampling node $\boldsymbol{y}_\ell$, a Taylor expansion of degree $m-1$, $m \in \mathbb{N}$, at a closest rank-1 lattice node is performed. This results in a total arithmetic complexity of $\mathcal{O}(m^d\,(L + M\log M + |H_N^d|))$, where $M \in \mathbb{N}$ is the size of the rank-1 lattice. We show error estimates for the approximation error of the presented method. Note, that we have the inclusion $\tilde{H}_n^d \subset H_{2^{n-1}}^d \subset \tilde{H}_{n-1+2d}^d$, see [8, Lemma 2.1].

In Section II, we give a short overview over Taylor expansion of trigonometric polynomials and define rank-1 lattices. We show that trigonometric polynomials can be evaluated at rank-1 lattice nodes using a one-dimensional FFT. The proposed method is presented in Section III as well as error estimates for symmetric hyperbolic cross index sets H_N^d. Results of numerical tests are presented in Section IV. Finally, we summarize the results in Section V.

II. Prerequisite

A. Taylor expansion

We approximate a function $f : \mathbb{T}^d \to \mathbb{C}$ by

$$f(\boldsymbol{x}) \approx s_m(\boldsymbol{x}) := \sum\nolimits_{0\le|\boldsymbol{s}|<m} \frac{D^{\boldsymbol{s}} f(\boldsymbol{a})}{\boldsymbol{s}!}(\boldsymbol{x}-\boldsymbol{a})^{\boldsymbol{s}},$$

where $m \in \mathbb{N}$, $D^{\boldsymbol{s}} f := \frac{\partial^{s_1}}{\partial {x_1}^{s_1}} \ldots \frac{\partial^{s_d}}{\partial {x_d}^{s_d}}$, $\boldsymbol{x} := (x_1,\ldots,x_d)^\top$, $\boldsymbol{s} := (s_1,\ldots,s_d) \in \mathbb{N}_0^d$, $|\boldsymbol{s}| := |s_1| + \ldots + |s_d|$, $D^{\boldsymbol{0}} f := f$, $\boldsymbol{s}! := s_1! \cdot \ldots \cdot s_d!$, $\boldsymbol{x}^{\boldsymbol{s}} := {x_1}^{s_1} \cdot \ldots \cdot {x_d}^{s_d}$.

For a trigonometric polynomial f from (1), we have $D^{\boldsymbol{s}} f(\boldsymbol{x}) = \sum_{\boldsymbol{l} \in \mathcal{I}_N} (-2\pi \mathrm{i} \boldsymbol{l})^{\boldsymbol{s}} \hat{f}_{\boldsymbol{l}} \, \mathrm{e}^{-2\pi \mathrm{i} \boldsymbol{l} \boldsymbol{x}}$ and thus,

$$s_m(\boldsymbol{x}) = \sum_{0 \leq |\boldsymbol{s}| < m} \frac{(\boldsymbol{x} - \boldsymbol{a})^{\boldsymbol{s}}}{\boldsymbol{s}!} \sum_{\boldsymbol{l} \in \mathcal{I}_N} (-2\pi \mathrm{i} \boldsymbol{l})^{\boldsymbol{s}} \, \hat{f}_{\boldsymbol{l}} \, \mathrm{e}^{-2\pi \mathrm{i} \boldsymbol{l} \boldsymbol{a}}. \quad (2)$$

B. Rank-1 lattice

Definition II.1 (rank-1 lattice)**.** *Let $M \in \mathbb{N}$, $\boldsymbol{z} \in \mathbb{Z}^d$. We define the rank-1 lattice $\Lambda(\boldsymbol{z}, M) \subset \mathbb{T}^d$ of size M with generating vector $\boldsymbol{z} \in \mathbb{Z}^d$ by $\Lambda(\boldsymbol{z}, M) := \{\boldsymbol{x}_k := ((k\boldsymbol{z}) \mod M)/M\}_{k=0}^{M-1}$.* □

Definition II.2 (mesh norm)**.** *Let the metric $\mu(\boldsymbol{x}, \boldsymbol{y}) := \min_{\boldsymbol{k} \in \mathbb{Z}^d} \|\boldsymbol{x} - \boldsymbol{y} + \boldsymbol{k}\|_\infty$ be given for $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{T}^d$. We define the mesh norm δ of an arbitrary point set $\mathcal{X} := \{\boldsymbol{x}_k\}_{k=0}^{M-1} \subset \mathbb{T}^d$ by $\delta := 2 \max_{\boldsymbol{x} \in \mathbb{T}^d} \min_{\boldsymbol{x}_k \in \mathcal{X}} \mu(\boldsymbol{x}_k, \boldsymbol{x})$.* □

For an arbitrary point set $\mathcal{X} \subset \mathbb{T}^d$ of size $|\mathcal{X}| = M$, we have $\delta \geq 1/\sqrt[d]{M}$, see e.g. [9, Lemma 3.1]. The following Lemma shows the existence of a rank-1 lattice $\Lambda(\boldsymbol{z}, M)$ of size M, such that the mesh norm $\delta \leq C_d/\sqrt[d]{M}$, where $C_d > 1$ is a constant depending only on d, i.e., we have $\delta \sim 1/\sqrt[d]{M}$.

Lemma II.3. *Let $b \in \mathbb{N}$, $b \geq 3$. Then, there exists a rank-1 lattice $\Lambda(\boldsymbol{z}, M)$ of size $M = b(b+1)$ for $d = 2$ and $b^d \cdot 2^{\frac{d(d-1)}{2}-1} < M \leq b^d \cdot 2^{d(d-2)}$ for $d \geq 3$ with generating vector $\boldsymbol{z} \in \mathbb{Z}^d$, such that the mesh norm $\delta \leq C_d/\sqrt[d]{M}$, where $C_d > 1$ is a constant depending only on d.*

Proof: In the case $d = 2$, we choose the rank-1 lattice size $M := b \cdot (b+1)$ and the generating vector $\boldsymbol{z} := (b, b+1)^\top$. Since b and $b+1$ are relatively prime to each other, there exists a bijective mapping between the rank-1 lattice nodes $\boldsymbol{x}_k := (k\boldsymbol{z} \mod M)/M$, $k = 0, \ldots, M-1$, and the grid $(j_1/(b+1), j_2/b)^\top$, $j_1 = 0, \ldots, b$ and $j_2 = 0, \ldots, b-1$, cf. [10]. Obviously, the mesh norm $\delta = 1/b \leq \frac{2}{\sqrt{3}}/\sqrt{M}$.

In the case $d = 3$, we set $v_1 := 2b+1$ and $v_2 := 2b$. Due to Bertrand's postulate there exists a prime number $p_3 \in \mathbb{N}$, $b \leq p_3 < 2b$. We choose $v_3 \in \{p_3, \ldots, v_2 - 1\}$, such that v_3 is relatively prime to v_1 and v_2. We set the rank-1 lattice size $M := v_1 \cdot v_2 \cdot v_3$ and the generating vector $\boldsymbol{z} := (M/v_1, M/v_2, M/v_3)^\top$. Then, the mesh norm $\delta \leq 1/v_3 \leq 1/b \leq 2/\sqrt[3]{M}$ and the rank-1 lattice size $M = (2b+1) \cdot 2b \cdot v_3 \geq (2b+1) \cdot 2b \cdot b > b^3 \cdot 2^2$.

In the case $d \geq 4$, we set $v_1 := b \cdot 2^{d-2} + 1$ and $v_2 := b \cdot 2^{d-2}$. We apply Bertrand's postulate $d-2$ times and choose $v_3, \ldots, v_d$, such that $v_1, \ldots, v_d$ are relatively prime to each other and $v_3 > \ldots > v_d \geq b$. We choose the rank-1 lattice size $M := \prod_{t=1}^d v_t$ and the generating vector $\boldsymbol{z} := (M/v_1, \ldots, M/v_d)^\top$. This yields that the mesh norm $\delta \leq 1/v_d \leq 1/b \leq 2^{d-2}/\sqrt[d]{M}$ and the rank-1 lattice size $M \geq (2^{d-2}b + 1) \cdot 2^{d-2} b \cdot \prod_{t=3}^d (2^{d-t} b) > b^d \cdot 2^{\frac{d(d-1)}{2}-1}$. ■

The following Lemma shows that rank-1 lattices exist where the constant C_d is arbitrarily close to 1 for constant d and increasing rank-1 lattice size M.

Lemma II.4. *For each constant $C_d > 1$, there exists a parameter $M^* \in \mathbb{N}$, such that for all $M' \geq M^*$ we can construct a rank-1 lattice $\Lambda(\boldsymbol{z}, M)$ of size $M \in \left(M', (C_d)^d M'\right]$ with mesh norm $\delta < C_d/\sqrt[d]{M}$.*

Proof: Let $R_{c,d}$ be the dth c-Ramanujan prime [11], i.e., the smallest integer such that there are at least d primes in the interval $(cx, x]$ for all $x \geq R_{c,d}$, where $c \in (0,1)$. For arbitrary constant $C_d > 1$, we set $c := (C_d)^{-1}$, $M^* := \left((C_d)^{-1} R_{(C_d)^{-1},d}\right)^d$ and $x := C_d \sqrt[d]{M'}$, $M' \geq 1$. Then, there are at least d primes $v_1, \ldots, v_d$ in the interval $(\sqrt[d]{M'}, C_d \sqrt[d]{M'}]$ for all $M' \geq M^*$. We choose the rank-1 lattice size $M := \prod_{t=1}^d v_t$ and the generating vector $\boldsymbol{z} := (M/v_1, \ldots, M/v_d)^\top$. Consequently, we have $M' < M \leq (C_d)^d M'$ and $\delta < 1/\sqrt[d]{M'} \leq C_d/\sqrt[d]{M}$. ■

C. Evaluation at rank-1 lattice nodes (rank-1 lattice FFT)

We consider the evaluation of a trigonometric polynomial $g : \mathbb{T}^d \to \mathbb{C}$ supported on the frequency index set $\mathcal{I}_N \subset \mathbb{Z}^d \cap [-N, N]^d$, $g(\boldsymbol{x}) := \sum_{\boldsymbol{l} \in \mathcal{I}_N} \hat{g}_{\boldsymbol{l}} \, \mathrm{e}^{-2\pi \mathrm{i} \boldsymbol{l} \boldsymbol{x}}$, $\hat{g}_{\boldsymbol{l}} \in \mathbb{C}$, at rank-1 lattice nodes $\boldsymbol{x}_k \in \Lambda(\boldsymbol{z}, M)$. As presented in [8], we have

$$g(\boldsymbol{x}_k) = g(k\boldsymbol{z}/M) = \sum_{j=0}^{M-1} \left(\sum_{\substack{\boldsymbol{l} \in \mathcal{I}_N \\ \boldsymbol{l}\boldsymbol{z} \equiv j \pmod M}} \hat{g}_{\boldsymbol{l}} \right) \mathrm{e}^{-2\pi \mathrm{i} \frac{kj}{M}}$$

and the outer sum is a one-dimensional discrete Fourier transform of length M. Using a one-dimensional FFT, the trigonometric polynomial g can be evaluated at all rank-1 lattice nodes in $\mathcal{O}(M \log M + |\mathcal{I}_N|)$ arithmetic operations.

Setting the Fourier coefficients $\hat{g}_{\boldsymbol{l}} := (-2\pi \mathrm{i} \boldsymbol{l})^{\boldsymbol{s}} \hat{f}_{\boldsymbol{l}}$, where $\hat{f}_{\boldsymbol{l}}$ are the Fourier coefficients of a trigonometric polynomial f from (1), yields $g(\boldsymbol{x}_k) = D^{\boldsymbol{s}} f(\boldsymbol{x}_k)$. Thus, for fixed $\boldsymbol{s} \in \mathbb{N}_0^d$, the mixed derivatives $D^{\boldsymbol{s}} f(\boldsymbol{x})$ of the trigonometric polynomial f can be evaluated at all rank-1 lattice nodes $\boldsymbol{x}_k$, $k = 0, \ldots, M-1$, in $\mathcal{O}(M \log M + |\mathcal{I}_N|)$ arithmetic operations.

III. NFFT based on Taylor expansion and rank-1 lattice FFT

A. Method

Let a frequency index set $\mathcal{I}_N \subset \mathbb{Z}^d \cap [-N, N]^d$ and a rank-1 lattice $\Lambda(\boldsymbol{z}, M)$ of size M be given. We replace the expansion point $\boldsymbol{a}$ in (2) by a closest rank-1 lattice node $\boldsymbol{x}_{k'} = \arg\min_{\boldsymbol{x}_k \in \Lambda(\boldsymbol{z}, M)} \mu(\boldsymbol{x}, \boldsymbol{x}_k)$, and obtain the Taylor expansion

$$s_m(\boldsymbol{x}) = \sum_{0 \leq |\boldsymbol{s}| < m} \frac{(\boldsymbol{x} - \boldsymbol{x}_{k'})^{\boldsymbol{s}}}{\boldsymbol{s}!} \sum_{\boldsymbol{l} \in \mathcal{I}_N} (-2\pi \mathrm{i} \boldsymbol{l})^{\boldsymbol{s}} \, \hat{f}_{\boldsymbol{l}} \, \mathrm{e}^{-2\pi \mathrm{i} \boldsymbol{l} \boldsymbol{x}_{k'}}. \quad (3)$$

Assuming that a closest rank-1 lattice node $\boldsymbol{x}_{k'}$ is known for each sampling node $\boldsymbol{y}_\ell$, the Taylor expansion s_m in (3) can be calculated in $\mathcal{O}\left(m^d(L + M \log M + |\mathcal{I}_N|)\right)$ arithmetic operations for all sampling nodes $\boldsymbol{y}_\ell$, $\ell = 0, \ldots, L-1$.

For symmetric hyperbolic cross index sets $\mathcal{I}_N = H_N^d$, $N \in \mathbb{N}$, $N \geq 2$, we have $|H_N^d| \leq C_{\mathrm{H}} N \log^{d-1} N$ for $N \geq 2$ with a constant $C_{\mathrm{H}} > 0$, see e.g. [12]. Choosing the rank-1 lattice size $M \sim |H_N^d|$, we obtain an arithmetic complexity of $\mathcal{O}\left(m^d(L + N \log^d N)\right)$.

B. Error estimates for symmetric hyperbolic cross index sets

Theorem III.1. *Let a trigonometric polynomial $f : \mathbb{T}^d \to \mathbb{C}$ supported on the symmetric hyperbolic cross index set $\mathcal{I}_N = H_N^d$, $f(\boldsymbol{x}) = \sum_{\boldsymbol{l}\in H_N^d} \hat{f}_{\boldsymbol{l}}\, \mathrm{e}^{-2\pi \mathrm{i}\boldsymbol{l}\boldsymbol{x}}$, $\hat{f}_{\boldsymbol{l}} \in \mathbb{C}$, $N \in \mathbb{N}$, be given. Furthermore, let $\Lambda(\boldsymbol{z}, M)$ be a rank-1 lattice with mesh norm δ. Then, for the approximation of the trigonometric polynomial f by a truncated Taylor series $s_m(\boldsymbol{x}) := \sum_{|\boldsymbol{s}|=0}^{m-1} \frac{D^{\boldsymbol{s}} f(\boldsymbol{x}_{k'})}{\boldsymbol{s}!}(\boldsymbol{x} - \boldsymbol{x}_{k'})^{\boldsymbol{s}}$ of degree $m-1$ from (3), where $m \in \mathbb{N}$ and $\boldsymbol{x}_{k'} = \arg\min_{\boldsymbol{x}_k \in \Lambda(\boldsymbol{z},M)} \mu(\boldsymbol{x}, \boldsymbol{x}_k)$, the remainder $R_m(\boldsymbol{x}) := f(\boldsymbol{x}) - s_m(\boldsymbol{x})$ is bounded by*

$$|R_m(\boldsymbol{x})| \leq \frac{d^m \pi^m}{m!}\, \delta^m N^{m-\alpha} \sum_{\boldsymbol{l}\in H_N^d} |\hat{f}_{\boldsymbol{l}}|\, r(\boldsymbol{l})^\alpha,$$

where $\alpha \in [0, m]$ is the smoothness parameter.

Proof: Let $\boldsymbol{\xi}(t) := \boldsymbol{x}_{k'} + t(\boldsymbol{x} - \boldsymbol{x}_{k'})$, $t \in [0,1]$. The remainder $R_m(\boldsymbol{x})$ can be written (cf. [13, Ch. 1]) in the form $R_m(\boldsymbol{x}) = m \int_0^1 (1-t)^{m-1} \sum_{|\boldsymbol{s}|=m} D^{\boldsymbol{s}} f(\boldsymbol{\xi}(t)) \frac{(\boldsymbol{x} - \boldsymbol{x}_{k'})^{\boldsymbol{s}}}{\boldsymbol{s}!} \mathrm{d}t$. Then,

$$\begin{aligned}
&|R_m(\boldsymbol{x})| \\
&\leq m \int_0^1 (1-t)^{m-1} \sum_{|\boldsymbol{s}|=m} |D^{\boldsymbol{s}} f(\boldsymbol{\xi}(t))| \frac{|(\boldsymbol{x} - \boldsymbol{x}_{k'})^{\boldsymbol{s}}|}{\boldsymbol{s}!} \mathrm{d}t \\
&\leq \max_{t\in[0,1]} \sum_{|\boldsymbol{s}|=m} \left| \sum_{\boldsymbol{l}\in H_N^d} (-2\pi\mathrm{i}\boldsymbol{l})^{\boldsymbol{s}} \hat{f}_{\boldsymbol{l}}\, \mathrm{e}^{-2\pi\mathrm{i}\boldsymbol{l}(\boldsymbol{\xi}(t))} \right| \frac{|(\boldsymbol{x} - \boldsymbol{x}_{k'})^{\boldsymbol{s}}|}{\boldsymbol{s}!} \\
&\leq \sum_{|\boldsymbol{s}|=m} \frac{|(\boldsymbol{x} - \boldsymbol{x}_{k'})^{\boldsymbol{s}}|}{\boldsymbol{s}!} \sum_{\boldsymbol{l}\in H_N^d} |(-2\pi\mathrm{i}\boldsymbol{l})^{\boldsymbol{s}}||\hat{f}_{\boldsymbol{l}}|.
\end{aligned}$$

Since $\mu(\boldsymbol{x}, \boldsymbol{x}_{k'}) \leq \delta/2$ and by applying the multinomial theorem, we get

$$\begin{aligned}
|R_m(\boldsymbol{x})| &\leq \sum_{|\boldsymbol{s}|=m} \frac{\left(\frac{\delta}{2}\right)^{|\boldsymbol{s}|}}{\boldsymbol{s}!} \sum_{\boldsymbol{l}\in H_N^d} |(-2\pi\mathrm{i}\boldsymbol{l})^{\boldsymbol{s}}|\, |\hat{f}_{\boldsymbol{l}}| \\
&\leq \pi^m \delta^m \sum_{\boldsymbol{l}\in H_N^d} |\hat{f}_{\boldsymbol{l}}| \sum_{|\boldsymbol{s}|=m} \frac{|l_1|^{s_1} \cdot \ldots \cdot |l_d|^{s_d}}{\boldsymbol{s}!} \\
&\leq \pi^m \delta^m \sum_{\boldsymbol{l}\in H_N^d} |\hat{f}_{\boldsymbol{l}}| \frac{\|\boldsymbol{l}\|_1^m}{m!}.
\end{aligned}$$

Introducing weights $r(\boldsymbol{l})^\alpha$, $0 \leq \alpha \leq m$, we obtain

$$\begin{aligned}
|R_m(\boldsymbol{x})| &\leq \pi^m \delta^m \sum_{\boldsymbol{l}\in H_N^d} |\hat{f}_{\boldsymbol{l}}|\, r(\boldsymbol{l})^\alpha \frac{\|\boldsymbol{l}\|_1^m}{r(\boldsymbol{l})^\alpha\, m!} \\
&\leq \frac{\pi^m \delta^m}{m!} \sum_{\boldsymbol{l}\in H_N^d} |\hat{f}_{\boldsymbol{l}}|\, r(\boldsymbol{l})^\alpha \frac{d^m\, r(\boldsymbol{l})^m}{r(\boldsymbol{l})^\alpha} \\
&\leq \frac{d^m \pi^m \delta^m}{m!} \left(\sum_{\boldsymbol{l}\in H_N^d} |\hat{f}_{\boldsymbol{l}}|\, r(\boldsymbol{l})^\alpha \right) \max_{\boldsymbol{l}\in H_N^d} r(\boldsymbol{l})^{m-\alpha} \\
&= \frac{d^m \pi^m}{m!}\, \delta^m N^{m-\alpha} \sum_{\boldsymbol{l}\in H_N^d} |\hat{f}_{\boldsymbol{l}}|\, r(\boldsymbol{l})^\alpha.
\end{aligned}$$

■

Corollary III.2. *Let a hyperbolic cross index set $\mathcal{I}_N = H_N^d$, $N \in \mathbb{N}$, $N \geq 2$, and a rank-1 lattice $\Lambda(\boldsymbol{z}, M)$ of size $M := C_{\mathrm{L}} N \log^{d-1} N \sim |H_N^d|$ for some constant $C_{\mathrm{L}} \geq 1$ be given, where the generating vector $\boldsymbol{z}$ is chosen as in the proof of Lemma II.3. Then,*

$$\begin{aligned}
|R_m(\boldsymbol{x})| &\leq \frac{d^m \pi^m}{m!}\, (C_d)^m M^{-m/d}\, N^{m-\alpha} \sum_{\boldsymbol{l}\in H_N^d} |\hat{f}_{\boldsymbol{l}}|\, r(\boldsymbol{l})^\alpha \\
&= \frac{d^m \pi^m}{m!} (C_d)^m \frac{N^{m-\alpha}}{(C_{\mathrm{L}} N \log^{d-1} N)^{\frac{m}{d}}} \sum_{\boldsymbol{l}\in H_N^d} |\hat{f}_{\boldsymbol{l}}|\, r(\boldsymbol{l})^\alpha
\end{aligned}$$

is valid for all smoothness parameters $\alpha \in [0, m]$, where $C_d > 1$ is the constant from Lemma II.3.

Proof: From Lemma II.3, we obtain that the mesh norm $\delta \leq C_d\, M^{-1/d}$. Applying Theorem III.1 yields the result. ■

Remark III.3. *If we choose the smoothness parameter $\alpha \in [\frac{d-1}{d}\, m, m]$, Corollary III.2 guarantees a decreasing relative error $|R_m(\boldsymbol{x})| / \left(\sum_{\boldsymbol{l}\in H_N^d} |\hat{f}_{\boldsymbol{l}}|\, r(\boldsymbol{l})^\alpha \right)$ for increasing refinement N. Setting the smoothness parameter $\alpha := m$ yields $|R_m(\boldsymbol{x})| \leq \frac{d^m \pi^m}{m!} (C_d)^m\, (C_{\mathrm{L}} N \log^{d-1} N)^{-\frac{m}{d}} \sum_{\boldsymbol{l}\in H_N^d} |\hat{f}_{\boldsymbol{l}}|\, r(\boldsymbol{l})^m$.*

Remark III.4. *The presented method can also be used for the approximate evaluation of trigonometric polynomials f supported on other frequency index sets. For instance, consider the case of l_1 balls, $\mathcal{I}_N = \{\boldsymbol{j} \in \mathbb{Z}^d \colon \|\boldsymbol{j}\|_1 \leq N\}$. In the proof of Theorem III.1, we introduce weights $\|\boldsymbol{l}\|_1^\alpha$ instead of $r(\boldsymbol{l})^\alpha$. Then, we obtain $|R_m(\boldsymbol{x})| \leq \frac{\pi^m}{m!}\, \delta^m\, N^{m-\alpha} \sum_{\boldsymbol{l}\in\mathcal{I}_N} |\hat{f}_{\boldsymbol{l}}|\, \|\boldsymbol{l}\|_1^\alpha$.*

IV. Numerical results

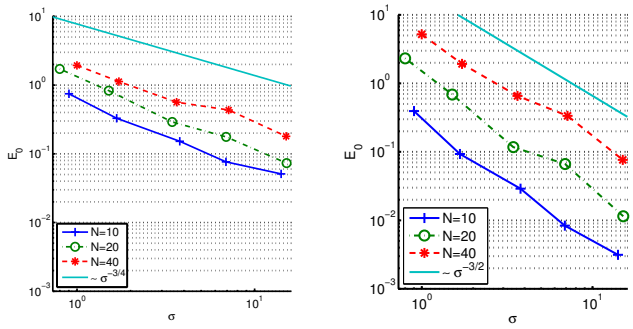
The Taylor expansion s_m in (3) was implemented in MATLAB for trigonometric polynomials f from (1) as described in Section III-A.

For symmetric hyperbolic cross index sets $\mathcal{I}_N = H_N^d$, numerical tests were performed. The generating vector $\boldsymbol{z}$ of each rank-1 lattice $\Lambda(\boldsymbol{z}, M)$ was chosen as in the proof of Lemma II.3. The maximum relative approximation error $E_\alpha := \max_{\boldsymbol{y}_\ell \in \mathcal{Y}} |R_m(\boldsymbol{y}_\ell)| / \left(\sum_{\boldsymbol{l}\in H_N^d} |\hat{f}_{\boldsymbol{l}}|\, r(\boldsymbol{l})^\alpha \right)$ was determined using $L = 100\,000$ uniformly random sampling nodes $\boldsymbol{y}_\ell \in \mathbb{T}^d$, $\mathcal{Y} := \{\boldsymbol{y}_\ell\}_{l=0}^{L-1}$.

