# An Extension of Averaged-Operator-Based Algorithms

Miguel Simões
*Instituto de Telecomunicações*
*Instituto Superior Técnico, Univ. Lisboa*
Lisbon, Portugal
miguel.simoes@lx.it.pt

José Bioucas-Dias
*Instituto de Telecomunicações*
*Instituto Superior Técnico, Univ. Lisboa*
Lisbon, Portugal
bioucas@lx.it.pt

Luis B. Almeida
*Instituto de Telecomunicações*
*Instituto Superior Técnico, Univ. Lisboa*
Lisbon, Portugal
luis.almeida@lx.it.pt

***Abstract*—Many of the algorithms used to solve minimization problems with sparsity-inducing regularizers are generic in the sense that they do not take into account the sparsity of the solution in any particular way. However, algorithms known as semismooth Newton are able to take advantage of this sparsity to accelerate their convergence. We show how to extend these algorithms in different directions, and study the convergence of the resulting algorithms by showing that they are a particular case of an extension of the well-known Krasnosel'skiĭ–Mann scheme.**

***Index Terms*—Convex nonsmooth optimization, primal–dual optimization, semismooth Newton method, forward–backward method, variable metric**

## I. Introduction

### A. Background

The objective functions of many signal-processing problems can be formulated as sums of two proper lower-semicontinuous convex functions: one that is smooth, $f : \mathbb{R}^n \to ]-\infty, +\infty]$, and another one that need not be smooth, $g : \mathbb{R}^n \to ]-\infty, +\infty]$. The resulting problem is

$$\underset{\mathbf{x}\in\mathbb{R}^n}{\text{minimize}} \quad f(\mathbf{x}) + g(\mathbf{x}). \tag{1}$$

Such problems are typically large-scale and can be solved by using splitting methods, which convert (1) into a sequence of separable subproblems. The (relaxed) *forward–backward* method [1], [2] is an example of such methods. Its iterations can be broken into a gradient (forward) step on $f$ and a proximal (backward) step on $g$, performed consecutively—see Algorithm 1, where $\text{prox}_{\tau g}$ denotes the *proximal operator* of function $g$, i.e., $\text{prox}_{\tau g}(\mathbf{x}) \triangleq \arg\min_{\mathbf{u}\in\mathbb{R}^n} \left\{ g(\mathbf{u}) + \frac{1}{2\tau}\|\mathbf{x}-\mathbf{u}\|^2 \right\}$ [3].

When analyzing the properties of many of these and other algorithms, it can be advantageous to use the theory of monotone operators [4]. Let $2^{\mathbb{R}^n}$ denote the power set of $\mathbb{R}^n$. A set-valued operator $A : \mathbb{R}^n \to 2^{\mathbb{R}^n}$ is said to be *monotone* if $\langle \mathbf{u}-\mathbf{v}, \mathbf{x}-\mathbf{y}\rangle \geq 0$ for all $(\mathbf{x},\mathbf{u}) \in \text{gra}\, A$ and $(\mathbf{y},\mathbf{v}) \in \text{gra}\, A$, where gra $A$ denotes the graph of A, and it is said to be *maximally monotone* if there exists no other monotone operator whose graph properly contains gra $A$. Monotone operators are connected to optimization problems as follows. Take, for example, (1). According to Fermat's rule, its solutions should satisfy the inclusion $0 \in \nabla f(\mathbf{x}) + \partial g(\mathbf{x})$, where the set-valued operator $\partial g : \mathbb{R}^n \to 2^{\mathbb{R}^n} : \mathbf{x} \to \partial g(\mathbf{x})$ denotes the *subdifferential* of $g$ (in the sense of Moreau and Rockafellar [5, Chapter 23]). The operators $\nabla f$ and $\partial g$ are examples of maximally-monotone operators [6, Theorem 20.40]. Problem (1) can be seen as a particular case of the problem of finding a zero of the sum of two monotone operators $A$ and $C$, i.e.,

$$\text{find } \mathbf{x} \in \mathbb{R}^n \quad \text{such that } 0 \in A(\mathbf{x}) + C(\mathbf{x}), \tag{2}$$