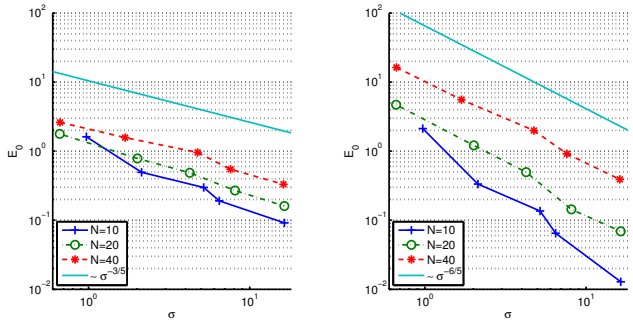
A. Decreasing error E_α for increasing rank-1 lattice size M

In this test case, we uniformly randomly chose the Fourier coefficients $\hat{f}_{\boldsymbol{l}} \in (0,1]/r(\boldsymbol{l})^\alpha$, $\boldsymbol{l} \in \mathcal{I}_N = H_N^d$. All tests were repeated five times using different Fourier coefficients $\hat{f}_{\boldsymbol{l}}$ and sampling nodes $\boldsymbol{y}_\ell$. Then, the average error of these five test runs was used.

We set the rank-1 lattice size $M := \sigma \cdot 2\,|H_N^d|$ with a factor $\sigma \geq \frac{1}{2}$. Due to Corollary III.2, the error E_α should decrease at least like $\sim \sigma^{-m/d}$ for increasing factor σ. In tests performed for the cases $d = 2, \ldots, 5$ and $m = 2, \ldots, 6$, this behaviour could be observed. Figure 1 shows the error E_0 for increasing values of factor σ for refinements $N = 10, 20, 40$ and $m = 3, 6$ in the four- and five-dimensional case as well as the lines $\sim \sigma^{-m/d}$.

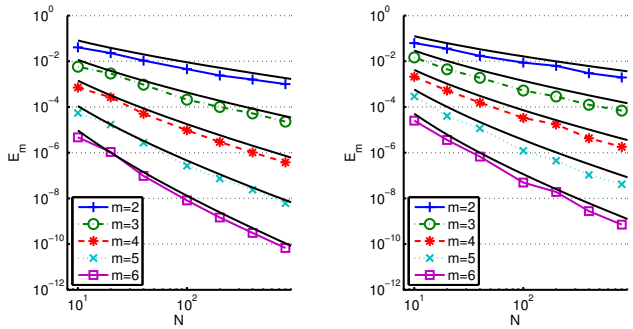


(a) $d = 4$, $m = 3$ (b) $d = 4$, $m = 6$



(c) $d = 5$, $m = 3$ (d) $d = 5$, $m = 6$

Fig. 1. Approximation error E_0 for increasing values of factor σ with rank-1 lattice size $M = \sigma\, 2|H_N^d|$ for Taylor expansions s_m of degree $m-1$, $m = 3, 6$, in the cases $d = 4, 5$.



(a) $d = 4$ (b) $d = 5$

Fig. 2. Approximation error E_m for increasing hyperbolic cross refinements N with rank-1 lattice size $M \approx 2|H_N^d|$ for Taylor expansions s_m of degree $m-1$, $m = 2, \ldots, 6$, and theoretical bounds $\sim (N \log^{d-1} N)^{-m/d}$ (solid lines without symbols) in the cases $d = 4, 5$.

B. Decreasing error E_m for increasing refinement N of the symmetric hyperbolic cross index set $\mathcal{I}_N = H_N^d$

In order to obtain a large error E_m, the Fourier coefficients $\hat{f}_{\boldsymbol{l}}$, $\boldsymbol{l} \in H_N^d$, were set to zero except $\hat{f}_{(\pm 1,0,\ldots,0)^\top} = 1$, $\hat{f}_{(0,\pm 1,0,\ldots,0)^\top} = 1$, $\ldots$, $\hat{f}_{(0,\ldots,0,\pm 1)^\top} = 1$ and $\hat{f}_{(\pm N,0,\ldots,0)^\top} = 1/N^m$, $\hat{f}_{(0,\pm N,0,\ldots,0)^\top} = 1/N^m$, $\ldots$, $\hat{f}_{(0,\ldots,0,\pm N)^\top} = 1/N^m$. We set the rank-1 lattice size $M \approx 2|H_N^d|$. Test cases included Taylor expansion degrees $m-1$, $m = 2, \ldots, 6$, and refinements up to $N = 10^4$ for $d = 2$, up to $N = 10^3$ for $d = 3$ and up to $N = 800$ for $d = 4, 5$. Remark III.3 states, that the error E_m should decrease at least like $\sim (N \log^{d-1} N)^{-\frac{m}{d}}$. In the results of the performed tests, a decrease of $\sim (N \log^{d-1} N)^{-\frac{m}{d}}$ could be observed. Figure 2 shows the results for the cases $d = 4, 5$.

V. Conclusion

Based on rank-1 lattice methods and Taylor expansion, we presented a method for the fast approximate evaluation of trigonometric polynomials f with frequencies supported on symmetric hyperbolic cross index sets $\mathcal{I}_N = H_N^d$ with refinement N at arbitrary sampling nodes $\boldsymbol{y}_\ell \in \mathbb{T}^d$, $\ell = 0, \ldots, L-1$. We showed conditions which guarantee a decreasing approximation error $|R_m(\boldsymbol{x})| / \left(\sum_{\boldsymbol{l} \in H_N^d} |\hat{f}_{\boldsymbol{l}}|\, r(\boldsymbol{l})^\alpha \right)$ for increasing refinement N. In particular for smoothness parameter $\alpha = m$, a rank-1 lattice $\Lambda(\boldsymbol{z}, M)$ of size $M \sim |H_N^d|$ exists, such that the approximation error decreases at least like $\sim (N \log^{d-1} N)^{-m/d}$ for increasing refinement N. For such a rank-1 lattice of size $M \sim |H_N^d|$, the total arithmetic complexity of the presented method is $\mathcal{O}(m^d L + m^d\, N \log^d N)$. The results of the numerical tests confirmed the theoretical upper bounds.

Acknowledgment

The author thanks Daniel Potts, Stefan Kunis and Lutz Kämmerer for numerous valuable discussions on the presented subject as well as the referees for their valuable suggestions. Moreover, he gratefully acknowledges support by German Research Foundation within the project PO 711/10-2.

References

[1] J. Keiner, S. Kunis, and D. Potts, "Using NFFT3 - a software library for various nonequispaced fast Fourier transforms," *ACM Trans. Math. Software*, vol. 36, pp. Article 19, 1 – 30, 2009.

[2] C. Anderson and M. Dahleh, "Rapid computation of the discrete Fourier transform," *SIAM J. Sci. Comput.*, vol. 17, pp. 913 – 919, 1996.

[3] S. Kunis, "Nonequispaced fast Fourier transforms without oversampling," *Proc. Appl. Math. Mech.*, vol. 8, pp. 10 977 – 10 978, 2008.

[4] M. Döhler, S. Kunis, and D. Potts, "Nonequispaced hyperbolic cross fast Fourier transform," *SIAM J. Numer. Anal.*, vol. 47, pp. 4415 – 4428, 2010.

[5] K. Hallatschek, "Fouriertransformation auf dünnen Gittern mit hierarchischen Basen," *Numer. Math.*, vol. 63, pp. 83 – 97, 1992.

[6] V. Gradinaru, "Fourier transform on sparse grids: Code design and the time dependent Schrödinger equation," *Computing*, vol. 80, pp. 1 – 22, 2007.

[7] L. Kämmerer and S. Kunis, "On the stability of the hyperbolic cross discrete Fourier transform," *Numer. Math.*, vol. 117, pp. 581 – 600, 2011.

[8] L. Kämmerer, S. Kunis, and D. Potts, "Interpolation lattices for hyperbolic cross trigonometric polynomials," *J. Complexity*, vol. 28, pp. 76 – 92, 2012.

[9] D. Potts and M. Tasche, "Numerical stability of nonequispaced fast Fourier transforms," *J. Comput. Appl. Math*, vol. 222, pp. 655 – 674, 2008.

[10] G. Steidl and M. Tasche, "Index transforms for multidimensional DFT's and convolutions," *Numer. Math.*, vol. 56, pp. 513 – 528, 1989.

[11] N. Amersi, O. Beckwith, S. J. Miller, R. Ronan, and J. Sondow, "Generalized Ramanujan Primes," *ArXiv e-prints*, Aug. 2011.

[12] L. Kämmerer, "Reconstructing hyperbolic cross trigonometric polynomials by sampling along rank-1 lattices," *submitted*, 2012. [Online]. Available: http://www.tu-chemnitz.de/~lkae/paper/Kae2012.pdf

[13] L. Hörmander, *The analysis of linear partial differential operators: Distribution theory and Fourier analysis*. Springer-Verlag, 1990.

Decoupling of Fourier Reconstruction System for Shifts of Several Signals

Yosef Yomdin*†, Niv Sarig*‡ and Dmitry Batenkov*§
*Department of Mathematics,Weizmann Institute of Science, Rehovot 76100, Israel
†Email: yosef.yomdin@weizmann.ac.il
‡Email: niv.sarig@weizmann.ac.il
§Email: dima.batenkov@weizmann.ac.il

***Abstract*—We consider the problem of "algebraic reconstruction" of linear combinations of shifts of several signals $f_1, \dots, f_k$ from the Fourier samples. For each $r = 1, \dots, k$ we choose sampling set S_r to be a subset of the common set of zeroes of the Fourier transforms $\mathcal{F}(f_\ell),\ \ell \neq r$, on which $\mathcal{F}(f_r) \neq 0$. We show that in this way the reconstruction system is reduced to k separate systems, each including only one of the signals f_r. Each of the resulting systems is of a "generalized Prony" form. We discuss the problem of unique solvability of such systems, and provide some examples.**

I. Introduction

In this paper we consider reconstruction of signals of the following a priori known form:

$$F(x) = \sum_{j=1}^{k} \sum_{q=1}^{q_j} a_{jq} f_j(x - x_{jq}), \tag{1.1}$$

with $a_{jq} \in \mathbb{R},\ x_{jq} = (x_{jq}^1, \dots, x_{jq}^n) \in \mathbb{R}^n$. We assume that the signals $f_1, \dots, f_k : \mathbb{R}^n \to \mathbb{R}$ are known (in particular, their Fourier transforms $\mathcal{F}(f_j)$ are known), while $a_{jq},\ x_{jq}$ are the unknown signal parameters, which we want to find from Fourier samples of F. We explicitly assume here that $k \geq 2$. So the usual methods which allow one to solve this problem "in closed form" in the case of shifts of a single function (see [6], [2], [16]) are not directly applicable. Still, we shall show that in many cases an explicit reconstruction from a relatively small collection of Fourier samples of F is possible. Practical importance of signals as above is well recognized in the literature: for some discussions and similar settings see, e.g. [6], [8], [13].

We follow a general line of the "Algebraic Sampling" approach (see [6], [15], [3] and references therein), i.e. we reconstruct the values of the unknown parameters, solving a system of non-linear equations, imposed by the measurements (system (2.1) below). The equations in this system appear as we equate the "symbolic" expressions of the Fourier samples, obtained from (1.1), to their actual measured values.

Our specific strategy is as follows: we choose a sampling set $S_r \subset \mathbb{R}^n,\ r = 1, \dots, k$, in a special way, in order to "decouple" (2.1), and to reduce it to k separate systems, each including only one of the signals f_r. To achieve this goal we take S_r to be a subset of the common set of zeroes of the Fourier transforms $\mathcal{F}(f_\ell),\ \ell \neq r$.

The decoupled systems turn out to be of a "generalized Prony" type:

$$\sum_{j=1}^{N} a_j y_j^{s_\ell} = m_\ell, \quad \ell = 1, 2, \dots,\ s_\ell \in S \subset \mathbb{R}^n. \tag{1.2}$$

The standard Prony system, where the sample set S is the set of integer points in a cube of a prescribed size, allows for a solution "in closed form" (see, for example, [2], [14], [16], [17] and references therein). We are not aware of any method for an explicit solution of generalized Prony systems. However, "generic" solution methods can be applied. Their robustness can be estimated via Turán-Nazarov inequality for exponential polynomials and its discrete version ([7], [12]). Some initial results in this direction have been presented in [16], [2]. Below we further extend these results, restricting ourselves to the uniqueness problem only.

II. Reconstruction System and its Decoupling

For F of the form (1.1) and for any $s = (s^1, \dots, s^n) \in \mathbb{R}^n$ we have for the sample of the Fourier transform $\mathcal{F}(F)$ at s

$$\begin{aligned} \mathcal{F}(F)(s) &= \int_{\mathbb{R}^n} e^{-2\pi i s x} F(x) dx \\ &= \sum_{j=1}^{k} \sum_{q=1}^{q_j} a_{jq} e^{-2\pi i s x_{jq}} \mathcal{F}(f_j)(s). \end{aligned}$$

So taking samples at the points $s_\ell = (s_\ell^1, \dots, s_\ell^n)$ of the sample set $S = \{s_1, \dots, s_m\}$, and denoting the vector $e^{-2\pi i x_{jq}} = (e^{-2\pi i x_{jq}^1}, \dots, e^{-2\pi i x_{jq}^n})$ by $y_{jq} = (y_{jq}^1, \dots, y_{jq}^n)$ we get our reconstruction system in the form

$$\sum_{j=1}^{k} \sum_{q=1}^{q_j} a_{jq} \mathcal{F}(f_j)(s_\ell) y_{jq}^{s_\ell} = \mathcal{F}(F)(s_\ell),\ \ell = 1, \dots, m, \tag{2.1}$$

in the standard multi-index notations. In system (2.1) the right hand sides $\mathcal{F}(F)(s_\ell)$ are the known measurements, while the

This research was supported by the Adams Fellowship Program of the Israeli Academy of Sciences and Humanities, ISF Grant No. 639/09, and by the Minerva foundation. We would like to thank the referees for useful corrections and remarks.

Fourier samples $\mathcal{F}(f_j)(s_\ell)$ are known by our assumptions. The unknowns in (2.1) are the amplitudes a_{jq} and the shifts x_{jq}, encoded in the vectors y_{jq}.

In the case $k = 1$ we could divide the equations in (2.1) by $\mathcal{F}(f_1)(s_\ell)$ and obtain directly a Prony-like system. However, for $k \geq 2$ this transformation usually is not applicable. Instead we "decouple" system (2.1) with respect to the signals $f_1, \dots, f_k$ using the freedom in the choice of the sample set S. Let

$$Z_\ell = \left\{x \in \mathbb{R}^n,\ \mathcal{F}(f_\ell)(x) = 0\right\}$$

denote the set of zeroes of the Fourier transform $\mathcal{F}(f_\ell)$. For each $r = 1, \dots, k$ we take the sampling set S_r to be a subset of the set

$$W_r = (\bigcap_{\ell \neq r} Z_\ell) \setminus Z_r$$

of common zeroes of the Fourier transforms $\mathcal{F}(f_\ell),\ \ell \neq r$, but not of $\mathcal{F}(f_r)$. For such S_r all the equations in (2.1) vanish, besides those with $j = r$. Hence we obtain:

Proposition 2.1: Let for each $r = 1, \dots, k$ the sampling set S_r satisfy

$$S_r = \{s_{r1}, \dots, s_{rm_r}\} \subset W_r.$$

Then for each r the corresponding system (2.1) on the sample set S_r takes the form

$$\sum_{q=1}^{q_r} a_{rq} y_{rq}^{s_{r\ell}} = c_{r\ell}(F),\ \ell = 1, \dots, m_r, \qquad (2.2)$$

where $c_{r\ell}(F) = \mathcal{F}(F)(s_{r\ell})/\mathcal{F}(f_r)(s_{r\ell})$. □

So (2.1) is decoupled into k generalized Prony systems (2.2), each relating to the shifts of the only signal f_r. The problem is that some (or all) of the sets W_r may be too small, and the resulting systems (2.2) will not allow us to reconstruct the unknowns a_{rq} and y_{rq}. Another problem is instability of zero finding, which may lead to only approximate zeroes of Fourier transforms. We have at present only initial results outlying applicability of the Fourier decoupling method ([16]). In a "good" case where the zero sets Z_ℓ of the Fourier transforms $\mathcal{F}(f_\ell),\ \ell = 1, \dots, k$, are nonempty $n-1$-dimensional hypersurfaces meeting one another transversally, still for $k > n + 1$ the intersection of $Z_\ell,\ \ell \neq r$, is empty. So the resulting systems (2.2) contain no equations. Hence we can apply the above decoupling only for $k \leq n + 1$.

Some specific examples, as well as investigation of the conditions on $f_1, \dots, f_k$ which provide solvability of systems (2.2) were presented in [16]. In one-dimensional case $(n = 1, k = 2)$ these conditions can be given explicitly. In this case $W_1 = W_1(f_1, f_2)$ consists of zeroes of $\mathcal{F}(f_2)$ which are not zeroes of $\mathcal{F}(f_1)$, and $W_2 = W_2(f_1, f_2)$ consists of zeroes of $\mathcal{F}(f_1)$ which are not zeroes of $\mathcal{F}(f_2)$. The following result has been proved (for real Prony systems) in [16]. Here we extend it to the case of system (2.2) which has purely imaginary exponents. The constant $2N$ below is sharp, in contrast with the constant $C(n, d)$ in (multidimensional) Theorem 4.1 below.

Let in (1.1) $n = 1, k = 2$, and let $q_1 = q_2 = N$. Assume that for the signals f_1, f_2 in (1.1) each of the sets W_1 and W_2 contains at least $2N$ elements. Let $D_j,\ j = 1, 2$, be the length of the shortest interval Δ_j such that $S_j = \Delta_j \cap W_j$ contains exactly $2N$ elements, and let $\rho_j = \frac{1}{D_j}$.

Theorem 2.1: For shifts x_{jq} in the interval $[0, \rho_j),\ j = 1, 2$, systems (2.2) with the sampling sets S_1, S_2 are uniquely solvable.

Proof: Let us fix $j = 1$. The proof for $j = 2$ is the same. Substituting $y_{1q} = e^{-2\pi i x_{1q}}$ associates to a solution $(a_{1q}, y_{1q}),\ q = 1, \dots, N$, of (2.2) an exponential polynomial $H(s) = \sum_{q=1}^{N} a_{1q} e^{-2\pi i x_{1q} s}$ with purely imaginary exponents. If (2.2) has two different solutions, the corresponding exponential polynomials $H_1(s)$ and $H_2(s)$ are equal for each $s \in S_1$. Hence S_1 is a set of zeroes of $H_2(s) - H_1(s)$, which is an exponential polynomial of the order at most $2N$. On the other hand, by Langer's lemma (Lemma 1.3 in [12]) such polynomial can have in each interval of length D at most $2N - 1 + \frac{\rho D}{2\pi}$ zeroes, where ρ is the maximum of the absolute values of the exponents. In our case $D = D_1$ and $\rho < 2\pi\rho_1 = \frac{2\pi}{D_j}$. Hence $\frac{\rho D}{2\pi}$ is strictly less than 1, and so the number of zeroes of $H_2 - H_1$ is at most $2N - 1$, in contradiction with the assumptions. □

III. Examples

Some examples of Fourier decoupling have been presented in [16]. In these examples the sets W_r are "large enough" to reduce the problem (with the number of allowed shifts fixed but arbitrarily large) to a set of decoupled standard Prony systems.

In dimension one we can take, for example, f_1 to be the characteristic function of the interval $[-1, 1]$, while $f_2(x) = \delta(x - 1) + \delta(x + 1)$. So we consider signals of the form

$$F(x) = \sum_{q=1}^{N} [a_{1q} f_1(x - x_{1q}) + a_{2q} f_2(x - x_{2q})]. \qquad (3.1)$$

Easy computations show that

$$\mathcal{F}(f_1)(s) = \sqrt{\frac{2}{\pi}} \frac{\sin s}{s}$$

and

$$\mathcal{F}(f_2)(s) = \sqrt{\frac{2}{\pi}} \cos s.$$

So the zeros of the Fourier transform of f_1 are the points $\pi n,\ n \in \mathbb{Z} \setminus \{0\}$ and those of f_2 are the points $(\frac{1}{2} + n)\pi,\ n \in \mathbb{Z}$. These sets do not intersect, so $W_1 = \{\pi n\}$, and $W_2 = \{(\frac{1}{2} + n)\pi\}$. Since W_1 and W_2 are just shifted integers $\mathbb{Z}$, the generalized Prony systems in (2.2) are actually the standard ones. For f_2 the system (2.2) takes the form

$$\frac{\mathcal{F}(F)(\pi n)}{\sqrt{\frac{2}{\pi}}(-1)^n} = \sum_{q=1}^{N} a_{2q} (y_{2q})^{\pi n},\ n \in \mathbb{Z}.$$

If we denote $M_n = \frac{\mathcal{F}(F)(\pi n)}{\sqrt{\frac{2}{\pi}}(-1)^n}$, $A_q = a_{2q}(y_{2q})^\pi$ and $\eta_q = (y_{2q})^\pi$ we get the usual Prony system

$$M_n = \sum_{q=0}^{N} A_q \eta_q^n \ , \ n \in \mathbb{Z}.$$

For f_1 we get

$$\frac{\mathcal{F}(F)((\frac{1}{2}+n)\pi)}{\sqrt{\frac{2}{\pi}}\frac{(-1)^{n+1}}{(\frac{1}{2}+n)\pi}} = \sum_{q=1}^{N} a_{1q}(y_{1q})^{(\frac{1}{2}+n)\pi} \ , \ n \in \mathbb{Z}.$$

In this case we denote $\mu_n = \frac{\mathcal{F}(F)((\frac{1}{2}+n)\pi)}{\sqrt{\frac{2}{\pi}}\frac{(-1)^{n+1}}{(\frac{1}{2}+n)\pi}}$, $\alpha_q = a_{1q}(y_{1q})^{\frac{\pi}{2}}$ and $\xi_q = (y_{1q})^\pi$ and we get again the usual Prony system

$$\mu_n = \sum_{q=1}^{N} \alpha_q \xi_q^n, \ n \in \mathbb{Z}.$$

Solving these two systems by any standard method will give us the translations and amplitudes of the functions f_1, f_2. Notice that a possible non-uniqueness of the solutions is imposed here by the substitutions $\eta_q = (y_{2q})^\pi$ and $\xi_q = (y_{1q})^\pi$.

In dimension two we can take, in particular, f_1, f_2, f_3 to be the characteristic functions of the three squares: $Q_1 = [-3,3]^2$, $Q_2 = [-5,5]^2$, and Q_3 which is the rotation of the square $[-\sqrt{2},\sqrt{2}]^2$ by $\frac{\pi}{4}$. So we put

$$\chi_j(x) = \begin{cases} 1 & x \in Q_j \\ 0 & x \notin Q_j \end{cases} \tag{3.2}$$

and consider signals of the form

$$F(x) = \sum_{j=1}^{3}\sum_{q=1}^{q_j} a_{jq}\chi_j(x - x_{jq}), \qquad \text{with } a_{jq} \in \mathbb{R},\ x_{jq} \in \mathbb{R}^3. \tag{3.3}$$

The following result is proved in [16]:

Proposition 3.1: The zero sets Z_1, Z_2 and Z_3 of the Fourier transforms of the three functions χ_1, χ_2 and χ_3 intersect each other in such a way that the decoupling procedure based on the sets $W_1 = (Z_2 \cap Z_3) \setminus Z_1, W_2 = (Z_3 \cap Z_1) \setminus Z_2$ and $W_3 = (Z_1 \cap Z_2) \setminus Z_3$ provides three standard Prony systems for the shifts of each of the functions.

Sketch of the proof: Simple calculation gives

$$\begin{aligned} \mathcal{F}(\chi_1)(\omega,\rho) &= 4\tfrac{\sin 3\omega}{\omega} \cdot \tfrac{\sin 3\rho}{\rho} \\ \mathcal{F}(\chi_2)(\omega,\rho) &= 4\tfrac{\sin 5\omega}{\omega} \cdot \tfrac{\sin 5\rho}{\rho} \\ \mathcal{F}(\chi_3)(\omega,\rho) &= 8\tfrac{\sin\frac{\omega+\rho}{2}}{\frac{\omega+\rho}{2}} \cdot \tfrac{\sin\frac{\omega-\rho}{2}}{\frac{\omega-\rho}{2}}. \end{aligned} \tag{3.4}$$

So Z_1 is the union of horizontal or vertical lines crossing the Fourier plane's axes at $(0, \frac{n\pi}{3})$ or $(\frac{n\pi}{3}, 0)$ respectively, for all non zero integer n. Similarly for Z_2, with the only difference that the lines cross the axes at $(0, \frac{n\pi}{5})$ or $(\frac{n\pi}{5}, 0)$.
Z_3 is the union of lines with slopes 1 or -1 crossing the ω axis at $2\pi n$ for some non zero integer n. Hence for any two integers n and m we have $(\frac{1+5n}{5}, \frac{1+5n}{5}) \in S_1, (\frac{1+3m}{3}, \frac{1+3m}{3}) \in S_2$ and since $\frac{1+3m}{3} \pm \frac{1+5n}{5}$ is not an integer, $(\frac{1+3m}{3}, \frac{1+5n}{5}) \in S_3$. These three points form a triangle which repeats itself as a periodic pattern. Appropriate transformations now bring the decoupled systems (2.2) to the form of the standard two-dimensional Prony system. See [16], [2] for a new approach to solving such systems and for the results of numerical simulations. □

IV. Uniqueness of Reconstruction

Application of Proposition 2.1 prescribes the choice of sample points from the common zeroes of the Fourier transforms $\mathcal{F}(f_j)$. So the geometry of the sample sets S_r may be complicated, and the known results on unique solvability of the standard Prony system ([2], [4], [14], [17]) are not directly applicable. Non-Uniform Sampling in Prony-type systems is also essential in other problems of algebraic signal reconstruction. In particular, recently it appeared as a key point in a proof of the Eckhoff conjecture, related to the accuracy of reconstruction of piecewise-smooth functions from their Fourier samples ([1]).

There are results on a behavior of exponential polynomials on arbitrary sets, which can provide important information on unique solvability and robustness of the generalized Prony system. In particular, this concerns the Turan-Nazarov inequality ([12]), and its extension to discrete sets obtained in [7]. In this last paper for each set S a quantity $\omega_D(S)$ has been introduced, measuring, essentially, the robustness of solvability of a generalized Prony system with the sample points $s_\ell \in S$. Here D comprises the "discrete" parameters of the Prony system to be solved. $\omega_D(S)$ can be explicitly estimated in terms of the metric entropy of S (see below), and we expect that in many important cases the quantity $\omega_D(W_r)$ for the zeroes sets W_r of the Fourier transforms $\mathcal{F}(f_j)$ can be effectively bounded from below. Some initial results and discussions in this direction, mainly in dimension one, are presented in [16], [3]. In the present paper we do not consider robustness of the Prony system, but provide a new multi-dimensional result on the uniqueness of solutions, in the lines of [16], [7] and Theorem 2.1 above.

Let us recall that for Z a bounded subset of $\mathbb{R}^n$, and for $\epsilon > 0$ the covering number $M(\epsilon, Z)$ is the minimal number of ϵ-balls in $\mathbb{R}^n$, covering Z. The ϵ-entropy $H(\epsilon, Z)$ is the binary logarithm of $M(\epsilon, Z)$.

Let $H(s) = \sum_{j=1}^{d} a_j e^{\lambda_j \cdot s}$, with $a_j \in \mathbb{R}$, $\lambda_j = (\lambda_{j1}, \ldots, \lambda_{jn}) \in \mathbb{R}^n$, be a real exponential polynomial in $s \in \mathbb{R}^n$. Denote $Z(H)$ the set of zeroes of H in $\mathbb{R}^n$, and let Q_R^n be the cube in $\mathbb{R}^n$ with the edge R. The following result is a special case of Lemma 3.3 proved in [7]:

Proposition 4.1: For each $R > 0$, and ϵ with $R > \epsilon > 0$ we have $M(\epsilon, Z(H) \cap Q_R^n) \le C(d,n)(\frac{R}{\epsilon})^{n-1}$. □

The explicit expression for $C(d,n)$ is given in [7], via Khovanski's bound ([9]) for "fewnomial" systems. Consider now a generalized Prony system (1.2) with a finite set S of samples allowed:

$$\sum_{j=1}^{N} a_j y_j^{s_\ell} = m_\ell,\ s_\ell \in S = \{s_1, \dots, s_m\} \subset \mathbb{R}^n. \qquad (4.1)$$

We shall consider only real solutions of (4.1) with y_j having all its coordinates positive.

Theorem 4.1: Let $S = \{s_1, \dots, s_m\} \subset Q_R^n$ be given, such that for a certain $\epsilon > 0$ we have $M(\epsilon, S) > C(2N, n)(\frac{R}{\epsilon})^{n-1}$. Then system (4.1) has at most one solution.

Proof: Associate to a solution $(a_j, y_j),\ j = 1, \dots, N$, of (4.1) an exponential polynomial $H(s) = \sum_{j=1}^{N} a_j e^{\lambda_j \cdot s}$, where $y_j = e^{\lambda_j},\ \lambda_j \in \mathbb{R}^n$. If (4.1) has two different solutions, the corresponding exponential polynomials $H_1(s)$ and $H_2(s)$ are equal for each $s = s_\ell \in S$. Hence S is a set of zeroes of $H_2(s) - H_1(s)$, which is an exponential polynomial of order at most $2N$. By Proposition 4.1 we have $M(\epsilon, S) \leq C(2N, n)(\frac{R}{\epsilon})^{n-1}$ for each $\epsilon > 0$, in contradiction with the assumptions of the theorem. □

Informally, Theorem 4.1 claims that finite sets S which cover (in a "resolution ϵ", for some $\epsilon > 0$), a significant part of the cube Q_R^n, are uniqueness sets of the Prony system. The condition of Theorem 4.1 on the sampling set S is quite robust with respect to the geometry of S, so we can explicitly verify it in many cases. In particular, for non-regular lattices we get the following result:

Definition 4.1: For fixed positive $\alpha < \frac{1}{2}$ and $h > 0$, a set $Z' \subset \mathbb{R}^n$ is called an (α, h)-net if it possesses the following property: there exists a regular grid Z with the step h in $\mathbb{R}^n$ such that for each $z' \in Z'$ there is $z \in Z$ with $||z' - z|| \leq \alpha h$, and for each $z \in Z$ there is $z' \in Z'$ with $||z' - z|| \leq \alpha h$.

Corollary 4.1: Let $Z' \subset \mathbb{R}^n$ be an (α, h)-net. Then for $R > C(2N)h(1-2\alpha)^{1-n}$ the set $S = Z \cap Q_R^n$ is a uniqueness set of the Prony system (4.1).

Proof: By definition, for each $z \in Z$ we can find $z' \in Z'$ inside the αh-ball around z. Clearly, any two such points are $h' = (1 - 2\alpha)h$-separated. So for each $\epsilon < h'$ we have $M(\epsilon, S) \geq |Z \cap Q_R^n| = (\frac{R}{h})^n$. We conclude that the inequality $(\frac{R}{h})^n > C(2N)(\frac{R}{h'})^{n-1}$, or $R > C(2N)h(1-2\alpha)^{1-n}$ implies the condition of Theorem 4.1. □

The condition of Theorem 4.1 can be verified in many other situations, under natural assumptions on the sample set S. In particular, using integral-geometric methods developed in [5], it can be checked for the zero sets of Fourier transforms of various types of signals. We plan to present these results separately.

Remark The restriction to only positive solutions of Prony system is very essential for the result of Theorem 4.1. Indeed, consider the Prony system

$$a_1 x_1^k + a_2 x_2^k = m_k,\ k = 0, 1, \dots. \qquad (4.2)$$

If we put $a_1 = 1,\ x_1 = 1,\ a_2 = -1,\ x_2 = -1$, then $m_k = 1^k - (-1)^k = 0$ for each even k. So the regular grid of even integers is not a uniqueness set for system (4.2). This fact is closely related to the classical Skolem-Mahler-Lech Theorem (see [10], [11], [18] and references therein) which says that the integer zeros of an exponential polynomial are the union of complete arithmetic progressions and a finite number of exceptional zeros. So such sets may be non-uniqueness sample sets for complex Prony systems.

The proof of the Skolem-Mahler-Lech Theorem is relied on non-effective arithmetic considerations. Recently the problem of obtaining effective such theorem was discussed in [18]. This problem may turn to be important for understanding of complex solutions of Prony systems. One can wonder whether the methods of Khovanskii ([9]) and Nazarov ([12]), as well as their combination in [7], can be applied here.

REFERENCES

[1] D. Batenkov. Complete Algebraic Reconstruction of Piecewise-Smooth Functions from Fourier Data. *submitted. Arxiv:1211.0680.*

[2] D. Batenkov, N. Sarig, and Y. Yomdin. An "algebraic" reconstruction of piecewise-smooth functions from integral measurements. *Functional Differential Equations*, 19(1-2):9–26, 2012.

[3] D. Batenkov and Y. Yomdin. Algebraic reconstruction of piecewise-smooth functions from Fourier data. *Proc. of Sampling Theory and Applications (SAMPTA)*, 2011.

[4] D. Batenkov and Y. Yomdin. On the accuracy of solving confluent Prony systems. *SIAM J. Appl. Math.*, 73(1):134–154, 2013.

[5] G. Comte and Y. Yomdin. Rotation of trajectories of Lipschitz vector fields *J. Differential Geom. 81 (2009), no. 3*, 601–630.

[6] P.L. Dragotti, M. Vetterli and T. Blu. *Sampling Moments and Reconstructing Signals of Finite Rate of Innovation: Shannon Meets Strang-Fix*, IEEE Transactions on Signal Processing, Vol. 55, Nr. 5, Part 1, pp. 1741-1757, 2007.

[7] O. Friedland and Y. Yomdin. An observation on Turán-Nazarov inequality. *arXiv preprint arXiv:1107.0039*, 2011.

[8] K. Gedalyahu, R. Tur, and Y.C. Eldar. Multichannel sampling of pulse streams at the rate of innovation. *IEEE Transactions on Signal Processing*, 59(4):1491–1504, 2011.

[9] A. G. Khovanskii. *Fewnomials*. Translated from the Russian by Smilka Zdravkovska. *Translations of Mathematical Monographs*, 88. American Mathematical Society, Providence, RI, 1991. viii+139 pp.

[10] C. Lech. A Note on Recurring Series. *Ark. Mat. 2, 417-421, 1953.*

[11] G. Myerson and A. J. van der Poorten, Some Problems Concerning Recurrence Sequences. *Amer. Math. Monthly* 102, 698-705, 1995.

[12] F.L. Nazarov. Local estimates of exponential polynomials and their applications to inequalities of uncertainty principle type. *St Petersburg Mathematical Journal*, 5(4):663–718, 1994.

[13] T. Peter, D. Potts, and M. Tasche. Nonlinear approximation by sums of exponentials and translates. *SIAM Journal on Scientific Computing*, 33(4):1920, 2011.

[14] B.D. Rao and K.S. Arun. Model based processing of signals: A state space approach. *Proceedings of the IEEE*, 80(2):283–309, 1992.

[15] N. Sarig and Y. Yomdin. Signal Acquisition from Measurements via Non-Linear Models. *Mathematical Reports of the Academy of Science of the Royal Society of Canada*, 29(4):97–114, 2008.

[16] N. Sarig. *Algebraic reconstruction of "shift-generated" signals from integral measurements*. PhD thesis, Weizmann Institute of Science, 2010.

[17] P. Stoica and R.L. Moses. *Spectral analysis of signals*. Pearson/Prentice Hall, 2005.

[18] T. Tao. http://terrytao.wordpress.com/2007/05/25/open-question-effective-skolem-mahler-lech-theorem/comment-46954.

DIGITAL CALIBRATION OF SAR ADC

Yun Chiu[1], Wenbo Liu[2], Pingli Huang[3], Foti Kacani[1], Gary Wang[1], Brian Elies[1], and Yuan Zhou[1]

[1]Analog and Mixed-Signal Lab
Texas Analog Center of Excellence
University of Texas at Dallas
Email: chiu.yun@utdallas.edu

[2]Apple Inc.
Mountain View, CA

[3]Analog Devices Inc.
Wilmington, MA

ABSTRACT

Four techniques for digital background calibration of SAR ADC are presented and compared. Sub-binary redundancy is the key to the realization of these techniques. Some experimental and simulation results are covered to support the effectiveness of these techniques.

Keywords—SAR ADC, digital background calibration, DAC mismatch, bit weight, sub-binary redundancy

1. INTRODUCTION

The return of switched-capacitor successive-approximation-register (SAR) analog-to-digital converter (ADC) has revealed the potential of the SAR conversion architecture in scaled technology for low-power operation [1]-[6]. Without the need for precision amplification, the analog operation of a SAR ADC is mostly switching type, similar to digital logic circuits. In addition, the single zero-crossing comparator employed in the bit-decision cycles is largely immune to offset errors, further reducing the analog design effort. One important trend in recent SAR works is the proliferation of the so-called *sub-binary* SAR architecture, which has fueled a continuous improvement on the conversion speed as well as the robustness of SAR.

In scaled technology, the signal-to-noise ratio (SNR) and linearity performance of SAR ADC are largely limited by the decreasing supply voltage and the static component mismatch errors of the digital-to-analog converter (DAC) used to produce the successive decision thresholds during the bit cycles. Consequently, while many recent SAR works have reported outstanding power efficiency, few demonstrate >10 effective number of bits (ENOB) [1]-[6].

In this paper, a few digital background calibration techniques aiming at lifting the static DAC mismatch errors in SAR conversion are presented. It will be shown that with these linearization techniques resolutions beyond 12 bits are achievable with a single SAR structure employing small capacitors.

2. SUB-BINARY SAR AND REDUNDANCY

A conventional SAR ADC employs a conversion algorithm termed binary search, in which the analog search range is halved in each successive bit-decision cycle. Exactly N steps are needed to resolve an N-bit word. The successive search ranges are usually set by a binary-weighted DAC which produces one analog level in each cycle to be compared with the sampled input. For example, during the MSB cycle the DAC produces a level corresponding to the code 10···0 or 01···1, one of the two codes closest to the midpoint of the ADC full scale V_{FS}. When the component matching of the DAC is ideal, the difference between these two codes is only one LSB, thus either choice yields no detectable difference to the ADC outcome. However, when mismatch is present the two choices will lead to drastically different results, which are better explained by the conversion curves illustrated in Fig. 1 for two scenarios. In the first case, the DAC MSB component is greater in value than the summation of all lower-rank components, resulting in two disjoint segments of the conversion curve (Fig. 1a). Since the analog input levels between the codes 01···1 and 10···0 all resolve to

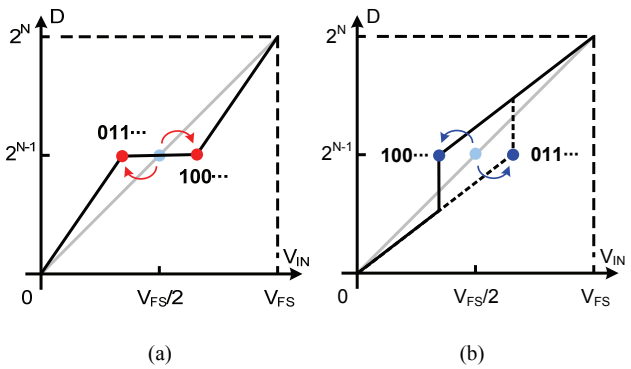


Fig. 1. SAR conversion curve due to MSB component mismatch: (a) super-binary and (b) sub-binary. The grey line indicates the ideal curve. V_{in} is the analog input and D is the raw decision vector ($D = \{d_i, i = 0…N-1\}$).

one digital LSB, the loss of information is irrecoverable from the digital domain alone. We term this scenario *super-binary*. The opposite case is shown in Fig. 1b, in which the analog levels corresponding to the codes 01···1 and 10···0 are swapped, resulting in an overlapped analog input range sandwiched by the two codes. We term this scenario *sub-binary*. Typically, depending on the pre-choice of the decision threshold, i.e., 01···1 or 10···0, during bit cycles, either the upper (solid) or lower (dashed) curve but not both will be exercised. The vertical discontinuity of the conversion curve thus leads to the absence of a chunk of digital levels termed missing codes.

A unique feature of the sub-binary conversion scheme is that if both the upper and lower segments within the overlapped range can be artificially enabled, any analog level inside this range can be mapped to two digital codes differing by the vertical distance between the two segments. We term this phenomenon *decision redundancy* or *architectural redundancy*. With a sub-binary architecture, it can be shown that the conversion nonlinearity due to missing codes can be fully corrected in digital domain under certain assumptions if the optimal bit weights are known [3],

$$V_{in} = V_R \sum_{i=0}^{N-1} \frac{C_i}{C_{tot}} (2d_i - 1) + QN, \qquad (1)$$

where V_{in} is the sampled input, V_R is the reference voltage, QN is the quantization noise, and the ratio of the i^{th} capacitor (C_i) to the total capacitance of the DAC (C_{tot}) defines the i^{th} bit weight w_i. Eq. (1) essentially guarantees that any two analog input levels at least one analog LSB apart will resolve to two distinct digital codes, or, equivalently, a digital-domain error correction is possible.

We note here that the sub-binary redundancy can be alternatively engineered using unit-element DAC by manipulating the SAR logic [1]. The binary-to-thermometer decoder required in a unit-element DAC is, however, undesirable due to its timing overhead and the extra logic needed to compute the redundant decision thresholds. A sub-binary DAC approach with hardcoded redundancy is preferred for high conversion speed [3], [8], [9]. Furthermore, the structure of a binary DAC can be retained while still providing redundancy by inserting additional decision steps into the binary search process periodically with some overhead to the SAR logic [2], [4]. Lastly, regardless of the exact redundancy form, bit-weight calibration according to Eq. (1) is dictated when random component mismatch is present, especially for a resolution of 10-bit and beyond. Section 3 will cover cases on how redundancy can be exploited to identify the bit weights.

3. DIGITAL BIT-WEIGHT CALIBRATION

3.1 Offset Double Conversion (ODC) [3]

This technique is derived from the superposition rule of linear systems. As shown in Fig. 2, a single SAR ADC digitizes each analog sample twice with two analog offsets, +Δ and –Δ, resulting in two raw codes, D_+ and D_-, respectively. Using identical bit weights, $W = \{w_i, i = 0, \cdots, N-1\}$, we first calculate the weighted sum of all bits for D_+ and D_-, denoted as d_+ and d_-, respectively. This actually realizes Eq. (1). The difference ε between d_+ and d_- is then obtained after removing 2δ (the digital version of Δ)—this difference should be ideally zero with optimal weights, guaranteed by the linearity of the ADC. A non-zero ε simply indicates incomplete learning of all bit weights and will direct the calibration to continue to adjust W until ε is driven to zero; at which point, the average of d_+ and d_- yields a proper digitization of V_{in}.

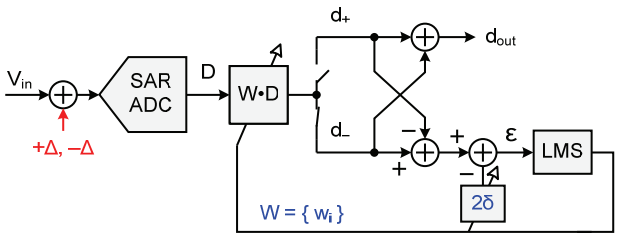
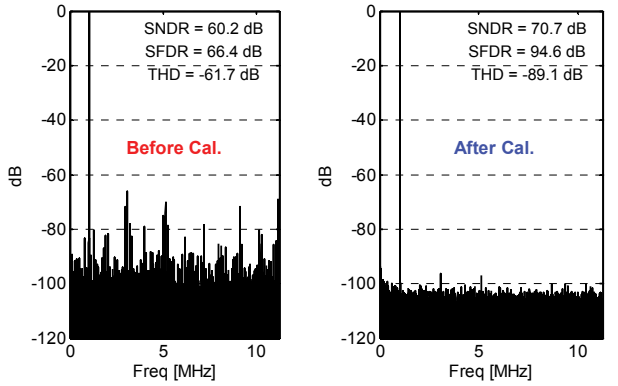


Fig. 2. ODC bit-weight calibration of SAR ADC



(a) (b)

Fig. 3. Measured output spectrum of a prototype 12-bit, 45-MS/s SAR ADC with ODC calibration: (a) before and (b) after calibration

The downside of ODC is that the conversion speed is halved in the background mode, while a benefit is that both the quantization noise and the comparator noise are attenuated by 3 dB in power due to averaging. The implementation of the offset injection is also very simple, i.e., one small capacitor and some digital logic. The convergence time of this technique compared to the correlation-based calibration [9], [10] is significantly shorter due to the deterministic and zero-forcing nature of the algorithm.

A prototype 12-bit SAR ADC was fabricated in a 1.2-V, 0.13-μm, 8M-1P CMOS process [3]. The active area of the ADC is 0.06 mm^2 and the total power consumption is 3.0 mW including calibration logic. The minimum capacitor size is set to 0.5 fF in this design. Driven by a 98% V_{FS}, 1.1-MHz sine wave at its input, the measured output power spectral density (PSD) of the prototype before and after calibration are shown in Fig. 3. With the calibration,

the SNDR, SFDR, THD were improved from 60.2, 66.4, 61.9 dB to 70.7, 94.6, 89.1 dB, respectively. The linearity improvement was nearly 30 dB. The convergence time was discovered to be inversely proportional to the magnitude of Δ—it takes about 22,000 samples to reach steady state when Δ is set to 25 LSBs.

3.2 Independent Component Analysis (ICA) [11]

Another SAR bit-weight calibration technique not subject to speed reduction is shown in Fig. 4, in which a pseudorandom bit sequence (PRBS) T of magnitude Δ is injected to the ADC input and gets digitized along with the analog input V_{in}. The digital output, obtained through a weighted sum of the individual bits of the raw digital output D, represents a digitization of $V_{in} + T{\cdot}\Delta$. If the ADC is ideal, the PRBS can be removed digitally, resulting in a digital output d_{out} (representing V_{in}) that is independent of T. When the optimal bit weights $W = \{w_i, i = 0, \cdots, N\text{-}1\}$ are unknown, the conversion process is nonlinear and the PRBS removal will be incomplete. Thus, the residual PRBS information in d_{out} can be exploited to infer the optimal bit weights.

A technical difficulty here is how to identify all N bit weights with the information of a single PRBS. Conventionally, estimating multiple model parameters dictates multiple PRBS injections, potentially degrading the ADC dynamic range and complicating the analog circuitry for injection. This is where ICA comes into picture [12]. As illustrated in Fig. 4, the technique operates on the bitwise correlation between T and the digital bits obtained through a digital re-quantizer, which mimics the SAR operation to decompose d_{out} back to its sub-binary format as D. This new digital output, termed $\hat{d}$, is correlated to T at bit level to direct the learning of all the bit weights. Since $\hat{d}_i$ and T are both one-bit signals, the digital logic implementing the correlation is simply an XOR gate.

A prototype 12-bit SAR ADC was fabricated in a 90-nm CMOS process [11]. The die area of the ADC is 0.05 mm^2. At 50 MS/s, the ADC consumes 3.3 mW from a 1.2-V supply. The PRBS injection is realized by one low-rank DAC capacitor at full sample rate. Fig. 5 presents the measured dynamic performance of the prototype. The calibration improves the SNDR and SFDR by more than 10 and 25 dB across the Nyquist band, respectively. The convergence time is around 10 million samples, or 0.2 seconds at 50 MS/s, with gear shifting applied to the LMS iterations.

3.3 Redundant Double Conversion (RDC)

The double conversion calibration illustrated in Fig. 2 can also be realized without explicit offset injection. Instead, the internal redundancy of a sub-binary SAR is exploited to facilitate the double conversion. As shown in Fig. 6, each sample is digitized twice, one using a sequence of decision thresholds corresponding to the DAC code 01···1 and the other 10···0. The effect of this, taking the MSB for example, is to create a bit-weight error detection window

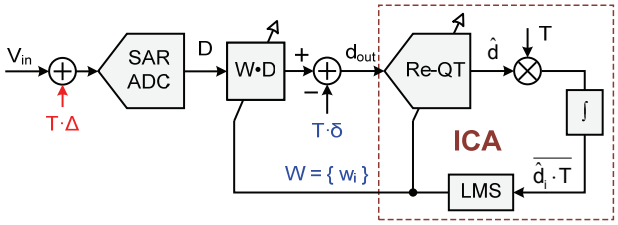


Fig. 4. ICA bit-weight calibration of SAR ADC

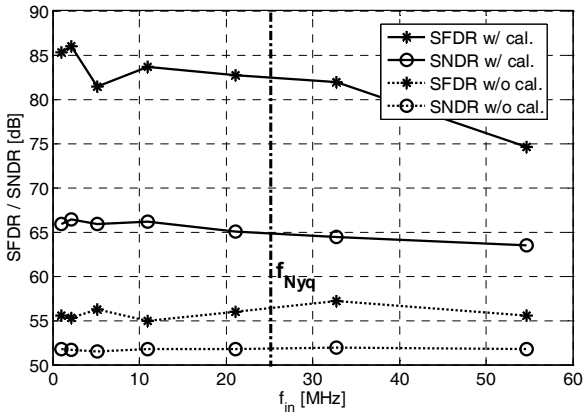


Fig. 5. Measured dynamic performance of a 90-nm, 12-bit, 50-MS/s SAR ADC prototype with ICA calibration

as large as the redundancy region shown in Fig. 1b. the two digitization outcomes d_+ and d_- are compared constantly for any difference—while a zero difference ε implies that either the bit weights are ideal or V_{in} is out of the redundancy region, a nonzero ε only means that the bit weights are not optimal and further adaptation is needed. All bit weights can be learned this way by altering the internal decision thresholds accordingly. Fig. 6 shows the two configurations in A and B, respectively.

RDC shares the same drawback of ODC, i.e., the ADC throughput is halved when operating in the background mode. However, there is no offset injection in RDC, thus there is no need to remove it digitally.

A prototype SAR ADC was fabricated in a 65-nm, 1.2-V CMOS process. The active area of the ADC is 0.05 mm^2. Fig. 5 shows the ADC output spectra before and after calibration for a full-scale 3-MHz sine-wave input. With the calibration, the SNDR and SFDR were improved from 30.0 and 31.4 dB to 71.4 and 94.1 dB, respectively. The LMS loop learns the optimal bit weights in less than 50,000 iterations, which is less than 3 μs with this ADC.

3.4 Internal Redundancy Dithering (IRD)

The 2× speed penalty associated with the RDC technique can be lifted if the shuffling of the internal configurations of A and B is controlled by a PRBS, leading to the fourth technique covered in this paper, the IRD. A system diagram of IRD is shown in Fig. 8, in which a back-end digital processor is employed to identify the MSB weight by correlating the corrected ADC output with the PRBS. Similar to the ICA case, to identify multiple bit weights,

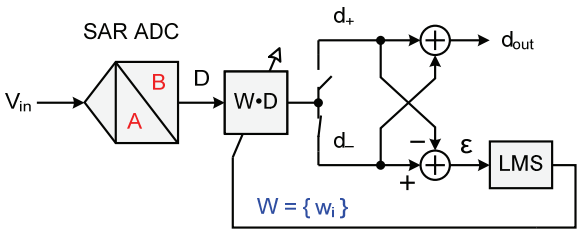
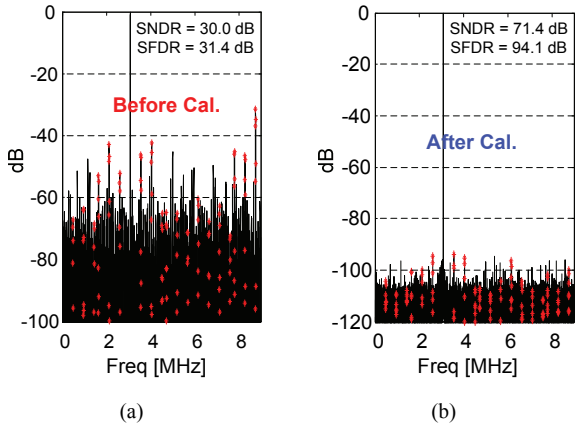


Fig. 6. RDC bit-weight calibration of SAR ADC. The labels A and B on the SAR symbol indicate that the bit-decision threshold corresponds to the DAC codes 01···1 and 10···0, respectively.