if one makes $A = \partial g$ and $C = \nabla f$. Problem (2) may be solved using a generalized version of Algorithm 1, in which Line 5 is replaced with

$$\mathbf{x}^{k+1} \leftarrow \mathbf{x}^k + \lambda^k \left( J_{\tau A}\left(\mathbf{x}^k - \tau C\left(\mathbf{x}^k\right)\right) - \mathbf{x}^k \right), \tag{3}$$

where $J_{\tau A} \triangleq (\text{Id} + \tau A)^{-1}$ is the *resolvent* of operator $A$ and Id denotes the *identity* operator. Note that $J_{\tau\partial g} = \text{prox}_{\tau g}$ [6, Example 23.3].

Problem (2) can alternatively be written as the problem of finding a fixed point of the operator $R \triangleq J_{\tau A} \circ (\text{Id} - \tau C)$:

$$\text{find } \mathbf{x} \in \mathbb{R}^n \quad \text{such that } R(\mathbf{x}) = \mathbf{x}. \tag{4}$$

In general, the solutions of a convex optimization problem correspond to the fixed points of a certain operator, and an iterative optimization algorithm corresponds to a fixed-point method. We can rewrite (3) as

$$\mathbf{x}^{k+1} \leftarrow T_{\lambda^k}\left(\mathbf{x}^k\right) \triangleq \mathbf{x}^k + \lambda^k (R\left(\mathbf{x}^k\right) - \mathbf{x}^k). \tag{5}$$

We say that an operator $R : \mathbb{R}^n \to \mathbb{R}^n$ is nonexpansive if $\|\mathbf{u}-\mathbf{v}\| \leq \|\mathbf{x}-\mathbf{y}\|$ for all $(\mathbf{x},\mathbf{u}) \in \text{gra}\, R$ and $(\mathbf{y},\mathbf{v}) \in \text{gra}\, R$. Let $R$ be a generic nonexpansive operator and let $\lambda \in ]0,1[$. Then

**Algorithm 1:** Relaxed forward–backward method.

1 Choose $\mathbf{x}^0 \in \mathbb{R}^n,\ \tau > 0$;
2 $k \leftarrow 1$;
3 **while** *stopping criterion is not satisfied* **do**
4 Choose $\lambda^k > 0$;
5 $\mathbf{x}^{k+1} \leftarrow \mathbf{x}^k + \lambda^k \left(\text{prox}_{\tau g}\left(\mathbf{x}^k - \tau\nabla f\left(\mathbf{x}^k\right)\right) - \mathbf{x}^k\right)$;
6 $k \leftarrow k+1$;
7 **end**

This work was supported by the Fundação para a Ciência e Tecnologia within the Portuguese Ministry for Science, Technology and Higher Education under Project UID/EEA/50008/2013 and Grant BPD/N.º 134 - 16/10/2017.

ISBN 978-90-827970-1-5 

the operator $T \triangleq (\mathrm{Id} - \lambda) + \lambda R$ is said to be $\lambda$-*averaged*. It obeys the following contractive property [6, Proposition 4.25]:

$$\begin{aligned}&\|T(\mathbf{x}) - T(\mathbf{y})\|^2 \\ &\leq \|\mathbf{x}-\mathbf{y}\|^2 - \frac{1-\lambda}{\lambda}\|(\mathrm{Id}-T)(\mathbf{x}) - (\mathrm{Id}-T)(\mathbf{y})\|^2 \quad (6)\end{aligned}$$

for all $\mathbf{x}$, $\mathbf{y} \in \mathbb{R}^n$. In particular, when $\lambda = 1/2$, $T$ is said to be *firmly nonexpansive*. The resolvents of maximally-monotone operators are firmly-nonexpansive [6, Corollary 23.8]. Iteration (5) is known as the *Krasnosel'skiĭ–Mann* scheme and is the basis of not only the forward–backward method but also other optimization algorithms, such as the Douglas–Rachford one [4], [6]. It can be shown that, under certain conditions, the Krasnosel'skiĭ–Mann scheme converges to Fix $R$, where Fix $R$ denotes the set of fixed points of $R$.

The convergence rate of the forward–backward method (Algorithm 1) can be shown to be sublinear, or, under certain assumptions, to be linear. This rate can often be improved by incorporating *second-order* information about $f$ if this function is twice-differentiable. The local convergence rate of second-order methods is superlinear or even quadratic. As an example, consider the second-order version of Algorithm 1, which is given by replacing Line 5 with the iteration $\mathbf{x}^{k+1} \leftarrow \mathbf{x}^k + \lambda^k\left(\mathrm{prox}_g^{\mathbf{B}^k}\left(\mathbf{x}^k - \left[\mathbf{B}^k\right]^{-1}\nabla f\left(\mathbf{x}^k\right)\right) - \mathbf{x}^k\right)$ [7]–[9], where $\mathbf{B}^k$ is a positive-definite (PD) matrix (the Hessian of $f$ or an approximation of it) and $\mathrm{prox}_g^{\mathbf{B}^k}$ denotes the proximal operator of $g$ relative to the norm $\|\cdot\|_{\mathbf{B}^k}^2$, i.e., $\mathrm{prox}_g^{\mathbf{B}^k}(\mathbf{x}) \triangleq \arg\min_{\mathbf{u}\in\mathbb{R}^n}\left\{g(\mathbf{u}) + \frac{1}{2}\|\mathbf{x}-\mathbf{u}\|_{\mathbf{B}^k}^2\right\}$. More generally, and from an operator-centric perspective, by using second-order methods such as these, one is actually solving a left-preconditioned version of (2), in the sense that instead of directly tackling that problem we are considering problems that share the same set of solutions but may be more convenient to solve:

$$\text{find } \mathbf{x} \in \mathbb{R}^n \quad \text{such that } 0 \in \mathbf{U}A(\mathbf{x}) + \mathbf{U}C(\mathbf{x}), \quad (7)$$

where $\mathbf{U}$ is a PD operator. In what follows, we denote positive definiteness by $\mathbf{U} \succ 0$ and positive semidefiniteness by $\mathbf{U} \succeq 0$.

### B. Contributions

The basis of this work is the study of the following alternative scheme to (5):

$$\mathbf{x}^{k+1} = T_{\mathbf{\Lambda}^k}\left(\mathbf{x}^k\right) \triangleq \mathbf{x}^k + \mathbf{\Lambda}^k\left(R\left(\mathbf{x}^k\right) - \mathbf{x}^k\right), \quad (8)$$

where, for every $k$, $\mathbf{\Lambda}^k$ is a linear operator such that $\mathrm{Id} \succ \mathbf{\Lambda}^k \succ 0$. For convenience, we call the operators $T_{\mathbf{\Lambda}^k}$, *operator-weighted averaged operators*. It is clear that if, for all $k$, we make $\mathbf{\Lambda}^k = \lambda^k\mathrm{Id}$, we recover (5).

Iteration (8) can be interpreted in different ways. For example, if $\mathbf{\Lambda}^k$ is fixed, i.e., if, for all $k$, $\mathbf{\Lambda}^k = \mathbf{\Lambda}$, where $\mathbf{\Lambda} \succ 0$, that iteration can also be seen as a left-preconditioning scheme to solve (4):

$$\text{find } \mathbf{x} \in \mathbb{R}^n \quad \text{such that } \mathbf{\Lambda}R(\mathbf{x}) = \mathbf{\Lambda}\mathbf{x}. \quad (9)$$