(a) (b)

Fig. 7. Measured output spectrum of a prototype 65-nm, 12-bit, 36-MS/s SAR ADC with RDC calibration: (a) before and (b) after calibration

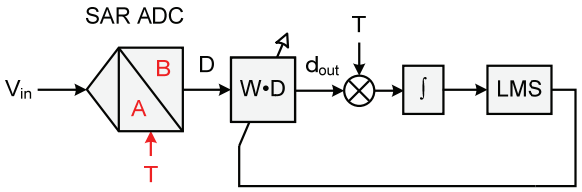


Fig. 8. IRD bit-weight calibration of SAR ADC. The labels A and B are defined the same as those of Fig. 6.

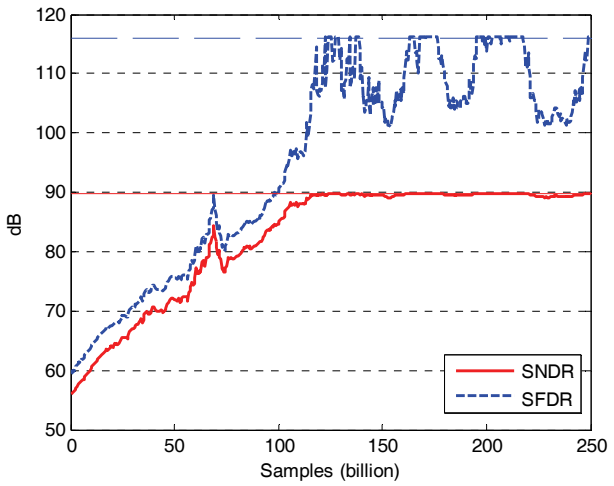


Fig. 9. Simulated learning curves of a SAR ADC employing the IRD bit-weight calibration technique

multiple correlations are required. In IRD, this implies that the PRBS responsible for decision threshold dithering will need to be different and uncorrelated for all the bits whose weights are to be identified. The convergence speed of the technique is expected to be slow due to the statistical fluctuation of the correlation process and the interaction between the learning loops of the multiple bit weights involved.

Fig. 9 illustrates the simulated SNDR and SFDR learning curves for the MSB learning case of a 15-bit SAR ADC. A significant linearity improvement of 30-40 dB is observed with calibration while the convergence time is 120 billion samples without gear shifting. The multi-bit case is currently under investigation.

4. CONCLUSION

Superior power efficiency and scalability continue to fuel the development of SAR converters in scaled technology. This paper reviews/presents a few techniques for digital background calibration of DAC bit-weight errors in SAR ADCs. With the dominant performance roadblock eliminated by calibration, SAR will become suitable and potentially dominate for broadband (100-200 MHz) and high-resolution (12-14 bits) applications such as wireless base-station and video streaming in advanced CMOS nodes.

REFERENCES

[1] F. Kuttner, A 1.2V 10b 20MS/s non-binary SAR ADC in 0.13µm CMOS, in *ISSCC*, *Dig. Tech. Papers*, Feb. 2002, pp. 176-177.

[2] C. C. Liu et al., A 10b 100MS/s 1.13mW SAR ADC with binary-scaled error compensation, in *ISSCC*, *Dig. Tech. Papers*, Feb. 2010, pp. 386-387.

[3] W. Liu et al., A 12b 22.5/45MS/s 3.0mW 0.059mm^2 CMOS SAR ADC achieving over 90dB SFDR, in *ISSCC*, *Dig. Tech. Papers*, Feb. 2010, pp. 380-381.

[4] C. Hurrell et al., An 18b 12.5MS/s ADC with 93dB SNR, *IEEE J. of Solid-State Circuits*, pp. 2647-2654, Dec. 2010.

[5] H. Wei et al., A 0.024mm^2 8b 400MS/s SAR ADC with 2b/cycle and resistive DAC in 65nm CMOS, in *ISSCC*, *Dig. Tech. Papers*, Feb. 2011, pp.188-190.

[6] Y. Zhu et al., A 34fJ 10b 500 MS/s partial-interleaving pipelined SAR ADC, in *VLSI Circuits*, *Dig. Tech. Papers*, Jun. 2012, pp. 90-91.

[7] Z. G. Boyacigiller et al., An error-correcting 14b/20µs CMOS A/D converter, in *ISSCC*, *Dig. Tech. Papers*, Feb. 1981, pp. 62-63.

[8] D. Draxelmayr, A self-calibration technique for redundant A/D converters providing 16b accuracy, in *ISSCC*, *Dig. Tech. Papers*, Feb. 1988, pp. 204-205.

[9] E. Siragusa and I. Galton, A digitally enhanced 1.8-V 15-bit 40-MSample/s CMOS pipelined ADC, *IEEE J. of Solid-State Circuits*, pp. 2126-2138, Dec. 2004.

[10] Y.-S. Shu and B.-S. Song, A 15b linear, 20MS/s, 1.5b/stage pipelined ADC digitally calibrated with signal-dependent dithering, in *VLSI Circuits*, *Dig. Tech. Papers*, Jun. 2006, pp. 218-219.

[11] W. Liu et al., A 12-bit 50-MS/s 3.3-mW SAR ADC with background digital calibration, in *Proc. CICC*, Sept. 2012, pp. 1-4.

[12] Y. Chiu et al., An ICA framework for digital background calibration of analog-to-digital converters, in press for *Sampling Theory in Signal and Image Processing (STSIP)*.

Trend of High-Speed SAR ADC towards RF Sampling

Mike Shuo-Wei Chen
Department of Electrical Engineering
University of Southern California, Los Angeles, USA
Email: swchen@usc.edu

ABSTRACT

One emerging trend of high-speed low-power ADC design is to leverage the successive approximation (SAR) topology. It has successfully advanced the power efficiency by orders of magnitude over the past decade. Given the nature of SAR algorithm, the conversion speed is intrinsically slow compared to other high-speed ADC architectures, and yet minimal static power is required due to the mostly digital implementation. This paper examines various speed enhancement techniques that enable SAR ADCs towards RF sampling, i.e. >GS/s sampling rate with >GHz input bandwidth, while maintaining low power and area consumption. It is expected to play a crucial role in the future energy-constrained wideband system.

I. INTRODUCTION

High-speed medium-resolution ADCs are widely adopted by electronic systems, such as instrumentations, disk read channel, high-speed serial links, optical communications, and wideband radios, etc. The ADCs in this category were initially dominated by Flash architecture with bipolar devices [1-5] in the 80s and early 90s due to the higher device speed. The first dramatic shift of paradigm began when the pervasive penetration of CMOS technology started in the late 90s. Despite that the device speed was not as high as bipolar devices, its low cost, wide adoption by digital and progressively improved speed have finally made high-speed CMOS ADC into reality [6-10]. From the architecture perspective, the Flash topology is preferred for high-speed operation since all the comparisons are accomplished within one clock cycle; however, the complexity increases exponentially with ADC resolution. It thus triggered other architectural possibilities in this realm, such as pipelined ADC architecture with time-interleaving [7, 11-15]. In the recent years, there is another major architectural shift towards high-speed SAR operation. Since the SAR architecture does not require linear analog amplification, it benefits more from the technology scaling. Much research has been engaged to push this power efficient architecture into high speed sampling regime, while it was conventionally limited to lower speed range, i.e. KS/s to MS/s, as illustrated in Fig. 1.

To prove the outstanding power efficiency of SAR architecture in relation to other ADC topologies, the performance of recent state-of-the-art high-speed ADCs is plotted in Fig. 2. The circled dots are the published ADC literatures based on SAR within the past decade, which indicate that the power efficiency has improved by orders of magnitude, particularly for the medium resolution high-speed operation up to tens GS/s sampling rate.

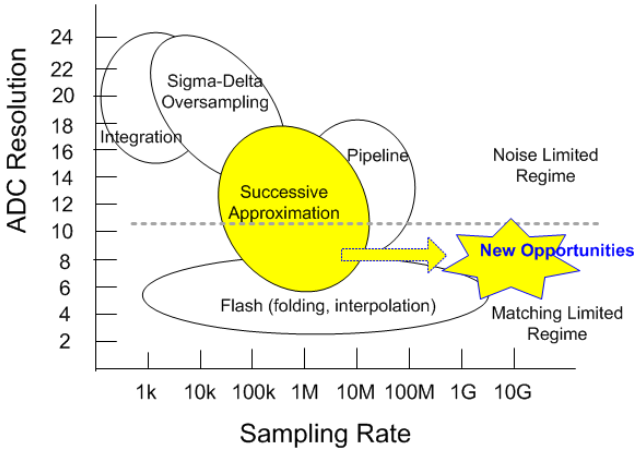


Figure 1 Emerging paradigm shift by leveraging SAR architecture

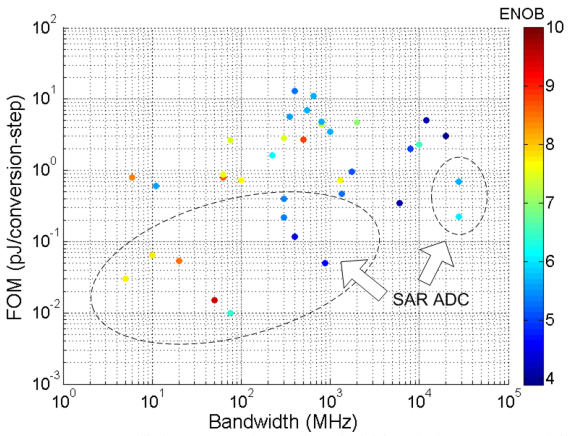


Figure 2 Power efficiency vs. input bandwidth of the recent published Nyquist ADCs

This paper overviews various critical techniques to enable this level of high-speed and low-power operation. An asynchronous SAR architecture will be described in section II, which effectively reduces the internal comparison time and complexity. Section III outlines a multi-bit per conversion cycle technique that reduces the required number of SAR comparison cycles. Pipelined (section IV) and time interleaved (section V) SAR further increases the sampling rate through pipelining multiple conversion stages or parallelizing an array of SAR ADCs. The paper will be concluded in section VI.

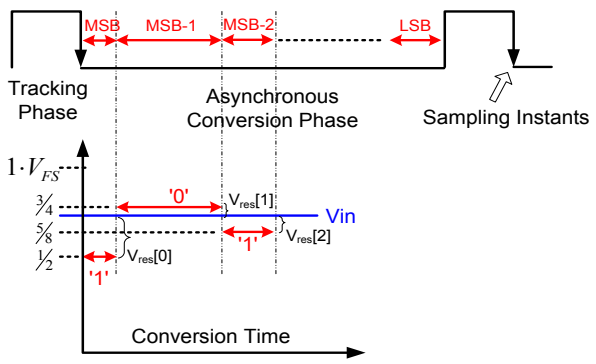


Figure 3 Concept of asynchronous SAR conversion

II. ASYNCHRONOUS SAR

The concept of asynchronous SAR architecture was first introduced in [16]. It aims to eliminate the conversion speed constraint of a conventional synchronous SAR ADC, which relies on an internal clock to divide the time into signal tracking and individual bit comparison phase from MSB to LSB. Since every clock cycle must tolerate the worst-case comparison time as well as the clock jitter, the overall conversion speed is constrained by design. The key idea of asynchronous SAR is to trigger the internal comparison from MSB to LSB like dominoes. Whenever the current comparison is complete, a ready signal is generated and triggers the following comparison immediately. The reduction in the overall comparison time is thus achieved due to the time savings in those faster conversion cycles, as shown in Fig. 3. Moreover, no high speed internal clock is needed, which leads to a low complexity implementation. Note that, a global clock running at the sampling rate is still required to perform uniform sampling.

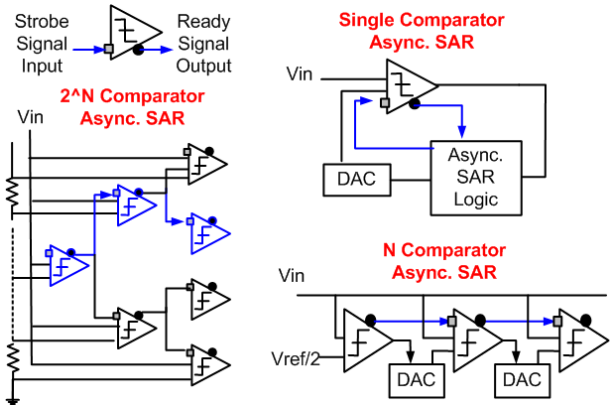


Figure 4 Potential implementations of asynchronous SAR

In terms of implementation, there are several variations to carry out the same asynchronous SAR algorithm including single, N and 2^N comparator configurations, as shown in Fig. 4. The single comparator configuration [16-18] consumes the least power and area among the three. However, if higher conversion speed is desired, the N or 2^N comparator configuration can be utilized to eliminate or reduce the time required for comparator reset and DAC settling [19, 20]. Note that, besides the additional hardware complexity, the offset voltage between the various comparators will degrade the ADC performance and hence extra calibrations are typically applied in the multi-comparator configurations.

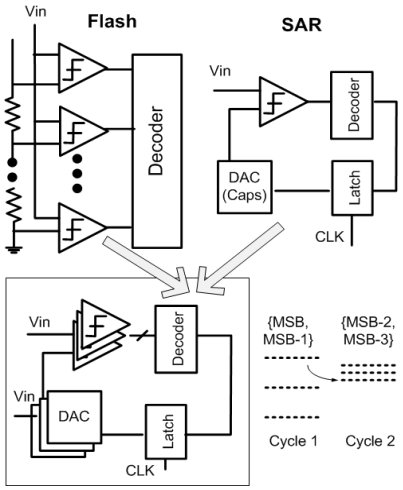


Figure 5 Multi-bit per cycle SAR architecture

III. MULTI-BIT PER CYCLE

Conventional SAR algorithm utilizes one comparison per cycle and hence requires at least N comparison cycles for an N-bit resolution. If more comparisons can be accomplished within one comparison cycle, the conversion time will be reduced proportionally, i.e. halved for 2bit/cycle case. Essentially, it is the combination of Flash and SAR ADC topology that are compromised in between the hardware complexity and sampling speed. The idea can be traced at least back to 60s' [21-23], where multiple reference DACs are built so that two bits are generated per cycle. In recent years, multiple capacitive DACs are utilized to sample the analog input and perform 2bit/cycle SAR algorithm by generating various reference levels [24]. To mitigate the drawback of additional capacitive loading, interpolation technique can be adopted with the mixture of resistive and capacitive DACs [25, 26]. The sampling speed of a single 2b/cycle SAR ADC has been demonstrated close-to 1GS/s with 6-8 bit resolution. Note that, the consequence of adopting such a Flash-like architecture is the vulnerability to the comparator offset, which leads to ADC nonlinearity. On the contrary, the comparator offset of a conventional 1b/cycle SAR only leads to the global offset without distortion. Therefore, the multi-bit per cycle SAR architecture is not as power efficient and most likely requires offset cancellation techniques. Another variation of a multi-bit per cycle SAR ADC is to utilize both voltage and time quantization that effectively provides multi-bit comparisons [27, 28]. It makes use of the input dependent delay of the comparator resolving time and allows SAR conversion to reduce switching activity and required conversion time.

IV. PIPELINING

Another common technique to improve the sampling speed is through pipelined conversion stages. Conventional pipelined ADC utilizes low resolution Flash ADC in each pipelined stage. The concept of pipelined SAR architecture is to replace the complex Flash ADC with power efficient SAR topology. As a result, one can allocate more quantization levels for each pipelined stage without much power penalty. Moreover, in the case of charge redistribution SAR ADC, the residue voltage is readily available on the capacitor network in the end of SAR

conversion, which can be re-used as part of the switch-capacitor residue amplifier [29, 30]. The architecture can also be extended to other low-power residue amplification techniques, such as an amplifier that dynamically charges up the second stage sampling capacitance depending on the residue voltage from the previous stage [19]. The drawback is the less accurate amplification and vulnerability to PVT variations which requires extra calibrations.

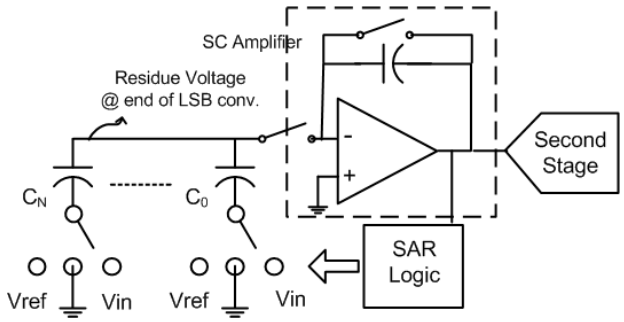


Figure 6 One embodiment of pipelined SAR

V. TIME INTERLEAVING

In the early 2000s, SAR ADC began its footprint in the high speed sampling regime (>hundreds MS/s) instead of operating in high resolution and lower sampling rate. In [31], it demonstrated that a 6bit, 600MS/s ADC is achievable via 8-way time interleaved SAR in 90nm CMOS with low power consumption. Ever since, the number of time interleaved SAR has been increasing consistently and a recent 8-bit, 56GS/s ADC was reported in [32] that consists of unprecedented 320-way 175MS/s SAR ADCs in 65nm CMOS. The ultra-high-speed ADC design has become somewhat similar to digital VLSI design, where massive parallelism is adopted for speed improvement. However, there are significant overheads associated massive time interleaved ADCs, including the capacitive loading of the sample-and-hold network, clock distributions, and mismatches in between the single ADCs. In this sub-section, several design techniques to alleviate these constraints will be reviewed.

First of all, the value of sampling capacitor should be minimized while maintaining sufficient matching accuracy. As more ADCs are time interleaved, the more sampling capacitors will load the previous driver stage and limit the achievable bandwidth. For example, if the tracking time is half of the entire sampling period, the driver of an M-way time interleaved ADC will be loaded with M/2 sampling capacitor at any given time. One way to alleviate the sampling capacitor loading issue is to divide the sampling switches into two stages. The first-stage front end sampling switches operate at a higher speed but with less capacitor loading [33]. However, the buffers in between the stages can become the linearity bottleneck. Another common approach is to reduce the tracking time so that only one sampling

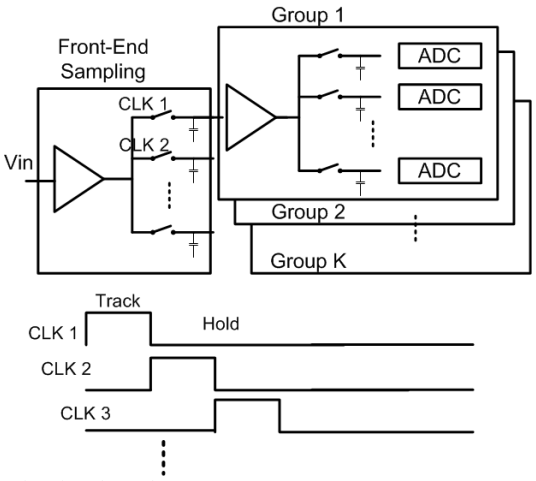


Figure 7 Time interleaved SAR ADC

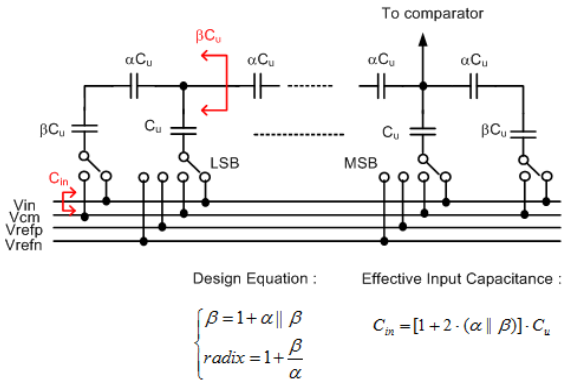


Figure 8 Series capacitor ladder network for S/H

capacitor is activated at a time [34, 35], as shown in Fig. 7. Besides cascading the sampling network, the sampling capacitance of each ADC should be minimized. For a medium resolution ADC, the sampling capacitance is not constrained by the KT/C noise, for instance, an 8-bit ADC requires merely on the order of 10fF total sampling capacitance with 1V input swing. Shown in Fig. 8, a series capacitor ladder network can be applied in both non-binary [16] and binary case [36]. Since the capacitors are connected in series, the total sampling capacitance seen by the input driver is substantially reduced and independent of ADC resolution, which is not the case in the conventional parallel connected capacitor array. Another benefit of using series connected capacitor network is the potential usage of a larger unit capacitor in order to satisfy the matching requirement. Finally, the mismatches between the interleaved ADCs typically require calibrations to compensate for offset, gain, and timing skews [13, 37-39].

VI. CONCLUSION

SAR ADC architecture presents a promising path for high speed and low power operation. Moreover, the nature of its mostly digital implementation will continue to favor the technology scaling in terms of the achievable speed and power consumption. Several outlined techniques, including asynchronous SAR, massive time interleaving, and pipelining, are expected to play a key role in driving even higher sampling rate and lower power consumption in the future. More architecture and circuit level innovations to further enhance SAR conversion speed and reduce the overhead of massive time interleaving are crucial to achieve this goal.

References

[1] A. Matsuzawa, M. Kagawa, M. Kanoh, K. Tatehara, T. Yamaoka and K. Shimizu, "A 10 b 30 MHz two-step parallel BiCMOS ADC with internal S/H," in Solid-State Circuits Conference, 1990. Digest of Technical Papers. 37th ISSCC., 1990 IEEE International, pp. 162-163, 1990.

[2] F. Murden and R. Gosser, "12b 50MSample/s two-stage A/D converter," in Solid-State Circuits Conference, 1995. Digest of Technical Papers. 42nd ISSCC, 1995 IEEE International, pp. 278-279, 379, 1995.

[3] B. Peetz, B.D. Hamilton and J. Kang, "An 8-bit 250 megasample per second analog-to-digital converter: operation without a sample and hold," Solid-State Circuits, IEEE Journal Of, vol. 21, pp. 997-1002, 1986.

[4] R. Petschacher, B. Zojer, B. Astegher, H. Jessner and A. Lechner, "A 10-b 75-MSPS subranging A/D converter with integrated sample and hold," Solid-State Circuits, IEEE Journal Of, vol. 25, pp. 1339-1346, 1990.

[5] K. Poulton, K.L. Knudsen, J.J. Corcoran, Keh-Chung Wang, R.B. Nubling, R.L. Pierson, M.-.F. Chang, P.M. Asbeck and R.T. Huang, "A 6-b, 4 GSa/s GaAs HBT ADC," Solid-State Circuits, IEEE Journal Of, vol. 30, pp. 1109-1118, 1995.

[6] B.P. Brandt and J. Lutsky, "A 75-mW, 10-b, 20-MSPS CMOS subranging ADC with 9.5 effective bits," Solid-State Circuits, IEEE Journal Of, vol. 34, pp. 1788-1795, 1999.

[7] K. Poulton, R. Neff, B. Setterberg, B. Wuppermann, T. Kopley, R. Jewett, J. Pernillo, C. Tan and A. Montijo, "A 20 GS/s 8 b ADC with a 1 MB memory in 0.18 μm CMOS," in Solid-State Circuits Conference, 2003. Digest of Technical Papers. ISSCC. 2003 IEEE International, pp. 318-496 vol.1, 2003.

[8] Shin-Il Lim, Seung-Hoon Lee and Sun-Young Hwang, "A 12 b 10 MHz 250 mW CMOS A/D converter," in Solid-State Circuits Conference, 1996. Digest of Technical Papers. 42nd ISSCC., 1996 IEEE International, pp. 316-317, 465, 1996.

[9] S. Tsukamoto, T. Endo and W.G. Schofield, "A CMOS 6b 400 M sample/s ADC with error correction," in Solid-State Circuits Conference, 1998. Digest of Technical Papers. 1998 IEEE International, pp. 152-153, 1998.

[10] Y. Tamba and K. Yamakido, "A CMOS 6 b 500 MSample/s ADC for a hard disk drive read channel," in Solid-State Circuits Conference, 1999. Digest of Technical Papers. ISSCC. 1999 IEEE International, pp. 324-325, 1999.

[11] W. Bright, "8 b 75 M sample/s 70 mW parallel pipelined ADC incorporating double sampling," in Solid-State Circuits Conference, 1998. Digest of Technical Papers. 1998 IEEE International, pp. 146-147, 428, 1998.

[12] C.S.G. Conroy, D.W. Cline and P.R. Gray, "A high-speed parallel pipelined ADC technique in CMOS," in VLSI Circuits, 1992. Digest of Technical Papers., 1992 Symposium on, pp. 96-97, 1992.

[13] Daihong Fu, K.C. Dyer, S.H. Lewis and P.J. Hurst, "A digital background calibration technique for time-interleaved analog-to-digital converters," Solid-State Circuits, IEEE Journal Of, vol. 33, pp. 1904-1911, 1998.