### C. Notation and outline

A detailed account of the notions listed in this section can be found in the work of Bauschke and Combettes [6]. We denote the *scalar product* of a Hilbert space by $\langle\cdot,\cdot\rangle$ and the associated *norm* by $\|\cdot\|$. The *range* of an operator $A$ is denoted by $\mathrm{ran}\, A$, and the *adjoint* of $A$ by $A^*$. We say that an operator $A : \mathbb{R}^n \to \mathbb{R}^n$ is *Lipschitz continuous* with constant $L > 0$ if $\|\mathbf{u}-\mathbf{v}\| \leq L\|\mathbf{x}-\mathbf{y}\|$, for all $(\mathbf{x},\mathbf{u}) \in \mathrm{gra}\, A$ and $(\mathbf{y},\mathbf{v}) \in \mathrm{gra}\, A$. Additionally, let $\Gamma_0(\mathbb{R}^n)$ denote the class of all proper lower-semicontinuous convex functions from $\mathbb{R}^n$ to $]-\infty,+\infty]$. Given two functions $f \in \Gamma_0(\mathbb{R}^n)$ and $g \in \Gamma_0(\mathbb{R}^n)$, their *infimal convolution* is denoted by $f \star_{\inf} g$. The Legendre–Fenchel *conjugate* of a function $f$ is denoted by $f^*$. The *indicator function* of a set $C \in \mathbb{R}^n$ is defined as $\delta_C(\mathbf{x}) \triangleq 0$ if $\mathbf{x} \in C$, $\delta_C(\mathbf{x}) \triangleq +\infty$ otherwise. We use the notation $\{\mathbf{x}^k\}$ as a shorthand for representing the sequence $\{\mathbf{x}^k\}_{k=1}^{+\infty}$. The space of *absolutely-summable sequences* in $\mathbb{R}$ is denoted by $\ell^1(\mathbb{N})$; the set of summable sequences in $[0,+\infty[$ is denoted by $\ell_+^1(\mathbb{N})$. Bold lowercase letters denote vectors and bold uppercase letters denote matrices. $[\mathbf{a}]_i$ denotes the $i$-th element of a vector $\mathbf{a}$, $[\mathbf{A}]_{:j}$ denotes the $j$-th column of a matrix $\mathbf{A}$, and $[\mathbf{A}]_{ij}$ denotes the element in the $i$-th row and $j$-th column of a matrix $\mathbf{A}$. $\mathbf{0}$ denotes a *zero* vector or matrix of appropriate size. The maximum and signum operators are denoted by $\max(\cdot)$ and $\mathrm{sgn}(\cdot)$, respectively.

The structure of this work is as follows. In Section II, we briefly discuss a class of algorithms known as semismooth Newton methods. In Section III, we study the scheme given by (8), and show how it can be used to solve a primal–dual problem first studied by Combettes and Pesquet [10]. In Section IV, we present a simple application of the proposed method to solve an inverse problem. Section V concludes. Due to space constraints, we omit the proofs of the results discussed in Section III; these proofs can be consulted elsewhere [11, Chapter 5].

## II. Semismooth Newton methods

*Semismooth Newton* methods were originally developed with the goal of using Newton-like methods to minimize certain nonsmooth functions at a superlinear convergence rate. To illustrate why these methods may be useful when solving problems of the form of (1), consider, as an example, that $f = \|\mathbf{y}-\mathbf{H}\cdot\|^2$, and $g = \mu\|\cdot\|_1$, where $\mathbf{y} \in \mathbb{R}^m$, $\mathbf{H} \in \mathbb{R}^{m\times n}$, and $\mu > 0$. For problems such as these, it was shown by Hintermüller [12] that some semismooth Newton methods are equivalent to some active-set methods. As we discuss in Section IV, the fact that these methods can be written as active-set ones allows for significant time savings when solving certain problems, namely the ones involving sparsity-inducing regularizers, as is the case of the $\ell_1$ norm.

Let $G : \mathbb{R}^n \to \mathbb{R}^n$ be an operator such that $G : \mathbf{x} \to \mathbf{x} - \mathrm{prox}_{\mu\|\mathbf{x}\|_1}(\mathbf{x} - 2\mu\mathbf{H}^*(\mathbf{H}\mathbf{x}-\mathbf{y}))$. The solution of the problem under consideration should satisfy the nonlinear equation $G(\mathbf{x}) = \mathbf{0}$, which is nonsmooth, since $\mathrm{prox}_{\mu\|\cdot\|_1}$ is not everywhere differentiable. There are, however, generalizations of the concept of differentiability that are applicable

ISBN 978-90-827970-1-5 © EURASIP 2018

to an operator such as $G$. One of them is the B(ouligand)-differential [13, Definition 4.6.2], which is defined as follows. Suppose that a generic operator $G : D \subset \mathbb{R}^n \to \mathbb{R}^m$ is locally Lipschitz, where $D$ is an open subset. Then by Rademacher's theorem, $G$ is differentiable almost everywhere in $D$. Let $C$ denote the subset of $\mathbb{R}^n$ consisting of the points where $G$ is differentiable (in the sense of Fréchet [6, Definition 2.45]). The B-differential of $G$ at $\mathbf{x}$ is $\partial_B G(\mathbf{x}) \triangleq \left\{\lim_{\mathbf{x}^j \to \mathbf{x}} \nabla G\left(\mathbf{x}^j\right)\right\}$, where $\{\mathbf{x}^j\}$ is a sequence such that $\mathbf{x}^j \in C$ for all $j$ and $\nabla G(\mathbf{x}^j)$ denotes the Jacobian of $G$ at $\mathbf{x}^j$.