[14] K.Y. Kim, N. Kusayanagi and A.A. Abidi, "A 10-bit, 100 MS/s CMOS A/D converter," in Custom Integrated Circuits Conference, 1996., Proceedings of the IEEE 1996, pp. 419-422, 1996.

[15] Yun-Ti Wang and B. Razavi, "An 8-bit 150-MHz CMOS A/D converter," Solid-State Circuits, IEEE Journal Of, vol. 35, pp. 308-317, 2000.

[16] S.-.M. Chen and R.W. Brodersen, "A 6-bit 600-MS/s 5.3-mW Asynchronous ADC in 0.13- CMOS," Solid-State Circuits, IEEE Journal Of, vol. 41, pp. 2669-2680, 2006.

[17] J. Craninckx and G. Van der Plas, "A 65fJ/Conversion-Step 0-to-50MS/s 0-to-0.7mW 9b Charge-Sharing SAR ADC in 90nm Digital CMOS," in Solid-State Circuits Conference, 2007. ISSCC 2007. Digest of Technical Papers. IEEE International, pp. 246-600, 2007.

[18] P. Harpe, Cui Zhou, Xiaoyan Wang, G. Dolmans and H. de Groot, "A 30fJ/conversion-step 8b 0-to-10MS/s asynchronous SAR ADC in 90nm CMOS," in Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2010 IEEE International, pp. 388-389, 2010.

[19] B. Verbruggen, J. Craninckx, M. Kuijk, P. Wambacq and G. Van der Plas, "A 2.6 mW 6 bit 2.2 GS/s Fully Dynamic Pipeline ADC in 40 nm Digital CMOS," Solid-State Circuits, IEEE Journal Of, vol. 45, pp. 2080-2090, 2010.

[20] G. Van der Plas and B. Verbruggen, "A 150 MS/s 133 W 7 bit ADC in 90 nm Digital CMOS," Solid-State Circuits, IEEE Journal Of, vol. 43, pp. 2631-2640, 2008.

[21] A.M. Dighe and A.R. Kelkar, "New strategies for fast ADC circuits," Instrumentation and Measurement, IEEE Transactions On, vol. 39, pp. 878-880, 1990.

[22] S.M. Bhandari and S. Aggarwal, "A successive double-bit approximation technique for analog/digital conversion," Circuits and Systems, IEEE Transactions On, vol. 37, pp. 856-858, 1990.

[23] T.C. Verster, "A Method to Increase the Accuracy of Fast-Serial-Parallel Analog-to-Digital Converters," Electronic Computers, IEEE Transactions On, vol. EC-13, pp. 471-473, 1964.

[24] Zhiheng Cao, Shouli Yan and Yunchu Li, "A 32mW 1.25GS/s 6b 2b/step SAR ADC in 0.13µm CMOS," in Solid-State Circuits Conference, 2008. ISSCC 2008. Digest of Technical Papers. IEEE International, pp. 542-634, 2008.

[25] Hegong Wei, Chi-Hang Chan, U-Fat Chio, Sai-Weng Sin, Seng-Pan U, R.P. Martins and F. Maloberti, "An 8-b 400-MS/s 2-b-Per-Cycle SAR ADC With Resistive DAC," Solid-State Circuits, IEEE Journal Of, vol. 47, pp. 2763-2772, 2012.

[26] Yuan-Ching Lien, "A 4.5-mW 8-b 750-MS/s 2-b/step asynchronous subranged SAR ADC in 28-nm CMOS technology," in VLSI Circuits (VLSIC), 2012 Symposium on, pp. 88-89, 2012.

[27] J. Guerber, H. Venkatram, M. Gande, A. Waters and U. Moon, "A 10-b Ternary SAR ADC With Quantization Time Information Utilization," Solid-State Circuits, IEEE Journal Of, vol. 47, pp. 2604-2613, 2012.

[28] A. Shikata, R. Sekimoto, T. Kuroda and H. Ishikuro, "A 0.5 V 1.1 MS/sec 6.3 fJ/Conversion-Step SAR-ADC With Tri-Level Comparator in 40 nm CMOS," Solid-State Circuits, IEEE Journal Of, vol. 47, pp. 1022-1030, 2012.

[29] M. Furuta, M. Nozawa and T. Itakura, "A 10-bit, 40-MS/s, 1.21 mW Pipelined SAR ADC Using Single-Ended 1.5-bit/cycle Conversion Technique," Solid-State Circuits, IEEE Journal Of, vol. PP, pp. 1-1, 2011.

[30] C.C. Lee and M.P. Flynn, "A SAR-Assisted Two-Stage Pipeline ADC," Solid-State Circuits, IEEE Journal Of, vol. 46, pp. 859-869, 2011.

[31] D. Draxelmayr, "A 6b 600MHz 10mW ADC array in digital 90nm CMOS," in Solid-State Circuits Conference, 2004. Digest of Technical Papers. ISSCC. 2004 IEEE International, pp. 264-527 Vol.1, 2004.

[32] I. Dedic, "56Gs/s ADC : Enabling 100GbE," in Optical Fiber Communication (OFC), collocated National Fiber Optic Engineers Conference, 2010 Conference on (OFC/NFOEC), pp. 1-3, 2010.

[33] K. Doris, E. Janssen, C. Nani, A. Zanikopoulos and G. van der Weide, "A 480mW 2.6GS/s 10b 65nm CMOS time-interleaved ADC with 48.5dB SNDR up to Nyquist," in Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2011 IEEE International, pp. 180-182, 2011.

[34] S.K. Gupta, M.A. Inerfield and Jingbo Wang, "A 1-GS/s 11-bit ADC With 55-dB SNDR, 250-mW Power Realized by a High Bandwidth Scalable Time-Interleaved Architecture," Solid-State Circuits, IEEE Journal Of, vol. 41, pp. 2650-2657, 2006.

[35] S.M. Louwsma, A.J.M. van Tuijl, M. Vertregt and B. Nauta, "A 1.35 GS/s, 10 b, 175 mW Time-Interleaved AD Converter in 0.13 µm CMOS," Solid-State Circuits, IEEE Journal Of, vol. 43, pp. 778-786, 2008.

[36] E. Alpman, H. Lakdawala, L.R. Carley and K. Soumyanath, "A 1.1V 50mW 2.5GS/s 7b Time-Interleaved C-2C SAR ADC in 45nm LP digital CMOS," in Solid-State Circuits Conference - Digest of Technical Papers, 2009. ISSCC 2009. IEEE International, pp. 76-77,77a, 2009.

[37] WenBo Liu, Yuchun Chang, Szu-Kang Hsien, Bo-Wei Chen, Yung-Pin Lee, Wen-Tsao Chen, Tzu-Yi Yang, Gin-Kou Ma and Yun Chiu, "A 600MS/s 30mW 0.13µm CMOS ADC array achieving over 60dB SFDR with adaptive digital equalization," in Solid-State Circuits Conference - Digest of Technical Papers, 2009. ISSCC 2009. IEEE International, pp. 82-83,83a, 2009.

[38] J.A. McNeill, K.Y. Chan, M.C.W. Coln, C.L. David and C. Brenneman, "All-Digital Background Calibration of a Successive Approximation ADC Using the "Split ADC" Architecture," Circuits and Systems I: Regular Papers, IEEE Transactions On, vol. PP, pp. 1-1, 2011.

[39] M. El-Chammas and B. Murmann, "A 12-GS/s 81-mW 5-bit Time-Interleaved Flash ADC With Background Timing Skew Calibration," Solid-State Circuits, IEEE Journal Of, vol. 46, pp. 838-847, 2011.

Multi-Step Switching Methods for SAR ADCs

Ying-Zu Lin†, Ya-Ting Shyu, Che-Hsun Kuo, Guan-Ying Huang, Chun-Cheng Liu and Soon-Jyh Chang††
National Cheng Kung University, Tainan, Taiwan
Email: †tibrius@gmail.com and ††soon@mail.ncku.edu.tw

***Abstract* - This paper presents multi-step capacitor switching methods for SAR ADCs based on precharge with floating capacitors and charge sharing. The proposed switching methods further reduce the transient power of the split monotonic switching method (an improved version of the monotonic switching method). Compared to the split monotonic switching, adding charge sharing achieves around 50% reduction in switching power. Using precharge with floating capacitors and charge sharing simultaneously, the switching power reduces around 75%. The proposed switching methods do not require additional intermediate reference voltages.**

I. Introduction

Deeply scaled CMOS technologies give analog-to-digital converter (ADC) designers low supply voltage and insufficient intrinsic gain. Among all kinds of ADCs, the successive-approximation-register (SAR) ADC seems to gain the most advantages in CMOS downscaling. A SAR ADC usually consists of sampling switches, a comparator, capacitor arrays and SAR logic. The SAR ADCs obtain digital representation of input signal by switching instead of amplifying in amplifier-based ADCs like the pipelined ADC. Improved metal implementation enhances metal-oxide-metal (MOM) capacitor matching. Digital SAR logic reaches higher speed and energy efficiency as CMOS technology continues to scale down. Recent publications show SAR ADCs achieve excellent power efficiency [1][2][3].

The accuracy of the SAR ADC mainly relies on capacitive digital-to-analog converter (DAC) design. In SAR ADCs, the binary-weighted DAC capture input signal on one side of capacitor arrays while C-2C and split DACs use both sides to process input signal. Thus, a binary-weighted DAC has better intrinsic linearity than the other two due to its better immunity against parasitic effects. Nonetheless, a SAR ADC using a binary-weighted DAC suffers from large input capacitance, resulting in large capacitor switching power which grows exponentially with ADC resolution.

This switching power of the DAC in SAR ADCs has been well analyzed in [4][5]. This paper proposes switching methods based on the split monotonic switching method [6]. The proposed methods further reduce the switching power. Section II introduces the monotonic switching method and its variant. Section III describes the switching techniques and the proposed methods. Section IV shows the analysis and behavioral simulation result. Section VI draws the conclusion.

II. The Monotonic Switching Method and Its Variant

Conventional SAR ADCs samples input signals onto bottom plates of the capacitor arrays. The switching power of capacitor arrays with bottom-plate signal sampling has been well analyzed in [4]. A SAR ADC using the monotonic switching method makes the top plates of the DACs connected to the comparator input and bottom plates to reference voltages [7]. The monotonic switching method samples input signals onto top plates of the capacitor arrays, as shown in Fig. 1. The advantages of this switching method include half unit capacitor count and one less switching compared to the conventional case. The main disadvantage of this method is the changing input common-mode voltage during bit cycling which affects the accuracy of the comparator. Thus, a variant of the monotonic switching method referred to as split monotonic switching method is invented [6]. Fig. 2 shows one of the capacitor arrays of the split monotonic switching method. A capacitor is split into two parts. For example, C_{1px} and C_{1py} in Fig. 2 are split by C_{1p} in Fig. 1. In the reset state, one is switched to the positive reference voltage V_{refp} and the other to the negative one V_{refn}. Once a compotator decision has made, the monotonic switching method only switches one capacitor in a capacitor array and the other capacitor array remains unchanged. For the split monotonic switching method, a capacitor array switches a sub-capacitor and the other array switches another sub-capacitor. Fig. 3 shows the waveforms at the top plates of the monotonic and split monotonic switching methods. The figure shows the split monotonic switching

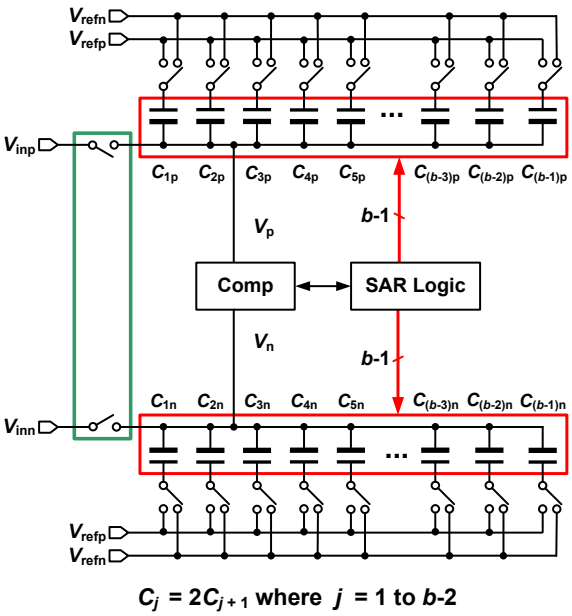


Fig. 1. A SAR ADC using the monotonic switching method.

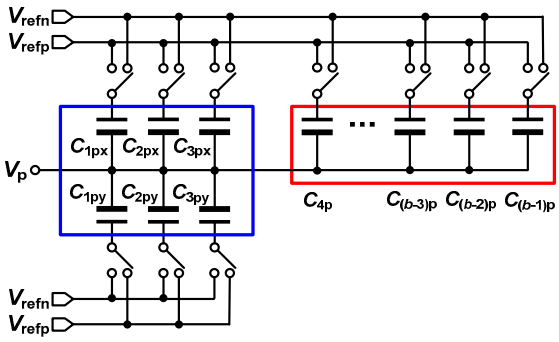


Fig. 2. The capacitor array of the split monotonic switching method.

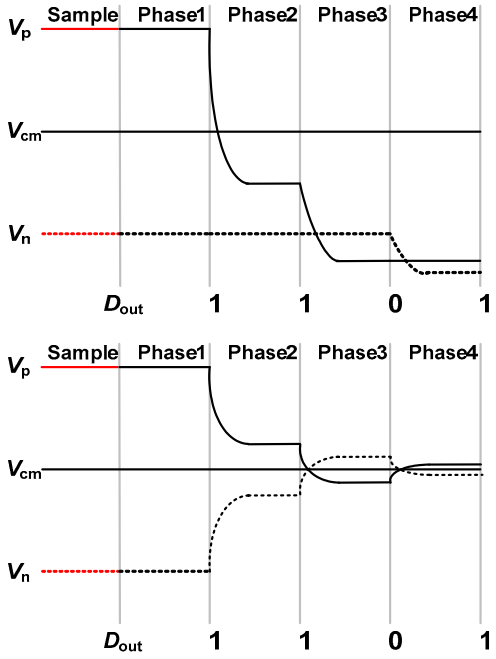


Fig. 3. The waveforms at top plates of the monotonic switching method (top) and the split monotonic switching method (bottom).

method has a constant common-mode voltage. Although doubled switches are required, the split monotonic method improves the accuracy of the SAR ADC. Generally, the split monotonic switching method only applies to the first several bits of a SAR ADC, and the rest bits perform monotonic switching. The compromised arrangement save hardware and enhance accuracy simultaneously.

The switching power of the capacitor array is proportional to the unit capacitance and the number of unit capacitors connected to the reference voltage. Intuitively, a high resolution capacitor array consumes more switching power than a lower one. For SAR ADCs with 10-bit or larger capacitor array, further switching power reduction is necessary to enhance power efficiency.

III. Proposed Multi-Step Switching Methods

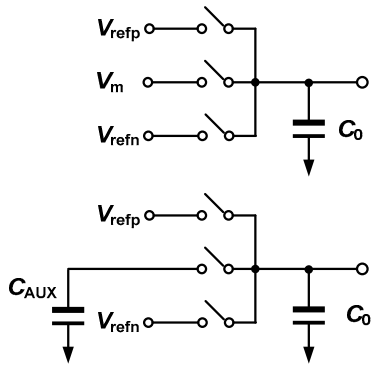


Fig. 4. Multi-step charging w/ external voltage (top) and that w/ a floating capacitor (bottom).

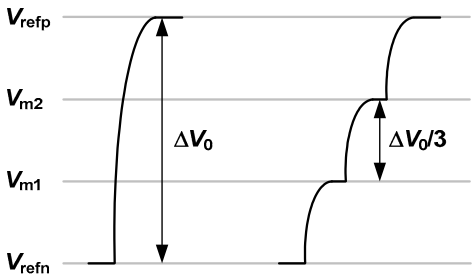


Fig. 5. Waveform of the one-step charging (left) and waveform of the multi-step charging (right).

This section shows two techniques reducing the energy consumed during capacitor transient activities. They are precharge with floating capacitors and charge sharing.

A. *Precharging with Floating Capacitors*

The power dissipation of a reference source arises from the charges to make a capacitor reaching the desired voltage level. In the split monotonic switching method, for each bit, a sub-capacitor is switched from V_{refp} to V_{refn}, and the other sub-capacitor is switched in the reversed direction. Charges from high voltage potential directly flowing to a lower voltage potential is energy inefficient. If the charges at the higher voltage potential help the charging of the capacitor at the lower potential, the charge recycling reduces the energy dissipation of the reference source. Generally, the charging of a capacitor is 'one-step.' However, a 'multi-step' charging is much more energy efficient. The multi-step charging idea is firstly mentioned in [8], and then applied to the drivers for LCD panels [9]. A SAR ADC employs this technique achieving excellent power efficiency [10]. This technique separates a 'one-step' charging into multi steps and multi phases, as shown in Fig. 5. If the voltage difference before and after charging is V_0 and the capacitance is C_0, the total energy consumed is $C_0V_0^2$ [4]. If the voltage difference is equally separated into n parts, the total energy is reduced to $C_0V_0^2/n$ where n is a natural number [8]. If the intermediate voltages are applied by power ICs, the efficiency loss during power conversion reduces the effectiveness of multi-step charging.

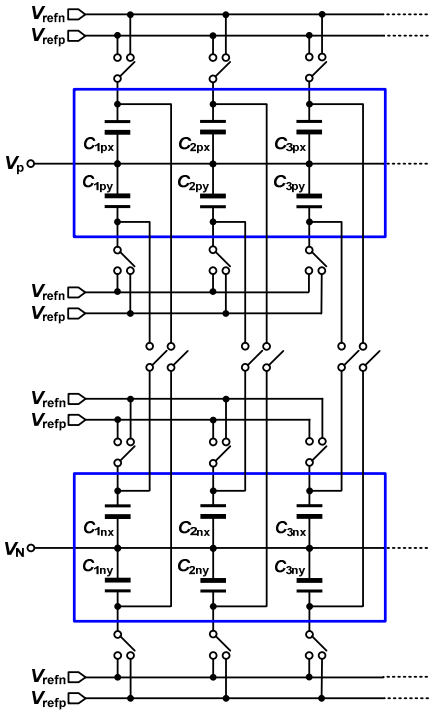


Fig. 6. The capacitor arrays of the split monotonic switching method combining charge sharing method.

Another approach is to use auxiliary capacitors to replace intermediate voltage sources. Fig. 4 depicts the cases using an external voltage (top) and a floating capacitor (bottom). In the bottom case, before switching to the highest voltage, the capacitor switches to the top plate of the floating capacitor. By repeating charging and discharging, the top plate of the auxiliary capacitor forms a stable intermediate voltage [9]. The auxiliary capacitor should be larger than the loading capacitor. A large ratio keeps the intermediate voltage stable.

In the multi-step charging procedures, the charges of intermediate steps are provided by the floating capacitors. Thus, the reference voltage only deals with the last charging. The more charging stages result in better energy efficiency. However, too many charging/discharging phases slow down the operation speed. Additional logic and driving circuits are also necessary to perform the multi-step charging/discharging.

B. *Charge Sharing*

Charge sharing is relatively intuitive. For example, a capacitor will discharge to the low voltage potential and the other identical capacitor will charge to the high potential. If we connect the top plates of the two capacitors together, the top plates will reach a middle voltage potential. Without dissipating charges from the reference voltage, the DAC array obtains free charges. Charge sharing needs less hardware than the precharge with floating capacitors. In split monotonic method, when a sub-capacitor switches upward, a sub-capacitor in the other array switches downward. The condition is perfectly suitable for charge sharing.

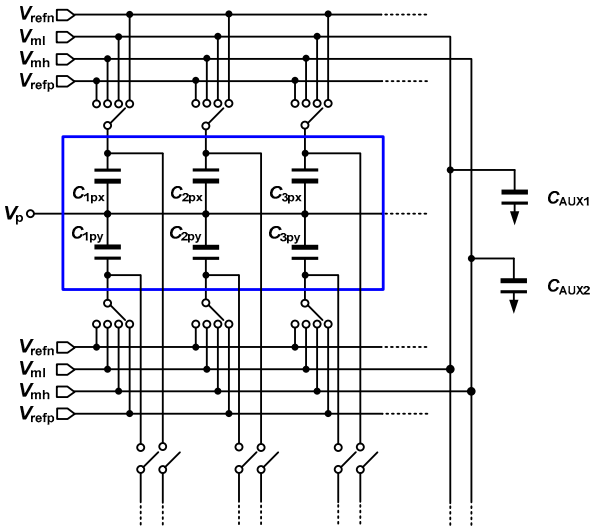


Fig. 7. The capacitor arrays of the split monotonic switching method combining charge sharing and precharge methods.

C. *Proposed Switching Methods*

Fig. 6 depicts the capacitor arrays of the split monotonic switching method combining the charge sharing technique. Note there is a switch placed between two sub-capacitors of the two arrays. The switches are used to perform charge sharing of two sub-capacitors. Fig. 7 shows the capacitor arrays of the split monotonic switching method combining charge sharing plus precharge with floating capacitors. Fig. 7 has three inter-stages. Floating capacitor performs the first and third charging and discharging; the charge sharing deals with the second one. Fig. 8 shows the waveforms of the bottom plates of the capacitor array using one charge sharing (left) and the waveforms using one charge sharing and two precharge (right). Note the inter-stage charging and discharging do not affect the final values.

Since the charge sharing and precharge with floating capacitors complicate logic design, the two techniques do not have to apply to the whole array. For a binary DAC array, the first 2- to 4-bit switching using the two techniques will save most of the switching power. For small capacitors, the two techniques are inefficient. The combined method is more suitable for low-speed high-resolution SAR ADCs. The method in Fig. 6 (only one charge sharing) is suitable for high-speed SAR ADCs.

V. Analysis and Behavioral Modeling

This section analyzes the switching power of the aforementioned switching methods. The switching power for each code of the monotonic switching can be expressed as

$$E_{\text{mono}}(n) = V_{\text{ref}} \times \left(V_{\text{x1}}(n) - V_{\text{x2}}(n)\right) \times C_{\text{ref}}(n) \quad (1)$$

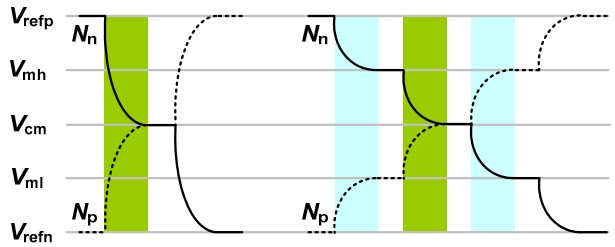


Fig. 8. The waveforms of the bottom plates of the capacitor array using one charge sharing (left) and the waveforms using one charge sharing and two precharge (right).

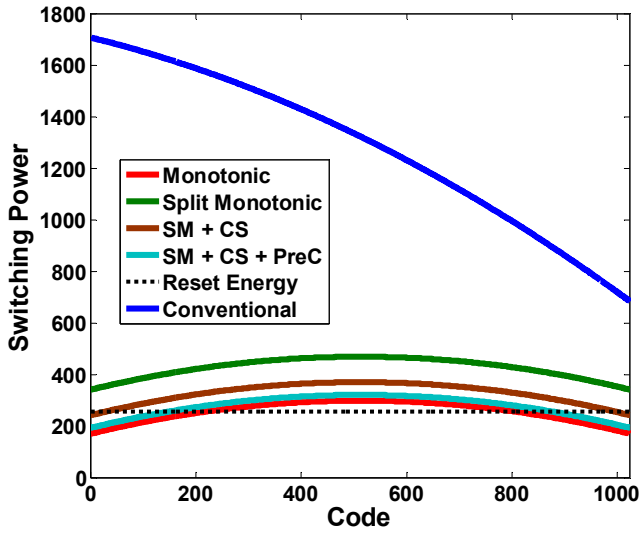


Fig. 9. Behavioral simulation result.

where $V_{x1}(n)$ and $V_{x2}(n)$ are the top-plate voltages before and after conversion for code n, respectively. C_{ref} is the total capacitance connected to V_{ref} after conversion. The switching power of the split monotonic switching can be expressed as

$$E_{mono\_split}(n) = V_{ref} \times \left(V_{x1}(n) - V_{x2}(n)\right) \times C_{ref}(n) + V_{ref}^2 \times C_{up}(n) \quad (2)$$

where C_{up} is the upward switching capacitance. The switching power of the split monotonic switching plus charge sharing and precharge with floating capacitors can be expressed as

$$E_{ms+cs+prec}(n) = V_{ref} \times \left(V_{x1}(n) - V_{x2}(n)\right) \times C_{ref}(n) + V_{ref} \times \left(V_{ref} - V_t \times \left[(C_{total} - C_{up}) / C_{total}\right]\right) \quad (3)$$

where $V_t = V_{ref}/2$ for charge sharing case and $V_t = 3V_{ref}/4$ for charge sharing plus precharge case. (1) – (3) only show the switching power during conversion. (2) and (3) must add energy consumption in sampling (reset) phase. The value is $V_{ref} \cdot C_{ref,smp}$ where $C_{ref,smp}$ means the total capacitance connected to V_{ref} during the sampling phase. Note the monotonic switching does not consume energy during reset.

We use behavioral modeling to demonstrate the switching power reduction of the proposed switching methods. Fig. 9 shows the behavioral simulation result of five cases: 1) conventional switching; 2) monotonic switching; 3) split monotonic switching; 4) split monotonic switching with charge sharing; and 5) split monotonic switching with charge sharing and precharge. During conversion, the split monotonic consumes less reference energy than the monotonic switching. However, the monotonic switching method is the smallest if the energy consumption during sampling (reset) phase is added. Note the proposed methods consumed the smallest energy from reference buffer during conversion. Thus the reference buffer design becomes easier. The charge sharing and charge sharing plus precharge help the split monotonic switching to achieve significant switching power reduction.

VI. Conclusion

This paper proposes multi-step charging and discharging methods for the SAR ADCs based on the split monotonic switching. Although the split monotonic consumes more total energy, it dissipates less energy during conversion. Shift the settling issue of the DAC from conversion to reset is beneficial for high-speed operation. The reset phase is generally much longer than a bit conversion phase. The proposed methods do not require additional voltages. For a 10-bit SAR ADC, the total capacitance using the split monotonic switching is smaller than 2.5pF. On-chip auxiliary capacitors (>25 pF) are possible. The multi-step switching method not only saves switching power but also hardware cost.

References

[1] V. Giannini, P. Nuzzo, V. Chironi, A. Baschirotto, G. Van der Plas, and J. Craninckx "An 820μW 9b 40MS/s noise-tolerant dynamic-SAR ADC in 90nm digital CMOS," *IEEE ISSCC Dig. Tech. Papers*, Feb. 2008, pp. 238-239.

[2] C.-C. Liu, S.-J. Chang, G.-Y. Huang, Y.-Z. Lin, C.-M. Huang and C.-H. Huang, "A 10b 100MS/s 1.13mW SAR ADC with binary scaled error compensation," *IEEE ISSCC Dig. Tech. Papers*, pp. 386-387, Feb. 2010.

[3] M. Yoshioka, K. Ishikawa, T. Takayama, S. Tsukamoto, "A 10b 50MS/s 820μW SAR ADC with on-chip digital calibration" *IEEE ISSCC Dig. Tech. Papers*, Feb. 2010, pp. 384-385.

[4] B. P. Ginsburg, A. P. Chandrakasan, "An energy-efficient charge recycling approach for a SAR converter with capacitive DAC," *IEEE Int. Symp. Circuits Syst.*, pp. 184-187, May 2005.

[5] J. Guerber, H. Venkatram, T. Oh, and U.-K. Moon, "Enhanced SAR ADC energy efficiency from the early reset merged capacitor switching algorithm," *IEEE Int. Symp. Circuits Syst.*, pp. 2361-2364, May 2012.

[6] C.-C. Liu, S.-J. Chang, G.-Y. Huang, Y.-Z. Lin and C.-M. Huang, "A 1V 11fJ/conversion-Step 10bit 10MS/s asynchronous SAR ADC in 0.18um CMOS," *IEEE Symposium on VLSI Circuits*, pp.241-242, 2010.

[7] C.-C. Liu, S.-J. Chang, G.-Y. Huang, and Y.-Z. Lin, "A 10-bit 50-MS/s SAR ADC with a monotonic capacitor switching procedure," *IEEE J. Solid-State Circuits*, vol. 34, no. 5, pp. 731-740, Apr. 2010.

[8] L. J. Svensson and J. G. Koller, "Driving a capacitive load without dissipating fCV^2," in *IEEE Symp. Low Power Electronics*, 1994.

[9] J.-S. Kim, D.-K. Jeong, and G. Kim, "A multi-level multi-phase charge-recycling method for low-power AMLCD column drivers," *IEEE J. Solid-State Circuits*, vol. 35, no. 1, pp. 74-84, Jan. 2000.

[10] M. van Elzakker, E. van Tuijl, P. Geraedts, D. Schinkel, E. A. M. Klumperink, an B. Nauta, "A 10-bit charge-redistribution ADC consuming 1.9 μW at 1 MS/s," *IEEE J. Solid-State Circuits*, vol. 45, no. 5, pp. 1007-1015, May. 2010.

ON THE USE OF REDUNDANCY IN SUCCESSIVE APPROXIMATION A/D CONVERTERS

Boris Murmann

Department of Electrical Engineering, Stanford University, Stanford, CA, USA
Email: murmann@stanford.edu

ABSTRACT

In practical realizations of sequential (or pipelined) A/D converters, some form of redundancy is typically employed to help absorb imperfections in the underlying circuits. The purpose of this paper is to review the various ways in which redundancy has been used in successive approximating register (SAR) ADCs, and to connect findings from the information theory community to ideas that drive modern hardware realizations.

Keywords— A/D conversion, redundancy, successive approximation, beta expansion

1. INTRODUCTION

Analog-to-digital (A/D) converters map continuous-time, continuous-amplitude signals into a discretized representation via sampling and quantization. In a typical hardware implementation, the precision of this mapping is impaired by nonidealities of the underlying electronic circuit, as for instance mismatch between nominally identical components. In practice, these nonidealities can be mitigated via a number of design techniques that can be categorized into the following groups: (1) precision analog design, (2) analog or digital calibration techniques, and (3) redundancy.

Precision analog design techniques aim at designing (or sizing) the circuit such that its precision matches the desired specifications by construction. While this approach can be practical, it sometimes causes significant overhead, for example in terms of power dissipation. To address this issue, calibrated A/D converters correct circuit imperfections by measuring the induced errors and by adjusting a correction circuit in the analog or digital domain. Introducing redundancy in the A/D conversion process is another popular solution, but it differs fundamentally from calibration in the sense that the errors are neither measured, nor corrected, but simply tolerated and rejected by the conversion algorithm. Many modern A/D converters utilize a combination of calibration and redundancy and employing redundancy is often required to make certain calibration techniques work.

To this author's best knowledge, the use of redundancy in A/D converters dates back to 1964 [1]. Since then, many variants of the idea have been proposed and used in practice. Most recently, however, there has been renewed interest in research on this topic for the successive approximation register (SAR) architecture, which has gained popularity due to its compatibility with nano-scale integrated circuit technologies [2]. As we will explain below, SAR ADCs can benefit from redundancy in a variety of intriguing ways, some of which have been discovered or applied only recently. Within this context, the purpose of this paper is to summarize the state-of-the-art in the design of SAR ADCs with redundancy.

2. IDEAL A/D CONVERSION AND BETA-EXPANSION

Ideal A/D conversion of a continuous input variable $0 \le x < 1$ can be viewed as a binary expansion of the form

$$\hat{x} = \sum_{k=1}^{N} b_k 2^{-k} \tag{1}$$

Here, $b_1, \ldots, b_N \in \{0, 1\}$ are the bits of the binary representation and $\hat{x} - x$ is the quantization error. The bits can be determined using a binary search algorithm that uses the initial guess $x_1 = 1/2$ and the recursion

$$x_k = x_{k-1} + s_k 2^{-k} \tag{2}$$

where

$$s_k = \begin{cases} +1 & x > x_k \\ -1 & x \le x_k \end{cases} \tag{3}$$

and $b_k = (s_k + 1)/2$. This process can be interpreted graphically using the decision tree shown in Figure 1 [3]. The dotted lines represent all possible paths for x_k, and the solid lines correspond to an example path for a specific input x. An important property of this conversion algorithm is that the path that leads to $\hat{x}$ is unique. This also implies that there exists a unique bit pattern for each input, and more importantly, any error in the bit decisions given by (3) will prevent us from achieving the best possible approximation.

Consider now a modification of (1) such that

$$\hat{x} = \alpha \sum_{k=1}^{N} b_k \beta^{-k} \tag{4}$$

where $1 < \beta < 2$ and $\alpha = \beta - 1$ is a scale factor that sets the full-scale range to unity. As explained in [4], this "beta-expansion" [5] contains redundancy, in the sense that multiple bit patterns can lead to an approximation within a certain error bound. This is illustrated graphically in Figure 2. Here, $\beta = 2^{3/4}$ and the algorithm uses $N = 4$ steps. After the last step, the obtained approximation is digitally mapped onto the closest level of an ideal 3-bit A/D converter.

As we can see from the pattern of all possible paths, there are multiple trajectories that terminate at the same $\hat{x}$. This means that certain decision errors can be absorbed without affecting the conversion result. For instance, as shown using the bold dashed line, a decision error in the

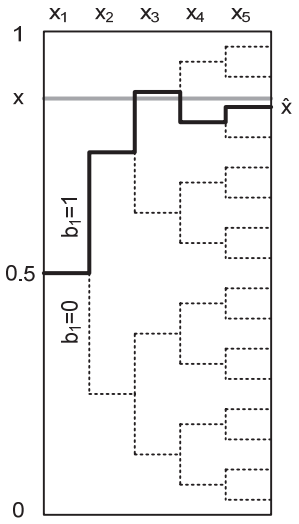


Figure 1: Graphical illustration of ideal sequential A/D conversion with 4-bit resolution. The algorithm resolves 4 bits using 4 steps (no redundancy).

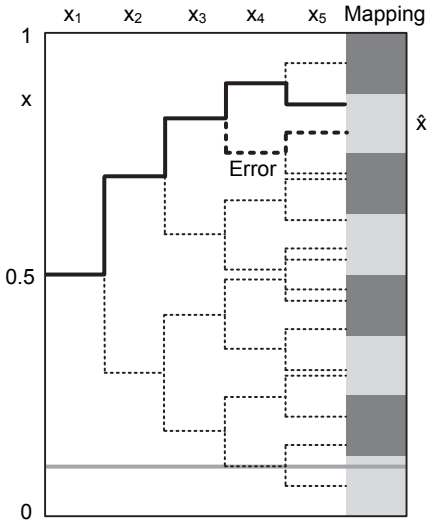


Figure 2: Graphical illustration of sequential A/D conversion with redundancy. The conversion resolves 3 bits using 4 steps.

third step will still lead to the correct conversion result. The cost for this error tolerance is two-fold: (1) the number of steps must be larger than the number of bits that are being resolved and (2) extra hardware is needed to map the raw bit pattern into the usual binary output.

The magnitude of the tolerable decision error in each step can be estimated by computing the difference between the current bit weight and the sum of all remaining weights. For example, consider the above-described converter with $\beta = 3/4$, resolving three bits in 4 steps. The first bit enters (1) with a weight of $\beta^{-1} = 0.595$. The sum of the remaining weights is $\beta^{-2} + \beta^{-3} + \beta^{-4} = 0.689$. As long as a close approximation (within the quantization error) is reachable by the sum of the last three weights, an error in the first bit decision will be inconsequential. The same idea applies to later bit decisions, with the main difference that the sum of the remaining weights, and therefore the correction range, is decreasing with each step. Detailed calculations of the tolerable decision errors for a variety of bit configurations are tabulated in [3]. For example, in a 10-bit, 12-step ADC, the tolerable decision errors normalized to the quantization step size are: 90, 51, 28, 16, 9, 5, 3, 1, 1 and 0 for all remaining decisions.

In recent literature, it is often overlooked that the concept of using $\beta < 2$ (or "radix < 2") has been used in hardware implementations long before detailed mathematical results – such as Daubechies' 2002 paper [4] – were available. In the context of SAR ADCs, using a reduced radix was first proposed in 1981 [6], and further popularized in [7], [8]. The latter reference is sometimes cited as the "first" even though it appeared more than twenty years after the original idea. What is even less known is that the original idea of using redundancy dates back to 1964 [1]. In this work, redundancy was introduced not by using $\beta < 2$, but instead by creating extra decision levels in (3). We will summarize this idea and other approaches that have evolved in the context of hardware design in the following section.

3. REDUNDANCY IN TODAY'S DESIGNS

A. Radix=2 Designs with Redundant Decision Levels

In the original work of [1], one extra decision level was used to create overlapping trajectories as in Figure 2. The design resolved two bits per step, which normally requires three decision levels. The added fourth decision level allowed the algorithm to absorb large comparison errors, enumerated in more detail in [1].

The idea of introducing redundant decision levels is still used today, and most widely exploited in pipeline ADCs [9], which can be described by a set of equations similar to (1) – (4). In this context, designers speak of a "1.5-bit" quantizer when one extra level is added to (3), since $\log_2(2+1) = 1.58$. The concept is also called "redundant signed digit (RSD)" conversion [10], akin to the redundant binary number system sometimes used in digital adders. The 1.5-bit concept has been re-introduced recently in SAR conversion, as described in [11].

B. Radix=2 Designs with Redundant Steps

Redundancy primarily helps absorb errors in the bit decisions (equation (3)). However, it is important to distinguish between two different ways in which such errors may be introduced. The first and most obvious is a direct error in the evaluation of the inequality. The second possibility is an error in x_k, which may occur in hardware realizations due to the finite speed at which (2) is computed ("DAC settling error" – see also Section IV). Such errors can be tolerated by designs with redundant decision levels, but it was shown in [12] that the introduction of one redundant step (and no extra decision levels) is also sufficient. The idea exploits the exponential nature of the settling errors and the fact that the impact of the errors reduces from cycle to cycle.

The work of [13] uses one redundant step in an even more intriguing way to mitigate the impact of random decision errors ("thermal noise"). It is noted that at most two out of all decisions (equation (2)) must resolve a very small difference that may be corrupted by noise. One of these critical decisions must be the last one, and the other one can be in any prior cycle. As shown in [13], this latter error can be elegantly corrected by introducing one extra conversion step. In hardware, this feature is then exploited by running all but the last two conversions with a very low-energy (but noisy) comparator, and expending significant energy to overcome noise only in the final decisions.

C. Radix<2 Designs

As discussed previously, the idea of using a radix of less than two ("beta-expansion") goes back to 1981 and is still used today [3]. One common challenge to this form of redundancy is that the radix must be known precisely to construct the proper conversion result. Of course, the radix must also be precisely set in radix = 2 topologies, but here this is naturally achieved by employing integer multiples of well-matched and nominally identical integrated circuit components.

In practice, the radix is typically measured using some form of calibration. In [14], it was shown that the radix (β) can be estimated by comparing the output of the converter for the inputs x and $1\text{-}x$. In a practical realization, such a calibration step would have to be performed with controlled input signals, thus interrupting the normal conversion operation. Such an approach is commonly called foreground or start-up calibration.

Reference [15] describes a method by which the radix can be continuously measured ("in the background") without interrupting normal conversion. The method is based on running two conversions of the same input with different additive perturbations. Based on the difference between the two results and its ideal value, an LMS loop updates the radix in the digital bit mapping until convergence is achieved. At first glance it seems expensive to run extra conversions for the sole purpose of measuring of the radix. However, the two measurements allow averaging of the thermal noise and hence the calibration is energy neutral (to first order).

With redundancy and radix calibration in place, the only remaining precision requirement in the hardware is that the computation of (2) must be sufficiently linear. However, as pointed out in [16], even nonlinearity could be compensated through calibration. Still, in typical realizations of SAR ADCs, where the computation of (2) relies on high-quality passive components, such issues have not yet proven to be significant. The situation is different in pipeline ADCs, where digital linearization techniques have been proposed to combat nonlinear effects in passives [17] and amplifiers [18].

5. A CLOSER LOOK AT DAC SETTLING ERRORS

Figure 3 shows a conceptional block diagram of a typical SAR ADC. The comparison level x_k in (2) is generated by a D/A converter, which is controlled by digital circuitry that implements the approximation

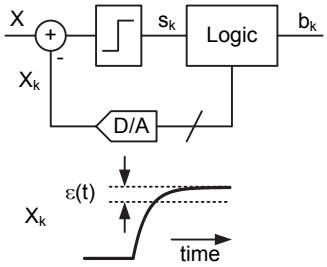


Figure 3: Conceptual block diagram of a SAR ADC.

algorithm. Since the speed of practical D/A converters is finite, x_k is usually not fully settled at the time the bit decision is made and this can lead to bit errors. Fortunately, and as already mentioned above, such errors are inconsequential with sufficient redundancy in place. The DAC error is indistinguishable from errors made in the quantizer itself. Especially for high-speed designs, this feature is being heavily exploited in today's designs [12].

In this context, it is interesting to invoke a comparison to pipeline ADCs, which also employ redundancy in their underlying quantizers. Figure 4(a) shows a block diagram of a pipeline ADC, which can be conceptually thought of as a "loop-unrolled" version of a SAR ADC. In other words, instead of performing (2) sequentially, the hardware is parallelized and pipelined to increase throughput. An interesting and important difference between the shown pipelined architecture and a SAR ADC is that DAC settling errors *cannot* be absorbed through redundancy. The reason is that the settling error is sampled and forward-propagated such that it results in a direct error that has no further time to decay. A clever workaround for this problem was only proposed recently in [19]. As shown in Figure 4(b), this design uses a feedforward path, which, after some delay injects a precise version of the fully settled DAC signal into the following stage. The feedforward path has extra time to settle, since its output is only needed after the succeeding stage's quantizer and DAC have processed their inputs. With this modification, the pipelined architecture can potentially benefit as much from redundancy as a SAR ADC, and high conversion rates are possible with relatively slow sub-D/A converters and amplifiers.

6. CONCLUSION

This paper has reviewed the state-of-the-art and historical background on the use of redundancy in SAR A/D converters. A general observation for most of the work in this area is that the practical exploration of ideas typically occurs well before the underlying mathematics has been thoroughly described. The development of a holistic theoretical framework that captures all variants of redundancy would be beneficial to the field.

REFERENCES

[1] T. C. Verster, "A Method to Increase the Accuracy of Fast-Serial-Parallel Analog-to-Digital Converters," *IEEE Trans. on Electronic Computers*, vol. EC-13, no. 4, pp. 471–473, Aug. 1964.

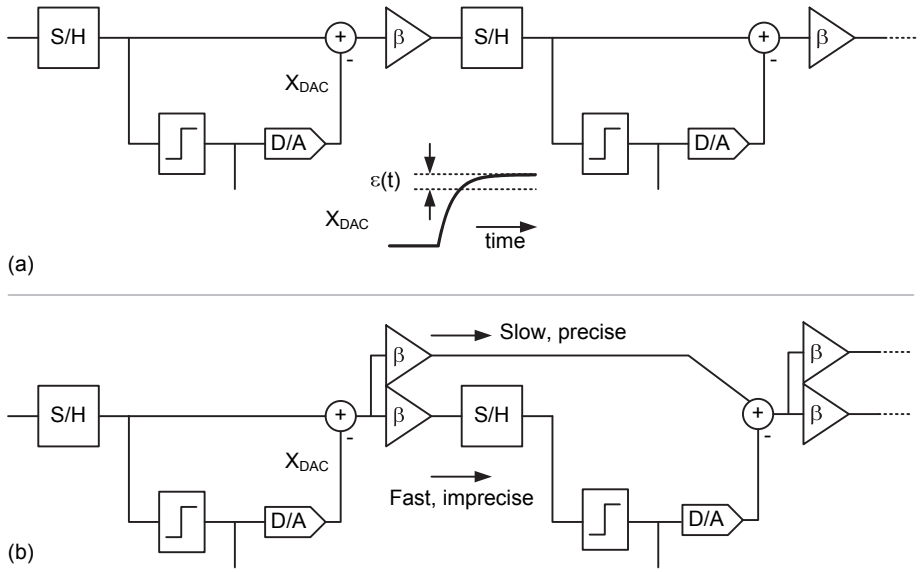


Figure 4: (a) Conventional pipeline ADC. (b) Addition of a feedforward path that allows the absorption of DAC settling errors in the converter's redundancy.

[2] D. Draxelmayr, "A 6b 600MHz 10mW ADC array in digital 90nm CMOS," in *ISSCC Dig. Techn. Papers*, 2006, pp. 264–265.

[3] T. Ogawa, H. Kobayashi, Y. Takahashi, N. Takai, M. Hotta, H. San, T. Matsuura, A. Abe, K. Yagi, and T. Mori, "SAR ADC Algorithm with Redundancy and Digital Error Correction," *IEICE Trans. on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E93.A, no. 2, pp. 415–423, Feb. 2010.

[4] I. Daubechies, R. DeVore, C. S. Gunturk, and V. Vaishampayan, "Beta expansions: a new approach to digitally corrected A/D conversion," in *Proc. IEEE ISCAS*, 2002, pp. 784–787.

[5] W. Parry, "On the Beta-Expansions of Real Numbers," *Acta Math. Acad. Sci. Hungar.*, vol. 11, pp. 401–416, 1960.

[6] Z. Boyacigiller, B. Weir, and P. Bradshaw, "An error-correcting 14b/20μs CMOS A/D converter," in *ISSCC Dig. Techn. Papers*, 1981, pp. 62–63.

[7] D. Draxelmayr, "A Self Calibration Technique for Redundant A/D Converters Providing 16b Accuracy," in *ISSCC Dig. Techn. Papers*, 1988, pp. 204–205.

[8] F. Kuttner, "A 1.2V 10b 20MSample/s non-binary successive approximation ADC in 0.13μm CMOS," in *ISSCC Dig. Techn. Papers*, 2002, pp. 136–137.

[9] S. H. Lewis, H. S. Fetterman, G. F. Gross, R. Ramachandran, and T. R. Viswanathan, "A 10-b 20-Msample/s analog-to-digital converter," *IEEE Journal of Solid-State Circuits*, vol. 27, no. 3, pp. 351–358, Mar. 1992.

[10] B. Ginetti, A. Vandemeulebroecke, and P. Jespers, "RSD cyclic analog-to-digital converter," in *Symp. VLSI Circuits Dig.*, 1988, pp. 125–126.

[11] R. Vitek, E. Gordon, S. Maerkovich, and A. Beidas, "A 0.015mm2 63fJ/conversion-step 10-bit 220MS/s SAR ADC with 1.5b/step redundancy and digital metastability correction," in *Proc. IEEE Custom Integrated Circuits Conference*, 2012, pp. 1–4.

[12] C.-C. Liu, S.-J. Chang, G.-Y. Huang, Y.-Z. Lin, C.-M. Huang, C.-H. Huang, L. Bu, and C.-C. Tsai, "A 10b 100MS/s 1.13mW SAR ADC with binary-scaled error compensation," in *ISSCC Dig. Techn. Papers*, 2010, pp. 386–387.

[13] V. Giannini, P. Nuzzo, V. Chironi, A. Baschirotto, G. Van der Plas, and J. Craninckx, "An 820μW 9b 40MS/s Noise-Tolerant Dynamic-SAR ADC in 90nm Digital CMOS," in *ISSCC Dig. Techn. Papers*, 2008, pp. 238–239.

[14] I. Daubechies and O. Yilmaz, "Robust and Practical Analog-to-Digital Conversion With Exponential Precision," *IEEE Trans. Information Theory*, vol. 52, no. 8, pp. 3533–3545, 2006.

[15] W. Liu, P. Huang, and Y. Chiu, "A 12b 22.5/45MS/s 3.0mW 0.059mm2 CMOS SAR ADC achieving over 90dB SFDR," in *ISSCC Dig. Techn. Papers*, 2010, pp. 380–381.

[16] J. Biveroni and H.-A. Loeliger, "On sequential analog-to-digital conversion with low-precision components," in *2008 Information Theory and Applications Workshop*, 2008.

[17] M. K. Mayes and S. W. Chin, "A 200 mW, 1 Msample/s, 16-b pipelined A/D converter with on-chip 32-b microcontroller," *IEEE J. Solid-State Circuits*, vol. 31, no. 12, pp. 1862–1872, Dec. 1996.

[18] B. Murmann and B. E. Boser, "A 12-bit 75-MS/s pipelined ADC using open-loop residue amplification," *IEEE Journal of Solid-State Circuits*, vol. 38, no. 12, pp. 2040–2050, Dec. 2003.

[19] Y. Chai and J.-T. Wu, "A 5.37mW 10b 200MS/s dual-path pipelined ADC," in *ISSCC Dig. Techn. Papers*, 2012, pp. 462–464.

Design Considerations of Ultra-Low-Voltage Self-Calibrated SAR ADC

Hai Huang [*], Xiaoyang Wang [*], and Qiang Li [*†]
[*] University of Electronic Science and Technology of China, Chengdu
[†] Department of Engineering, Aarhus University, Denmark

***Abstract*—This paper discusses the design of 0.5V 12bit successive approximation register (SAR) analog-to-digital converter (ADC) with focus on the considerations of self calibration at low supply voltage. Relationships among noises of comparators and overall ADC performance are studied. Moreover, an ultra–low-leakage switch is demonstrated in a 0.13µm CMOS process and an improved process of measuring mismatch is proposed to alleviate the charge injection of sampling switch. Simulation shows the ADC achieves an ENOB of 11.4b and a SFDR of 90dB near Nyquist rate with capacitor mismatch up to 3%. At 12b 1MS/s, the ADC exhibits an FOM of 13.2fJ/step under 0.5V supply voltage.**

I. Introduction

ENERGY- constrained applications such as mobile devices, wearable medical equipments, wireless sensor networks, etc., require power-efficient analog-to-digital converters (ADC) for long life span. Meanwhile, low voltage ADC was demanded by the continuous down-scaling of digital supply voltage for SoC integration. In these applications, successive approximation register (SAR) ADC is normally a dominant architecture due to its low power and mostly-digital characteristics [1] [2].

The achievable resolution of SAR ADC at normal voltage is mainly limited by capacitor matching. Benefiting from the down-scaling of CMOS technology, however, SAR ADC can incorporate additional calibration logic naturally. Several calibration techniques have been reported [3]-[6], whereas there is few designs discussing the design considerations of high resolution sub-1V ADCs with the calibration. Improved results of the calibrations in [3], [4] and [5] were reported with supply voltages above 1V. The ADC in [6] works at 0.5V supply voltage with an ENOB around 10b after the calibration.

Ultra-low-supply voltage introduces some additional serious challenges to the self-calibrated SAR ADC. The noises of the comparators play more dominant roles in limiting the ADC's performance at low voltage compared with that at normal voltage. And the improved performance due to the calibration is also degenerated significantly as the measurement of capacitor's mismatch is vulnerable to the leakage and the charge injection of sampling switches and the offset of the comparator.

This work was supported in part by the National Natural Science Foundation of China (61006027) and the New Century Excellent Talents (NCET) program of the Ministry of Education of China (NCET-10-0297).