The B-differential of an operator at a given point may not be unique: for example, take $\operatorname{prox}_{\mu\|\cdot\|_1}(\mathbf{x})$, which can be evaluated element-wise by computing $\max\left\{\left|[\mathbf{x}]_i\right| - \mu, 0\right\} \circ \operatorname{sgn}([\mathbf{x}]_i)$ for $i \in \{1, \cdots, n\}$. A possible $\mathbf{H} \in \partial_B \operatorname{prox}_{\mu\|\cdot\|_1}(\mathbf{x})$ is a binary diagonal matrix defined as [14, Proposition 3.3]

$$[\mathbf{H}]_{ii} = \begin{cases} 1 & \text{if } \left|[\mathbf{x}]_i\right| > \mu, \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

This generalization of the concept of differentiability can also be used to formulate the so-called semismooth Newton method, which is characterized by the iteration $\mathbf{x}^{k+1} \leftarrow \mathbf{x}^k - [\mathbf{H}^k]^{-1} G(\mathbf{x}^k)$, where $\mathbf{H}^k \in \partial_B G(\mathbf{x}^k)$. It can be shown that this method locally converges superlinearly for operators known as semismooth [15]. Let $\mathbf{x} \in D$ and $\mathbf{d} \in \mathbb{R}^n$; semismooth operators are operators that are directionally differentiable at a neighborhood of $\mathbf{x}$ and that, for any $\mathbf{H} \in \partial_B G(\mathbf{x} + \mathbf{d})$, satisfy the condition $\mathbf{H}\mathbf{d} - G'(\mathbf{x}; \mathbf{d}) = o(\|\mathbf{d}\|)$ for $\mathbf{d} \to \mathbf{0}$, where $G'(\mathbf{x}; \mathbf{d})$ denotes the directional derivative of $G$ at $\mathbf{x}$ along $\mathbf{d}$. Examples of semismooth functions are the Euclidean norm and piecewise-differentiable functions [16, Chapter 2], $\operatorname{prox}_{\mu\|\cdot\|_1}(\mathbf{x})$ being an example of the latter. Note that the semismooth Newton method is a particular case of (8), although we impose that $\mathrm{Id} \succ \mathbf{\Lambda}^k \succ 0$ in the latter equation, which is not necessarily true for this method.

## III. An extension of averaged-operator-based algorithms

In this section, we define operator-weighted averaged operators, and show that they have a contractive property. We also study the asymptotic behavior of fixed-point iterations of these operators. Such iterations can be seen as an extension of the Krasnosel'skiĭ–Mann scheme [cf. (5)]. We base our analysis on the fact that these iterations produce a sequence that is variable-metric Fejér monotone [17], [18]. We then present an algorithm that uses operator-weighted averaged operators, and that solves a primal–dual problem that encapsulates many problem formulations [10], [18].

### A. An extension of the Krasnosel'skiĭ–Mann scheme

**Definition III.1** (Operator-weighted averaged operators)**.** Let $D$ be a nonempty subset of $\mathbb{R}^n$, let $\epsilon \in ]0, 1[$, and let $\mathbf{\Lambda}$ be an operator in $\mathbb{R}^n$ such that

$$\mu\mathrm{Id} \succeq \mathbf{\Lambda} \succeq \alpha\mathrm{Id}, \quad \text{where } \mu,\, \alpha \in [\epsilon, 1 - \epsilon]. \tag{11}$$