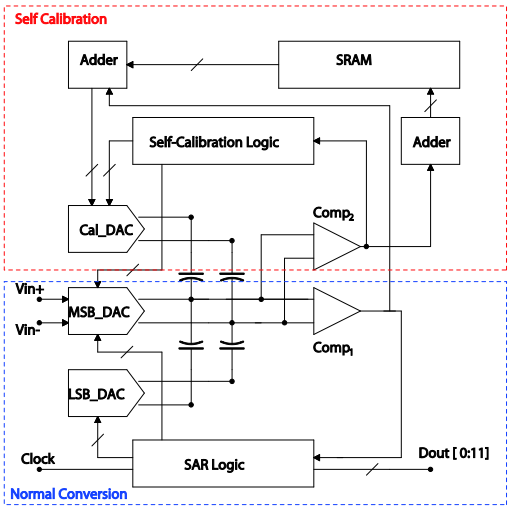


Fig. 1. The architecture of self-calibrated ADC.

This paper elaborates the design considerations of SAR ADC with self calibration at low voltage. Relationships among noises of comparators and overall ADC performance are studied. Meanwhile, we discuss challenges to circuit designs of ultra-low-voltage ADC with self calibration and present several solutions. Simulated in 0.13µm CMOS process, the implemented 0.5V ADC achieves 11.4b ENOB and 13.2fJ/step FOM with capacitors' mismatches up to 3%.

II. Self Calibration of SAR ADC

Fig. 1 shows the block diagram of the self-calibrated SAR ADC. The ADC consists of two comparators, a main DAC (splitting to MSB DAC and LSB DAC), a calibration DAC, SAR logic, self-calibration logic, adders and SRAM. The modules in the bottom of Fig. 1 are responsible for normal SAR conversion while those in the top are in charge of the self calibration.

Once the ADC is powered on, the measurement of the mismatch is started under the control of self-calibration logic. The measurement begins from the MSB capacitor in the MSB DAC and ends at the LSB capacitor. To begin with, all capacitors except the capacitor ready for the measurement in the main DAC samples reference voltage V_{ref} and then redistribute the charge with the capacitor ready for the measurement, resulting in the residual voltage related to the mismatch [3]. Then the calibration DAC digitizes the residual voltage in a SAR conversion making use of the precise comparator $Comp_2$.

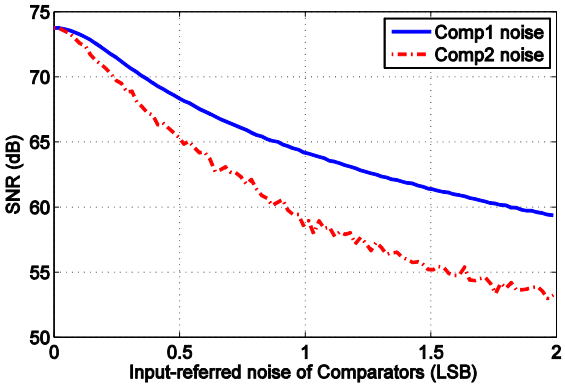


Fig. 2. SNR of 12bit self-calibrated ADC as functions of the comparators' noises.

The digital code of the residual voltage is processed by adders and stored in the on-chip SRAM, thus the measurement of one capacitor's mismatch is accomplished. This procedure repeats subsequently until the LSB capacitor in MSB DAC is completed.

After the measurement of the mismatch, the normal conversion begins and the mismatch-measurement block is powered off. During the normal conversion, the calibration DAC adjusts its connection according to the output of $Comp_1$ and the accumulation result of the data read from the SRAM. Effectively, the error voltage caused by capacitors' mismatches is compensated by the calibration DAC and the accurate successive approximation could be established.

III. Noises Considerations

For 12bit SAR ADC at 0.5V voltage, the 1LSB voltage is very small (244uV) and the input-referred noises of $Comp_1$ and $Comp_2$ would degenerate the ADC's performance drastically. As a result, it is very necessary to analyze how the two noises deteriorate the ADC's performance.

Because the $Comp_1$ is connected to the output of DAC in the normal conversion, its noise could be seen as one part of the input signal. As for the noise of $Comp_2$, it introduces some errors to the measurement of the capacitor's mismatch, resulting in the incomplete compensation to the capacitor's mismatch in the normal conversion. Consequently, the noise of $Comp_2$ could be converted to the mismatch-induced error of the ADC by some ratio. Therefore, the relationship between the two noises and the ADC's SNR would be demonstrated as

$$SNR = 10*\log\left(\frac{(\frac{2^N\Delta}{2\sqrt{2}})^2}{\frac{\Delta^2}{12}+V_{n,1}^2+V_e^2}\right) = 10*\log\left(\frac{(\frac{2^N\Delta}{2\sqrt{2}})^2}{\frac{\Delta^2}{12}+V_{n,1}^2+k^2V_{n,1}^2}\right). \quad (1)$$

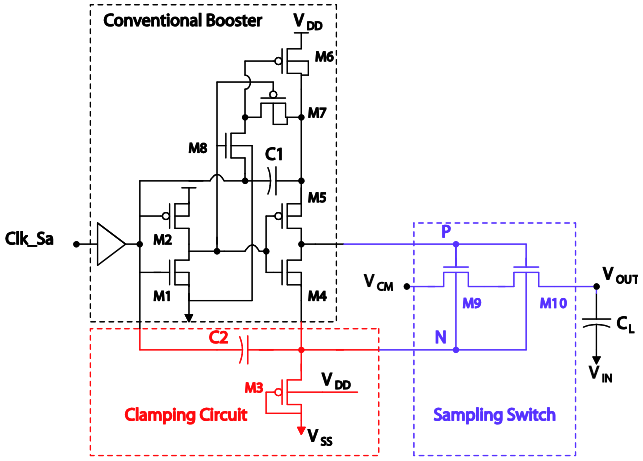


Fig. 3. Schematic of the sampling switch in SAR ADC.

Where $V_{n,1}$ and $V_{n,2}$ are the input-referred noises of $Comp_1$ and $Comp_2$, respectively. Δ is the value of 1LSB voltage and k is the conversion factor from $V_{n,2}$ to V_e (the mismatch-induced error).

In order to verify the analysis, a behavioral simulation of SAR ADC based on self calibration was performed. In the simulation, the values of unit capacitors in MSB DAC were taken to be Gaussian random variables with standard deviation of 3%.

Fig. 2 shows SNR of the ADC as functions of two comparators' noises. As expected, the noises of $Comp_1$ and $Comp_2$ both degrade the SNR nearly exponentially and the relation between noises and SNR approaches the relationship displayed by equation (1). Besides, the noise of $Comp_2$ deteriorates the SNR more drastically than that of $Comp_1$, which is due to the fact that the digital codes of mismatches including noises are add to compensate the error voltage in the normal conversion. However, $Comp_2$'s noise could be averaged out by multiple measurements of the same mismatch. In this design, we set the noise parameters of 60μV for $Comp_1$ and 40μV for $Comp_2$ to ensure more than 11bit ENOB.

IV. Circuit Considerations and Designs

At the circuit level, the challenges for the self-calibrated ADC under ultra low voltage lie in the leakage and the charge injection caused by the sampling switch and the input-referred offset of $Comp_2$. As a result, some techniques are presented to mitigate these interferences.

A. Ultra-low-leakage switch

For a SAR ADC operating at low voltage, the *on* resistance of sampling switch determines the bandwidth of SAR ADC while the leakage affects its linearity significantly. For self-calibrated SAR ADC, the *off* leakage would also distort the residual voltages in the measurements of mismatches and deteriorate the improved performance directly.

A ultra-low-leakage switch shown in Fig. 3 is proposed to mitigate the problem. The proposed switch is consisted of three parts, a conventional voltage booster, a voltage clamping circuit and a sampling switch. By producing a boosted voltage

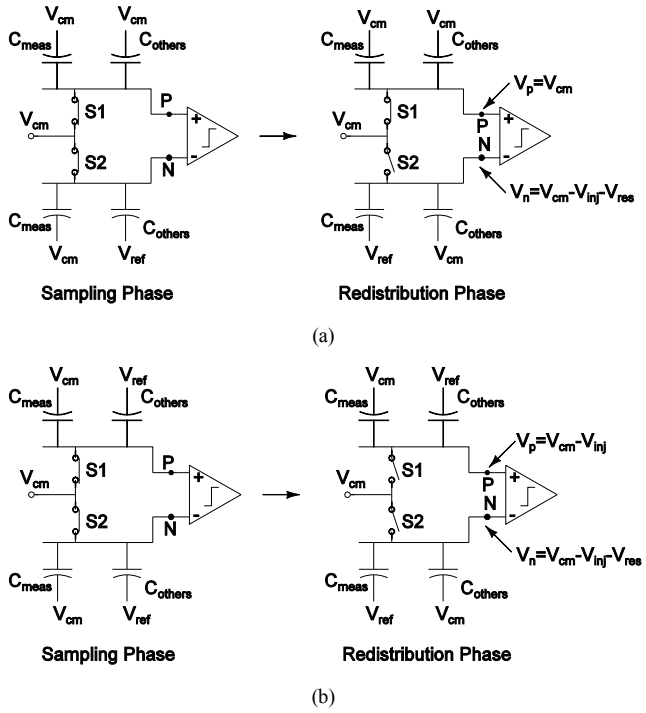


Fig. 4. The processes of digitizing the residual voltage: (a) the conventional and (b) the proposed. In the figure, V_{inj} is the voltage caused by charge injection and V_{res} is the residual voltage due to the mismatch between C_{meas} and C_{others}.

at the *on* phase and a negative voltage at the *off* phase, it can decrease the *off* leakage and increase the *on* conductance simultaneously. When the sampling clock Clk_Sa is "1", the voltage at node N is a small positive value and the voltage at node P is approximately $2V_{dd}$, which improves the *on* conductance. If Clk_Sa turns from "1" to "0", the voltages at nodes N and P turn to a same negative level. This would result in a negative *off* voltage at the gate and increase threshold voltage due to the negative voltage at the bulk, reducing more *off* current than conventional voltage boosting [7].

B. The charge injection of sampling switch

Fig. 4 (a) shows the conventional process of measuring the residual voltage on the bottom array. During the process, S2 is turned off after V_{ref} is sampled by C_{others} in the bottom array while S1 is turned on all the time. Therefore, charge injections caused by the switch S2 would introduce an offset to the measurement of the residual voltage.

At low supply voltage (0.5V), the amplitude of the residual voltage due to the mismatch is very small. For example, if the MSB capacitor includes 3% mismatch, the residual voltage will have the largest value of 3.75mV [3]. Because the residual voltage combined with the offset voltage is digitized by calibration DAC, the large offset voltage compared with the residual voltage would saturate measurement results and thus bring about the failed self calibration.

To mitigate this interference, we optimize the process of digitizing the residual voltage as shown in Fig. 4 (b). Take the measurement of residual voltage on the bottom array for example. In the sampling phase, S1 and S2 in Fig. 4 (b) turn on

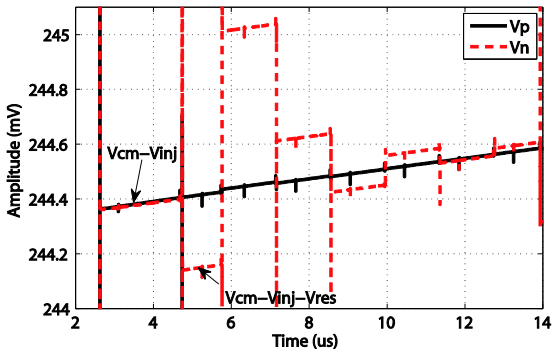


Fig. 5. The waveforms of V_p and V_n in Fig. 4 (b).

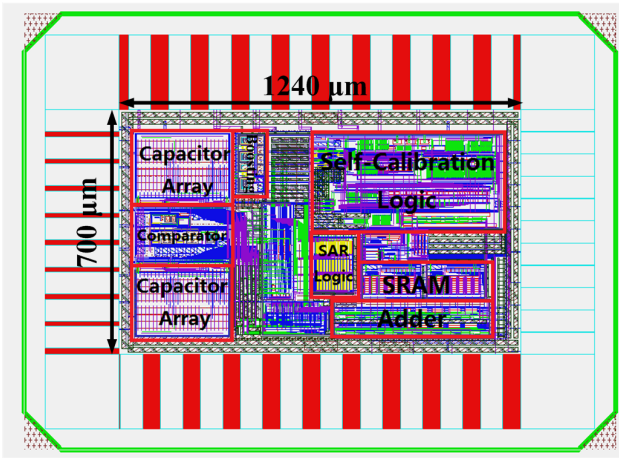


Fig. 6. Layout of the ADC.

at the same time and both the top and bottom array samples V_{ref}. When it comes to the redistribution phase, S1 and S2 turn off simultaneously. Since S1 and S2 have the same dimensions and circuit connections of the top and the bottom arrays are the same, the charge injections caused by S1 and S2 are nearly equal. Subsequently, only the bottom array are reversed between V_{cm} and V_{ref} and generates the residual voltage whereas connections of the top array remain unchanged. Finally, the residual voltage is digitized in a SAR conversion with the help of calibration DAC.

In the proposed process, the residual voltage is digitized through successive comparison with V_{cm}-V_{inj} instead of V_{cm}, so the offset caused by charge injection would be compensated effectively and the impact of the charge injection is alleviated.

Moreover, the improved process could alleviate the impact of the sampling switch's leakage. Fig. 5 shows the simulated transient voltage waveforms at nodes N and P in the measurement. Because S1 and S2 are both turned off and V_p is close to V_n, the error voltage due to S1's leakage approaches that of S2, as shown in Fig. 5. Consequently, the leakage induced error voltage of S2 could be compensated by that of S1 effectively.

C. Comparator with offset cancellation

To digitize the small residual voltage with enough accuracy, $Comp_2$'s resolution should be very high. In this work, $Comp_2$ was implemented by two pre-amplifiers and a latch to ensure enough gain. Meanwhile, the output offset cancelation was exploited in $Comp_2$ to diminish the offset voltage and avoid the saturation of measurement results.

V. Simulated Results and Discussions

A 12b 1MS/s SAR ADC at 0.5V has been implemented in a 0.13μm CMOS technology, occupying 1.2mm × 0.7mm active area. Fig. 6 shows the layout of the ADC. In the simulation at the circuit level, random capacitor mismatch up to 3% was adopted, which covers 99.7% of actual mismatch distribution with 1% of σ. And parasitic capacitors with the values of 3% are contained in the simulation.

A. Dynamic Performance

The dynamic performance of the ADC without calibration and that with calibration is shown in Fig. 7. The ENOB is improved from 9bit to 11.4bit by self calibration. What's more, SFDR of 90dB is achieved compared with 64.3dB before calibration. It could be clearly shown that the noises and circuit considerations in the preceding sections ensure the performance of self-calibrated ADC at low voltage.

B. Power Consumption and FOM

At sampling rate of 1MS/s, the ADC draws 35.7μW from a 0.5V supply voltage. The percentage of power consumption for digital logic, comparators and capacitor array are 37.2%, 56% and 6.7%, respectively. The less power consumption of capacitor array results from the small unit capacitance. In this design, more than half of the power is allocated to comparators due to the noise requirements.

The figure of merit (FOM) for Nyquist converters refers to the energy required to accomplish an effective conversion step. The FOM is defined as

$$FOM = \frac{Power}{2^{ENOB} \cdot f_s}. \qquad (2)$$

The FOM of the proposed ADC is 13.2fJ/conversion-step at 12b 1MS/s, which is a competitive value. The high power efficiency is partially a result of the low supply voltage that decreases the power of digital circuits quadratically. Meanwhile, due to the self calibration, the unit capacitance is able to be very small, decreasing the power of DAC significantly.

VI. Conclusions

This paper elaborates the analysis and design of a 12bit 1MS/s SAR ADC at 0.5V with self calibration. The effects of noises on the performance of self-calibrated ADC are demonstrated. Circuit considerations and designs on ultra-low voltage self-calibrated SAR ADC are also described. Simulated in a 0.13μm CMOS, the proposed ADC exhibits 11.4bit ENOB and 90dB SFDR with capacitor mismatch up to 3%. A remarkable power efficiency of 13.2fJ/step FOM has also been achieved.

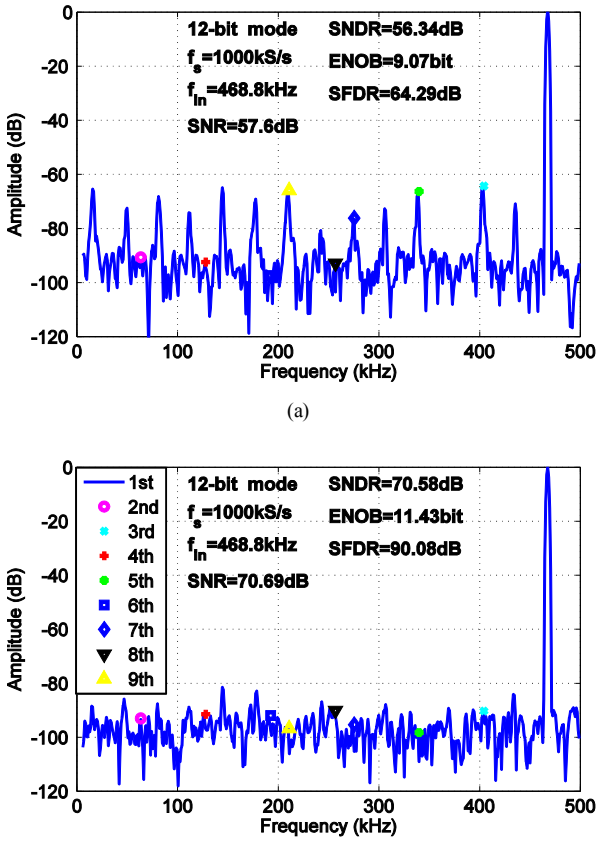


Fig. 7. Dynamic performance with a 468.8kHz input at 1MS/s sampling rate: (a) before calibration (b) after calibration.

References

[1] W. Liu, P. Huang, and Y. Chiu, "A 12-bit, 45-MS/s, 3-mW redundant successive-approximation-register analog-to-digital converter with digital calibration," *IEEE J. Solid-State Circuits*, vol. 46, pp. 2661-2672, Nov. 2011.

[2] X. Zhou and Q. Li, "A 160mV 670nW 8-bit SAR ADC in 0.13μm CMOS," *2012 IEEE Custom Integrated Circuits Conference*, pp. 1-4, Sep. 2012.

[3] H. S. Lee, D. A. Hodges, and P. R. Gray, "A self-calibrating 15-bit CMOS A/D converter," *IEEE J. Solid-State Circuits*, Vol. SC-19, pp.813-819, Dec. 1984

[4] M. Yoshioka, K. Ishikawa, T. Takayama, and S. Tsukamoto, "A 10b 50MS/s 820uW SAR ADC with On-Chip Digital Cali-bration," *IEEE Transactions on Biomedical Circuits and Systems*, Vol. 4, NO. 6, pp410-416, Dec. 2010.

[5] Y. Kuramochi, A. Matsuzawa, and M. Kawabata, "A 0.05-mm2 110-μW 10-b self-calibrating successive approximation ADC core in 0.18-μm CMOS," in *Proc. IEEE Asian Solid-State Circuits Conf.*, Nov. 2007, pp. 224–227.

[6] J-Y. Um, J-H Kim, J-Y Sim, and H-J Park, "Digital-Domain Calibration of Split-Capacitor DAC with no Extra Calibration DAC for a Differential-Type SAR ADC," in *Proc. IEEE Asian Solid-State Circuits Conf.*, Nov. 2011, pp. 77–80.

[7] H. Huang, K. Ao and Q. Li, "0.5V Rate-Resolution Scalable SAR ADC with 63.7dB SFDR," *2013 IEEE International Symposium on Circuits and Systems*, to appear.

Author Index

A B C D E F G H I J K L M N O P Q R S T U V W Y Z

Abreu, Daniel
Special Frames

Adcock, Ben
Optimal wavelet reconstructions from Fourier samples via generalized sampling

Overcoming the coherence barrier in compressed sensing

Aldroubi, Akram
Binary Reduced Row Echelon Form Approach for Subspace Segmentation

Allard, Bruno
Efficient Simulation of Continuous Time Digital Signal Processing RF Systems

Amini, Arash
Optimal Interpolation Laws for Stable AR(1) Processes

Angelini, Elsa
Video sampling and reconstruction using linear or non-linear Fourier measurements

Angeloni, Laura
Variation and approximation for Mellin-type operators

Anitori, Laura
Compressive CFAR Radar Processing

Antoine, Philippe
Compressive Acquisition of Sparse Deflectometric Maps

Arce, Gonzalo
Sparse 2D Fast Fourier Transform

Arias, Arnaud
Efficient Simulation of Continuous Time Digital Signal Processing RF Systems

Atallah, Elie
Recovery of bilevel causal signals with finite rate of innovation using positive sampling kernels

Au-Yeung, Enrico
Balayage and short time Fourier transform frames

Tight frames in spiral sampling

August, Yitzhak
Challenges in Optical Compressive Imaging and Some Solutions

Averbuch, Amir
Constructive sampling for patch-based embedding

Measure-based diffusion kernel methods

Missing Entries Matrix Approximation and Completion

Particle Filter Acceleration Using Multiscale Sampling Methods

Spline-based frames for image restoration

Using Affinity Perturbations to Detect Web Traffic Anomalies

Ayaz, Ulas
Sparse Recovery with Fusion Frames via RIP

Azghani, Masomeh
iterative methods for random sampling recovery and compressed sensing recovery

Baboulaz, Loic
Optimal Sampling Rates in Infinite-Dimensional Compressed Sensing

Baechler, Gilles
Multichannel ECG Analysis using VPW-FRI

Bah, Bubacarr
Energy-aware adaptive bi-Lipschitz embeddings

On construction and analysis of sparse matrices and expander graphs with applications to CS

Balazs, Peter
(Non-)Density Properties of Discrete Gabor Multipliers

A Review of the Invertibility of Frame Multipliers

Gabor dual windows using convex optimization

Bandeira, Afonso
Fundamental Limits of Phase Retrieval

Bar-Ilan, Omer
FRI-based Sub-Nyquist Sampling and Beamforming in Ultrasound and Radar

Baraniuk, Richard
Compressive CFAR Radar Processing

Test-size Reduction Using Sparse Factor Analysis

Bardaro, Carlo
Mellin analysis and exponential sampling. Part I: Mellin fractional integrals

Mellin analysis and exponential sampling. Part II: Mellin differential operators and sampling

Bartels, Andreas
Hybrid Regularization and Sparse Reconstruction of Imaging Mass Spectrometry Data

Batenkov, Dmitry
Algebraic signal sampling, Gibbs phenomenon and Prony-type systems

Decoupling of Fourier Reconstruction System for Shifts of Several Signals

Bayer, Dominik
(Non-)Density Properties of Discrete Gabor Multipliers

Beck, Amir
GESPAR: Efficient Sparse Phase Retrieval with Application to Optics

Becker, Stephen
Randomized Singular Value Projection

Benedetto, John
Balayage and short time Fourier transform frames

Bermanis, Amit
Constructive sampling for patch-based embedding

Measure-based diffusion kernel methods

Particle Filter Acceleration Using Multiscale Sampling Methods

Bernstein, Swanhild
A Lie group approach to diffusive wavelets

Beyrouthy, Taha
Non-uniform sampling pattern recognition based on atomic decomposition

Bidegaray-Fesquet, Brigitte
Level crossing sampling of strongly monoHölder functions

Bigot, Jérémie
Sampling by blocks of measurements in Compressed Sensing

Bilen, Cagdas
Blind Sensor Calibration in Sparse Recovery Using Convex Optimization

Blu, Thierry
Approximate FRI with Arbitrary Kernels

Localization of point sources in wave fields from boundary measurements using new sensing principle

Boche, Holger
Phase Retrieval via Structured Modulations in Paley-Wiener Spaces

Sampling and Reconstruction of Bandlimited BMO-Functions

Signal Analysis with Frame Theory and Persistent Homology

Bonanno, Gabriele
Joint reconstruction of misaligned images from incomplete measurements for cardiac MRI

Bonvilain, Agnès
Non-uniform sampling pattern recognition based on atomic decomposition

Borkholder, David
A Comparison of Reconstruction Methods for Compressed Sensing of the Photoplethysmogram

Bossert, Martin
On the Noise-Resilience of OMP with BASC-Based Low Coherence Sensing Matrices

Bostan, Emrah
MAP Estimators for Self-Similar Sparse Stochastic Models

Boufounos, Petros
Sparse Signal Reconstruction from Phase-only Measurements

Boyer, Claire
Sampling by blocks of measurements in Compressed Sensing

Brady, David
Coding and sampling for compressive tomography

Bullkich, Elad
Sub-Wavelength Coherent Diffractive Imaging based on Sparsity

Butzer, Paul
Mellin analysis and exponential sampling. Part I: Mellin fractional integrals

Mellin analysis and exponential sampling. Part II: Mellin differential operators and sampling

Cahill, Jameson
A note on scalable frames

Cahill, Jameson
Fundamental Limits of Phase Retrieval

Campman, Xander
Jointly filtering and regularizing seismic data using space-varying FIR filters

Carlini, Lina
Fast Maximum Likelihood High-density Low-SNR Super-resolution Localization Microscopy

Carrillo, Rafael
On Sparsity Averaging

Casey, Stephen
Signal Adaptive Frame Theory

Castro, Rui
On the Performance of Adaptive Sensing for Sparse Signal Inference

Cevher, Volkan
Energy-aware adaptive bi-Lipschitz embeddings

Randomized Singular Value Projection

Chang, Soon-Jyh
Multi-Step Switching Methods for SAR ADCs

Chardon, Gilles
Reconstruction of solutions to the Helmholtz equation from punctual measurements

Chauffert, Nicolas
From variable density sampling to continuous sampling using Markov chains