We say that an operator $T_{\mathbf{\Lambda}} : D \to \mathbb{R}^n$ is an operator-weighted averaged operator if there exists a nonexpansive operator $R : D \to \mathbb{R}^n$ such that

$$T_{\mathbf{\Lambda}} \triangleq (\mathrm{Id} - \mathbf{\Lambda}) + \mathbf{\Lambda} R. \tag{12}$$

We have proved the following results:

**Proposition III.2.** *Let $D$ be a nonempty subset of $\mathbb{R}^n$, let $\epsilon \in ]0, 1[$, let $\mathbf{\Lambda}$ be an operator in $\mathbb{R}^n$ satisfying* (11)*, let $R : D \to \mathbb{R}^n$ be a nonexpansive operator, and let $T_{\mathbf{\Lambda}} : D \to \mathbb{R}^n$ be an operator as defined in* (12)*. Then the operator $T_{\mathbf{\Lambda}}$ is $\mu$-averaged in the metric induced by $\mathbf{\Lambda}^{-1}$. In other words, the operator $T_{\mathbf{\Lambda}}$ verifies*

$$\begin{aligned} &\|T_{\mathbf{\Lambda}}(\mathbf{x}) - T_{\mathbf{\Lambda}}(\mathbf{y})\|^2_{\mathbf{\Lambda}^{-1}} \\ &\leq \|\mathbf{x} - \mathbf{y}\|^2_{\mathbf{\Lambda}^{-1}} - \frac{1-\mu}{\mu} \|(\mathrm{Id} - T_{\mathbf{\Lambda}})(\mathbf{x}) - (\mathrm{Id} - T_{\mathbf{\Lambda}})(\mathbf{y})\|^2_{\mathbf{\Lambda}^{-1}} \end{aligned}$$

*for all $\mathbf{x}$, $\mathbf{y} \in D$.*

**Theorem III.3.** *Let $D$ be a nonempty closed convex subset of $\mathbb{R}^n$, let $\epsilon \in ]0, 1[$, let $\{\eta^k\} \in \ell^1_+(\mathbb{N})$, let $\{\mathbf{\Lambda}^k\}$ be a sequence of PD operators in $\mathbb{R}^{n\times n}$ such that, for all $k \in \mathbb{N}$,*

$$\begin{cases} \mu^k \mathrm{Id} \succeq \mathbf{\Lambda}^k \succeq \alpha^k \mathrm{Id}, \\ \mu^k,\, \alpha^k \in [\epsilon, 1 - \epsilon], \\ \left(1 + \eta^k\right) \mathbf{\Lambda}^{k+1} \succeq \mathbf{\Lambda}^k, \end{cases} \tag{13}$$

*and let $R : D \to D$ be a nonexpansive operator such that* Fix $R \neq \emptyset$*. Additionally, let $\mathbf{x}^0 \in D$ and let, for all $k$, $\{\mathbf{x}^k\}$ be a sequence generated by* (8)*. Then $\{\mathbf{x}^k\}$ converges to a point in* Fix $R$*.*

### B. Primal–dual optimization algorithms

Combettes and Pesquet studied a primal–dual problem that generalizes many problems [10, Problem 4.1]. By being able to devise an algorithm to solve this problem, we are effectively tackling a large number of problems simultaneously (problem (1) is one of these). Let $m$, $n$, and $N$ be strictly-positive integers, let $g \in \Gamma_0(\mathbb{R}^n)$, let $\mu \in ]0, +\infty[$, let $f : \mathbb{R}^n \to ]-\infty, +\infty]$ be convex and differentiable with a $\mu^{-1}$-Lipschitzian gradient, and let $\mathbf{z} \in \mathbb{R}^n$. For every $j \in \{1, \ldots, N\}$, let $\mathbf{r}_j \in \mathbb{R}^{m_j}$, let $h_j \in \Gamma_0(\mathbb{R}^{m_j})$, let $\nu_j \in ]0, +\infty[$, let $l_j \in \Gamma_0(\mathbb{R}^{m_j})$ be $\nu_j$-strongly convex,[1] let $\mathbf{L}_j \in \mathbb{R}^{m_j \times n}$ such that $\mathbf{L}_j \neq 0$, and let $\omega_j$ be real numbers in $]0, 1]$ such that $\sum_{j=1}^N \omega_j = 1$. The problem is as follows:

**Problem III.4.** Solve the primal minimization problem,

$$\underset{\mathbf{x}\in\mathbb{R}^n}{\operatorname{minimize}}\; g(\mathbf{x}) + \sum_{j=1}^N \omega_j \left(h_j \mathbin{\square_{\inf}} l_j\right)(\mathbf{L}_j\mathbf{x} - \mathbf{r}_j) + f(\mathbf{x}) - \langle \mathbf{x}, \mathbf{z}\rangle\,,$$

together with its corresponding dual minimization problem,

$$\begin{aligned} \underset{\mathbf{d}_1\in\mathbb{R}^{m_1},\cdots,\mathbf{d}_j\in\mathbb{R}^{m_j}}{\operatorname{minimize}}\; &\left(g^* \mathbin{\square_{\inf}} h^*\right)\left(\mathbf{z} - \sum_{j=1}^N \omega_j \mathbf{L}_j^* \mathbf{d}_j\right) \\ &+ \sum_{j=1}^N \omega_j \left(h_j^*(\mathbf{d}_j) + l_j^*(\mathbf{d}_j) + \langle \mathbf{d}_j, \mathbf{r}_j\rangle\right). \end{aligned}$$

The sets of solutions to these primal and dual problems are denoted by $P$ and $D$, respectively.

[1] A function $l$ is said to be *$\nu$-strongly convex* if $l - \frac{\nu}{2}\langle \mathbf{x}, \mathbf{x}\rangle$ is convex, for some $\nu > 0$.

ISBN 978-90-827970-1-5 © EURASIP 2018

Consider Algorithm 2 to solve Problem III.4. In what follows, for all $j$, $\{\mathbf{U}^k\}$, $\{\mathbf{\Lambda}^k\}$, $\{\mathbf{U}_j^k\}$, $\{\mathbf{\Lambda}_j^k\}$ are sequences of linear operators, and $\{\mathbf{a}^k\}$, $\{\mathbf{b}_j^k\}$, $\{\mathbf{c}^k\}$, $\{\mathbf{e}_j^k\}$ are absolutely-summable sequences that can be used to model errors. Algorithm 2 is an extension of [18, Example 6.4]. The

**Algorithm 2:** An application of (8) to solve Problem III.4.

1 Choose $\mathbf{x}^0 \in \mathbb{R}^n$ and $\mathbf{d}_1^0 \in \mathbb{R}^{m_1}, \cdots, \mathbf{d}_j^0 \in \mathbb{R}^{m_j}$;
2 $k \leftarrow 1$;
3 **while** *stopping criterion is not satisfied* **do**
4  **for** $j = 1, \ldots, N$ **do**
5  Choose $\mathbf{U}_j^k$, $\mathbf{\Lambda}_j^k \succ 0$ s.t. $\mathbf{\Lambda}_j^k \prec \mathrm{Id}$;
6  $\mathbf{q}_j^k = \mathrm{prox}_{h_j^*}^{(\mathbf{U}_j^k)^{-1}}\big(\mathbf{d}_j^k + \mathbf{U}_j^k\big(\mathbf{L}_j\mathbf{x}^k - \nabla l_j^*\left(\mathbf{d}^k\right) - \mathbf{e}_j^k - \mathbf{r}_j\big)\big) + \mathbf{b}_j^k$;
7  $\mathbf{y}_j^k = 2\mathbf{q}_j^k - \mathbf{d}_j^k$;
8  $\mathbf{d}_j^{k+1} = \mathbf{d}_j^k + \mathbf{\Lambda}_j^k\left(\mathbf{q}_j^k - \mathbf{d}_j^k\right)$;
9  **end**
10  Choose $\mathbf{U}^k$, $\mathbf{\Lambda}^k \succ 0$ s.t. $\mathbf{\Lambda}^k \prec \mathrm{Id}$;
11  $\mathbf{p}^k = \mathrm{prox}_g^{(\mathbf{U}^k)^{-1}}\Big(\mathbf{x}^k - \mathbf{U}^k\big(\sum_{j=1}^N \omega_j \mathbf{L}_j^* \mathbf{y}_j^k + \nabla f\left(\mathbf{x}^k\right) + \mathbf{c}^k - \mathbf{z}\big)\Big) + \mathbf{a}^k$;
12  $\mathbf{x}^{k+1} = \mathbf{x}^k + \mathbf{\Lambda}^k\left(\mathbf{p}^k - \mathbf{x}^k\right)$;
13  $k \leftarrow k + 1$;
14 **end**

following corollary establishes some convergence properties of Algorithm 2.