Travelling salesman-based variable density sampling

Chen, Mike Shuo-Wei
Trend of High-Speed SAR ADC towards RF Sampling

Chen, Xuemei
A note on scalable frames

A null space property approach to compressed sensing with frames

Chernyakova, Tanya
FRI-based Sub-Nyquist Sampling and Beamforming in Ultrasound and Radar

Chiu, Yun
Digital Calibration of SAR ADC

Chou, Evan
Non-Convex Decoding for Sigma Delta Quantized Compressed Sensing

Christensen, Ole
On transformations between Gabor frames and wavelet frames

Ciuciu, Philippe
From variable density sampling to continuous sampling using Markov chains

Travelling salesman-based variable density sampling

Clausel, Marianne
Level crossing sampling of strongly monoHölder functions

Cluni, Federico
Multivariate sampling Kantorovich operators: approximation and applications to civil engineering

Cohen, Albert
Reconstruction of solutions to the Helmholtz equation from punctual measurements

Cohen, Deborah
Spectrum Reconstruction from Sub-Nyquist Sampling of Stationary Wideband Signals

Cohen, Oren
Sub-Wavelength Coherent Diffractive Imaging based on Sparsity

Condat, Laurent
Robust Spike Train Recovery from Noisy Data by Structured Low Rank Approximation

Conn, Nicholas
A Comparison of Reconstruction Methods for Compressed Sensing of the Photoplethysmogram

Costarelli, Danilo
Multivariate sampling Kantorovich operators: approximation and applications to civil engineering

Crochiere, Ronald
Multichannel ECG Analysis using VPW-FRI

Da Silva, Curt
Hierarchical Tucker Tensor Optimization - Applications to Tensor Completion

Dai, Wei
Tracking Dynamic Sparse Signals with Kalman Filters: Framework and Improved Inference

Dana, Hod
Sub-Wavelength Coherent Diffractive Imaging based on Sparsity

Danilouchkine, Mike
Jointly filtering and regularizing seismic data using space-varying FIR filters

Datta, Somantika
Tight frames in spiral sampling

Daudet, Laurent
Blind Sensor Calibration in Sparse Recovery Using Convex Optimization

Reconstruction of solutions to the Helmholtz equation from punctual measurements

De Vleeschouwer, Christophe
Quantized Iterative Hard Thresholding: Bridging 1bit and HighResolution Quantized Compressed Sensing

Degraux, Kévin
Quantized Iterative Hard Thresholding: Bridging 1bit and HighResolution Quantized Compressed Sensing

Deledalle, Charles-Alban
Stable Recovery with Analysis Decomposable Priors

Demanet, Laurent
Super-resolution via superset selection and pruning

Demaret, Laurent
Irregular Sampling of the Radon Transform of Bandlimited Functions

Wavelet Signs: A New Tool for Signal Analysis

Divekar, Atul
Using Correlated Subset Structure for Compressive Sensing Recovery

Doerfler, Monika
Frames of eigenspaces and localization of signal components

Tracing Sound Objects in Audio Textures

Dogan, Zafer
Localization of point sources in wave fields from boundary measurements using new sensing principle

Dragotti, Pier Luigi
Approximate FRI with Arbitrary Kernels

Dubois, Xavier
Compressive Acquisition of Sparse Deflectometric Maps

Eggenberger, Kurt
Reconstruction of Signals from Highly Aliased Multichannel Samples by Generalized Matching Pursuit

Ehler, Martin
Phase retrieval using time and Fourier magnitude measurements

Elad, Michael
Iterative Hard Thresholding with Near Optimal Projection for Signal Recovery

OMP with Highly Coherent Dictionaries

Eldar, Yonina
FRI-based Sub-Nyquist Sampling and Beamforming in Ultrasound and Radar

GESPAR: Efficient Sparse Phase Retrieval with Application to Optics

Spectrum Reconstruction from Sub-Nyquist Sampling of Stationary Wideband Signals

Sub-Wavelength Coherent Diffractive Imaging based on Sparsity

Fadili, Jalal
Robust Polyhedral Regularization

Stable Recovery with Analysis Decomposable Priors

Fageot, Julien
MAP Estimators for Self-Similar Sparse Stochastic Models

Fatemi, Mitra
Optimal Sampling Rates in Infinite-Dimensional Compressed Sensing

Fernandez-Granda, Carlos
Support detection in super-resolution

Fernández-Morales, Héctor
Generalized sampling in U-invariant subspaces

Fesquet, Laurent
Non-uniform sampling pattern recognition based on atomic decomposition

Fickus, Matthew
Characterizing completions of finite frames

Frossard, Pascal
Analysis of Hierarchical Image Alignment with Descent Methods

Tangent space estimation bounds for smooth manifolds

Führ, Hartmut
Orlicz Modulation Spaces

Gabriel, Turinici
A priori convergence of the Generalized Empirical Interpolation Method

Gamboa, Fabrice
From variable density sampling to continuous sampling using Markov chains

Gan, Lu
Deterministic Binary Sequences for Modulated Wideband Converter

García, Antonio
Generalized sampling in U-invariant subspaces

Gehm, Michael
Calibration—An open challenge in creating practical computational- and compressive-sensing systems

Giryes, Raja
Iterative Hard Thresholding with Near Optimal Projection for Signal Recovery

OMP with Highly Coherent Dictionaries

Goh, Say
On transformations between Gabor frames and wavelet frames

Golubov, Boris
Absolute Convergence of the Series of Fourier-Haar Coefficients

Gonzalez Gonzalez, Adriana
Compressive Acquisition of Sparse Deflectometric Maps

Goodman, Nathan
Measurement Structures and Constraints in Compressive RF Systems

Gribonval, Rémi
Blind Sensor Calibration in Sparse Recovery Using Convex Optimization

Gryboś, Anna
Recovery of Bandlimited Signal Based on Nonuniform Derivative Sampling

Guillemard, Mijail
Signal Analysis with Frame Theory and Persistent Homology

Hand, Paul
Conditions for Dual Certificate Existence in Semidefinite Rank-1 Matrix Recovery

Hansen, Anders
Optimal wavelet reconstructions from Fourier samples via generalized sampling

Overcoming the coherence barrier in compressed sensing

Hegde, Chinmay
The Constrained Earth Mover Distance Model, with Applications to Compressive Sensing

Heider, Sabine
A sparse Prony FFT

Heilemann, Mike
From super-resolution microscopy towards quantitative single-molecule biology

Heise, Bettina
Analogies and differences in optical and mathematical systems and approaches

Hernández-Medina, Miguel
Generalized sampling in U-invariant subspaces

Herrmann, Felix
Hierarchical Tucker Tensor Optimization - Applications to Tensor Completion

Hirabayashi, Akira
Robust Spike Train Recovery from Noisy Data by Structured Low Rank Approximation

Hogan, Jeffrey
Sampling aspects of approximately time-limited multiband and bandpass signals

Holden, Seamus
Optimisation and control of sampling rate in localisation microscopy

Holighaus, Nicki
Gabor dual windows using convex optimization

Howlett, Phil
Estimation of large data sets on the basis of sparse sampling

Huali, Wang
Deterministic Binary Sequences for Modulated Wideband Converter

Huang, Bo
STORM by compressed sensing

Huang, Guan-Ying
Multi-Step Switching Methods for SAR ADCs

Huang, Hai
Design Considerations of Ultra-Low-Voltage Self-Calibrated SAR ADC

Hur, Youngmi
The Design of Non-redundant Directional Wavelet Filter Bank Using 1-D Neville Filters

Indyk, Piotr
The Constrained Earth Mover Distance Model, with Applications to Compressive Sensing

Jacques, Laurent
Compressive Acquisition of Sparse Deflectometric Maps

Quantized Iterative Hard Thresholding: Bridging 1bit and HighResolution Quantized Compressed Sensing

Jeon, Daejong
Fast Maximum Likelihood High-density Low-SNR Super-resolution Localization Microscopy

Joannes, Luc
Compressive Acquisition of Sparse Deflectometric Maps

Jovanovic, Ivana
Localization of point sources in wave fields from boundary measurements using new sensing principle

Kabanava, Maryia
Recovery of cosparse signals with Gaussian measurements

Kahn, Jonas
Travelling salesman-based variable density sampling

Kamada, Masaru
Sparse Approximation of Ion-Mobility Spectrometry Profiles by Minutely Shifted Discrete B-splines

Kamilov, Ulugbek
MAP Estimators for Self-Similar Sparse Stochastic Models

Karandikar, Abhay
Finite Rate of Innovation Signals: Quantization Analysis with Resistor-Capacitor Acquisition Filter

Karseras, Evripidis
Tracking Dynamic Sparse Signals with Kalman Filters: Framework and Improved Inference

Kim, Kyung Sang
Fast Maximum Likelihood High-density Low-SNR Super-resolution Localization Microscopy

Kirshner, Hagai
Identification of Rational Transfer Functions from Sampled Data

Kivinukk, Andi
Approximation by Shannon sampling operators in terms of an averaged modulus of smoothness

The Variation Detracting Property of some Shannon Sampling Series and their Derivatives

Kontakis, Apostolos
Jointly filtering and regularizing seismic data using space-varying FIR filters

Koscielnik, Dariusz
Recovery of Bandlimited Signal Based on Nonuniform Derivative Sampling

Krahmer, Felix
Local coherence sampling for stable sparse recovery

Sigma-Delta quantization of sub-Gaussian compressed sensing measurements

Spectral properties of dual frames

The restricted isometry property for random convolutions

Krueger, Kyle
Sampling Techniques for Improved Algorithmic Efficiency in Electromagnetic Sensing

Kumar, Animesh
Bandlimited Signal Reconstruction From the Distribution of Unknown Sampling Locations

Finite Rate of Innovation Signals: Quantization Analysis with Resistor-Capacitor Acquisition Filter

Kunis, Stefan
A sparse Prony FFT

Phase retrieval using time and Fourier magnitude measurements

Kuo, Che-Hsun
Multi-Step Switching Methods for SAR ADCs

Kutyniok, Gitta
Perfect Preconditioning of Frames by a Diagonal Operator

Signal Analysis with Frame Theory and Persistent Homology

Spectral properties of dual frames

Kyrillidis, Anastasios
Randomized Singular Value Projection

Laga, Hamid
Estimation of large data sets on the basis of sparse sampling

Lakey, Joseph
Sampling aspects of approximately time-limited multiband and bandpass signals

Lazich, Dejan
On the Noise-Resilience of OMP with BASC-Based Low Coherence Sensing Matrices

Le Pelleter, Tugdual
Non-uniform sampling pattern recognition based on atomic decomposition

Le-Montagner, Yoann
Video sampling and reconstruction using linear or non-linear Fourier measurements

Leistedt, Boris
Fourier-Laguerre transform, convolution and wavelets on the ball

Lemvig, Jakob
Spectral properties of dual frames

Leung, Kin K.
Tracking Dynamic Sparse Signals with Kalman Filters: Framework and Improved Inference

Leus, Geert
Jointly filtering and regularizing seismic data using space-varying FIR filters

Levitina, Tatiana
On the Number of Degrees of Freedom of Band-Limited Functions

Li, Qiang
Design Considerations of Ultra-Low-Voltage Self-Calibrated SAR ADC

Lin, Ying-Zu
Multi-Step Switching Methods for SAR ADCs

Lin-Shi, Xuefang
Efficient Simulation of Continuous Time Digital Signal Processing RF Systems

Liu, Chun-Cheng
Multi-Step Switching Methods for SAR ADCs

Maday, Yvon
A priori convergence of the Generalized Empirical Interpolation Method

Maleki, Arian
Compressive CFAR Radar Processing

Manley, Suliana
Fast Maximum Likelihood High-density Low-SNR Super-resolution Localization Microscopy

Optimisation and control of sampling rate in localisation microscopy

Mantellini, Ilaria
Mellin analysis and exponential sampling. Part I: Mellin fractional integrals

Mellin analysis and exponential sampling. Part II: Mellin differential operators and sampling

Marvasti, Farokh
iterative methods for random sampling recovery and compressed sensing recovery

Marziliano, Pina
Multichannel ECG Analysis using VPW-FRI

Massopust, Peter
Wavelet Signs: A New Tool for Signal Analysis

Matusiak, Ewa
Tracing Sound Objects in Audio Textures

McClellan, James
Sampling Techniques for Improved Algorithmic Efficiency in Electromagnetic Sensing

McEwen, Jason
Fourier-Laguerre transform, convolution and wavelets on the ball

On Sparsity Averaging

Mendelson, Shahar
The restricted isometry property for random convolutions

Metsmägi, Tarmo
The Variation Detracting Property of some Shannon Sampling Series and their Derivatives

Min, Junhong
Fast Maximum Likelihood High-density Low-SNR Super-resolution Localization Microscopy

Minotti, Anna Maria
Multivariate sampling Kantorovich operators: approximation and applications to civil engineering

Miskowicz, Marek
Recovery of Bandlimited Signal Based on Nonuniform Derivative Sampling

Mixon, Dustin
Fundamental Limits of Phase Retrieval

Mohamadian, Habib
Analysis of Multistage Sampling Rate Conversion for Potential Optimal Factorization

Mönich, Ullrich
Sampling and Reconstruction of Bandlimited BMO-Functions

Morche, Dominique
Efficient Simulation of Continuous Time Digital Signal Processing RF Systems

Mroueh, Youssef
q-ary compressive sensing

Mula, Olga
A priori convergence of the Generalized Empirical Interpolation Method

Murmann, Boris
On the use of redundancy in successive approximation A/D converters

Nair, Amrish
Multichannel ECG Analysis using VPW-FRI

Nam, Sangnam
An Uncertainty Principle for Discrete Signals

Needell, Deanna
Super-resolution via superset selection and pruning

Using Correlated Subset Structure for Compressive Sensing Recovery

Neittaanmäki, Pekka
Spline-based frames for image restoration

Nelson, Aaron
Fundamental Limits of Phase Retrieval

Nestler, Franziska
Fast Ewald summation under 2d- and 1d-periodic boundary conditions based on NFFTs

Nguyen, Nam
Super-resolution via superset selection and pruning

Nguyen, Truong Thao
Finite-power spectral analytic framework for quantized sampled signals

Ohno, Masakazu
Sparse Approximation of Ion-Mobility Spectrometry Profiles by Minutely Shifted Discrete B-splines

Okoudjou, Kasso
Perfect Preconditioning of Frames by a Diagonal Operator

Olivo-Marin, Jean-Christophe
Video sampling and reconstruction using linear or non-linear Fourier measurements

Omer, Harold
Estimation of frequency modulations on wideband signals; applications to audio signal analysis

Osherovich, Eliahyu
Sub-Wavelength Coherent Diffractive Imaging based on Sparsity

Otten, Matern
Compressive CFAR Radar Processing

Ozbek, Ali
Reconstruction of Signals from Highly Aliased Multichannel Samples by Generalized Matching Pursuit

Ozdemir, Kemal
Reconstruction of Signals from Highly Aliased Multichannel Samples by Generalized Matching Pursuit

Pawlak, Mirek
Joint Signal Sampling and Detection

Pengo, Thomas
Optimisation and control of sampling rate in localisation microscopy

Perraudin, Nathanaël
Gabor dual windows using convex optimization

Pesenson, Isaac
Shannon Sampling and Parseval Frames on Compact Manifolds

Peyré, Gabriel
Robust Polyhedral Regularization

Stable Recovery with Analysis Decomposable Priors

Pfander, Goetz
Sparse Finite Gabor Frames for Operator Sampling

Philipp, Friedrich
Signal Analysis with Frame Theory and Persistent Homology

Plan, Yaniv
Structured-signal recovery from single-bit measurements

Pohl, Volker
Phase Retrieval via Structured Modulations in Paley-Wiener Spaces

Poon, Clarice
Optimal wavelet reconstructions from Fourier samples via generalized sampling

Overcoming the coherence barrier in compressed sensing

Poteet, Miriam
Characterizing completions of finite frames

Potts, Daniel
A sparse Prony FFT

Fast Ewald summation under 2d- and 1d-periodic boundary conditions based on NFFTs

Puy, Gilles
Blind Sensor Calibration in Sparse Recovery Using Convex Optimization

Joint reconstruction of misaligned images from incomplete measurements for cardiac MRI

Quick, Frank
Multichannel ECG Analysis using VPW-FRI

Ramesh, Gayatri
Recovery of bilevel causal signals with finite rate of innovation using positive sampling kernels

Ratiu, Alin
Efficient Simulation of Continuous Time Digital Signal Processing RF Systems

Rauh, Andre
Sparse 2D Fast Fourier Transform

Rauhut, Holger
Local coherence sampling for stable sparse recovery

Low-rank Tensor Recovery via Iterative Hard Thresholding

Recovery of cosparse signals with Gaussian measurements

Sparse Recovery with Fusion Frames via RIP

The restricted isometry property for random convolutions

Reinhardt, Martin
Analogies and differences in optical and mathematical systems and approaches

Rivenson, Yair
Challenges in Optical Compressive Imaging and Some Solutions

Robinson, Michael
The Nyquist theorem for cellular sheaves

Rolland, Robin
Non-uniform sampling pattern recognition based on atomic decomposition

Roman, Bogdan
Overcoming the coherence barrier in compressed sensing

Romero, José Luis
Frames of eigenspaces and localization of signal components

Rosasco, Lorenzo
q-ary compressive sensing

Rzepka, Dominik
Recovery of Bandlimited Signal Based on Nonuniform Derivative Sampling

Saab, Rayan
Sigma-Delta quantization of sub-Gaussian compressed sensing measurements

Sadeghi, Bashir
Shift-Variance and Cyclostationarity of Linear Periodically Shift-Variant Systems

Sadeghian, Ali
Energy-aware adaptive bi-Lipschitz embeddings

Salhov, Moshe
Constructive sampling for patch-based embedding

Salmon, Joseph
Stable Recovery with Analysis Decomposable Priors

Sarig, Niv
Decoupling of Fourier Reconstruction System for Shifts of Several Signals

Schausberger, Stefan
Analogies and differences in optical and mathematical systems and approaches

Schmidt, Ludwig
The Constrained Earth Mover Distance Model, with Applications to Compressive Sensing

Schnackers, Catherine
Orlicz Modulation Spaces

Schnass, Karin
Dictionary Identification Results for K-SVD with Sparsity Parameter 1

Schneider, Reinhold
Low-rank Tensor Recovery via Iterative Hard Thresholding

Schnitzbauer, Joerg
STORM by compressed sensing

Scott, Waymond
Sampling Techniques for Improved Algorithmic Efficiency in Electromagnetic Sensing

Segev, Mordechai
Sub-Wavelength Coherent Diffractive Imaging based on Sparsity

Sekmen, Ali
Binary Reduced Row Echelon Form Approach for Subspace Segmentation

Shabat, Gil
Missing Entries Matrix Approximation and Completion

Particle Filter Acceleration Using Multiscale Sampling Methods

Using Affinity Perturbations to Detect Web Traffic Anomalies

Shechtman, Yoav
GESPAR: Efficient Sparse Phase Retrieval with Application to Optics

Sub-Wavelength Coherent Diffractive Imaging based on Sparsity

Shmueli, Yaniv
Missing Entries Matrix Approximation and Completion

Particle Filter Acceleration Using Multiscale Sampling Methods

Using Affinity Perturbations to Detect Web Traffic Anomalies

Shoham, Shy
Sub-Wavelength Coherent Diffractive Imaging based on Sparsity

Shutin, Dmitriy
Incremental Sparse Bayesian Learning for Parameter Estimation of Superimposed Signals

Shyu, Ya-Ting
Multi-Step Switching Methods for SAR ADCs

Sidorenko, Pavel
Sub-Wavelength Coherent Diffractive Imaging based on Sparsity

Simon, Loic
Truncation Error in Image Interpolation

Sipola, Tuomo
Using Affinity Perturbations to Detect Web Traffic Anomalies

Soendergaard, Peter
Gabor dual windows using convex optimization

Stern, Adrian
Challenges in Optical Compressive Imaging and Some Solutions

Stifter, David
Analogies and differences in optical and mathematical systems and approaches

Stoeva, Diana
A Review of the Invertibility of Frame Multipliers

Stojanac, Zeljka
Low-rank Tensor Recovery via Iterative Hard Thresholding

Storath, Martin
Wavelet Signs: A New Tool for Signal Analysis

Strohmer, Thomas
Sparse MIMO Radar with Random Sensor Arrays and Kerdock Codes

Stuber, Matthias
Joint reconstruction of misaligned images from incomplete measurements for cardiac MRI

Studer, Christoph
Test-size Reduction Using Sparse Factor Analysis

Sudhakar, Prasad
Compressive Acquisition of Sparse Deflectometric Maps

Sun, Qiyu
Recovery of bilevel causal signals with finite rate of innovation using positive sampling kernels

Szameit, Alexander
Sub-Wavelength Coherent Diffractive Imaging based on Sparsity

Tamberg, Gert
Approximation by Shannon sampling operators in terms of an averaged modulus of smoothness

Tang, Zijian
Jointly filtering and regularizing seismic data using space-varying FIR filters

Tanner, Jared
On construction and analysis of sparse matrices and expander graphs with applications to CS

Tenneti, Srikanth
Finite Rate of Innovation Signals: Quantization Analysis with Resistor-Capacitor Acquisition Filter

Thomas, Jost
Incremental Sparse Bayesian Learning for Parameter Estimation of Superimposed Signals

Torokhti, Anatoli
Estimation of large data sets on the basis of sparse sampling

Torrésani, Bruno
Estimation of frequency modulations on wideband signals; applications to audio signal analysis

Tyagi, Hemant
Tangent space estimation bounds for smooth manifolds

Unnikrishnan, Jayakrishnan
On Optimal Sampling Trajectories for Mobile Sensing

Unser, Michael
Fast Maximum Likelihood High-density Low-SNR Super-resolution Localization Microscopy

Identification of Rational Transfer Functions from Sampled Data

MAP Estimators for Self-Similar Sparse Stochastic Models

Uriguen, Jose Antonio
Approximate FRI with Arbitrary Kernels

Vaiter, Samuel
Robust Polyhedral Regularization

Stable Recovery with Analysis Decomposable Priors

Van De Ville, Dimitri
Localization of point sources in wave fields from boundary measurements using new sensing principle

van Manen, Dirk-Jan
Reconstruction of Signals from Highly Aliased Multichannel Samples by Generalized Matching Pursuit

van Rossum, Wim
Compressive CFAR Radar Processing

Vandergheynst, Pierre
Joint reconstruction of misaligned images from incomplete measurements for cardiac MRI

Vassallo, Massimiliano
Reconstruction of Signals from Highly Aliased Multichannel Samples by Generalized Matching Pursuit

Vats, Divyanshu
Test-size Reduction Using Sparse Factor Analysis

Veit, Michael
A sparse Prony FFT

Verdier, Jacques
Efficient Simulation of Continuous Time Digital Signal Processing RF Systems

Vetterli, Martin
On Optimal Sampling Trajectories for Mobile Sensing

Optimal Sampling Rates in Infinite-Dimensional Compressed Sensing

Vinti, Gianluca
Multivariate sampling Kantorovich operators: approximation and applications to civil engineering

Variation and approximation for Mellin-type operators

Volkmer, Toni
Taylor and rank-1 lattice based nonequispaced fast Fourier transform

Volosivets, Sergey
Absolute Convergence of the Series of Fourier-Haar Coefficients

Vural, Elif
Analysis of Hierarchical Image Alignment with Descent Methods

Tangent space estimation bounds for smooth manifolds

Walnut, David
Sparse Finite Gabor Frames for Operator Sampling

Wang, Haichao
A null space property approach to compressed sensing with frames

Sparse MIMO Radar with Random Sensor Arrays and Kerdock Codes

Wang, Rongrong
A null space property approach to compressed sensing with frames

Wang, Wei
Incremental Sparse Bayesian Learning for Parameter Estimation of Superimposed Signals

Wang, Xiaoyang
Design Considerations of Ultra-Low-Voltage Self-Calibrated SAR ADC

Ward, John Paul
Identification of Rational Transfer Functions from Sampled Data

Ward, Rachel
Local coherence sampling for stable sparse recovery

Weiss, Pierre
From variable density sampling to continuous sampling using Markov chains

Sampling by blocks of measurements in Compressed Sensing

Travelling salesman-based variable density sampling

Wiaux, Yves
On Sparsity Averaging

Wiese, Thomas
Irregular Sampling of the Radon Transform of Bandlimited Functions

Wolf, Guy
Constructive sampling for patch-based embedding

Measure-based diffusion kernel methods

Yang, Fanny
Phase Retrieval via Structured Modulations in Paley-Wiener Spaces

Yavneh, Irad
Sub-Wavelength Coherent Diffractive Imaging based on Sparsity

Ye, Jong Chul
Fast Maximum Likelihood High-density Low-SNR Super-resolution Localization Microscopy

Ye, Zhengmao
Analysis of Multistage Sampling Rate Conversion for Potential Optimal Factorization

Yilmaz, Ozgur
Sigma-Delta quantization of sub-Gaussian compressed sensing measurements

Yomdin, Yosef
Algebraic signal sampling, Gibbs phenomenon and Prony-type systems

Decoupling of Fourier Reconstruction System for Shifts of Several Signals

Yu, Runyi
Shift-Variance and Cyclostationarity of Linear Periodically Shift-Variant Systems

Zayed, Ahmed
Fractional Prolate Spheroidal Wave Functions

Zheludev, Valery
Spline-based frames for image restoration

Zheng, Fang
The Design of Non-redundant Directional Wavelet Filter Bank Using 1-D Neville Filters

Zhu, Lei
STORM by compressed sensing

Zibulevsky, Michael
Sub-Wavelength Coherent Diffractive Imaging based on Sparsity

Zörlein, Henning
On the Noise-Resilience of OMP with BASC-Based Low Coherence Sensing Matrices