**Corollary III.5.** *Suppose that*

$$\mathbf{z} \in \mathrm{ran}\left(\partial g + \sum_{j=1}^N \omega_j \mathbf{L}_j^*\left(\partial h_j \star_{inf} \partial l_j\right)\left(\mathbf{L}_j \cdot - \mathbf{r}_j\right) + \nabla f\right)$$

*and set* $\beta \triangleq \min\{\mu, \nu_1, \ldots, \nu_N\}$. *Let* $\{\mathbf{U}^k\}$ *be a sequence of PD operators in* $\mathbb{R}^{n\times n}$ *and, for every* $j \in \{1, \ldots, N\}$, *let* $\{\mathbf{U}_j^k\}$ *be a sequence of PD operators in* $\mathbb{R}^{m_j \times m_j}$ *such that, for all* $k \in \mathbb{N}$,

$$\begin{cases} \mu_{\mathbf{U}}\mathrm{Id} \succeq \mathbf{U}^k \succeq \alpha_{\mathbf{U}}\mathrm{Id}, \\ \mu_{\mathbf{U}}\mathrm{Id} \succeq \mathbf{U}_j^k \succeq \alpha_{\mathbf{U}}\mathrm{Id}, \\ \mu_{\mathbf{U}},\, \alpha_{\mathbf{U}} \in \left]0, +\infty\right[, \end{cases} \tag{14}$$

*let* $\epsilon \in \left]0, \min\{1, \beta\}\right[$, *let* $\{\mathbf{\Lambda}^k\}$ *be a sequence of PD operators in* $\mathbb{R}^{n\times n}$, *and let* $\{\mathbf{\Lambda}_j^k\}$ *be a sequence of PD operators in* $\mathbb{R}^{m_j\times m_j}$ *such that, for all* $k$,

$$\begin{cases} \mathbf{\Lambda}^k\mathbf{U}^k = \mathbf{U}^k\mathbf{\Lambda}^k, \\ \mathbf{\Lambda}_j^k\mathbf{U}_j^k = \mathbf{U}_j^k\mathbf{\Lambda}_j^k, \\ \mu\mathrm{Id} \succeq \mathbf{\Lambda}^k \succeq \alpha\mathrm{Id}, \\ \mu\mathrm{Id} \succeq \mathbf{\Lambda}_j^k \succeq \alpha\mathrm{Id}, \\ \mu,\, \alpha \in [\epsilon, 1], \end{cases} \quad \textit{and} \quad \begin{cases} \mathbf{\Lambda}^{k+1}\mathbf{U}^{k+1} \succeq \mathbf{\Lambda}^k\mathbf{U}^k, \\ \mathbf{\Lambda}_j^{k+1}\mathbf{U}_j^{k+1} \succeq \mathbf{\Lambda}_j^k\mathbf{U}_j^k. \end{cases} \tag{15}$$

*Let, for all* $j$, $\{\mathbf{a}^k\}$, $\{\mathbf{b}^k\}$, $\{\mathbf{c}_j^k\}$, $\{\mathbf{e}_j^k\} \in \ell^1(\mathbb{N})$. *For every* $k$, *set* $\delta^k \triangleq \left(\sum_{j=1}^N \omega_j \left\|\sqrt{\mathbf{U}_j^k}\mathbf{L}_j\sqrt{\mathbf{U}^k}\right\|^2\right)^{-\frac{1}{2}} - 1$ *and suppose that* $\xi^k \triangleq \frac{\delta^k}{(1+\delta^k)\mu_{\mathbf{U}}} \geq \frac{1}{2\beta-\epsilon}$.

*Let* $\{\mathbf{x}^k\}$ *be a sequence generated by Algorithm 2. Then* $\mathbf{x}^k$ *converges to a point in* $P$ *and* $\left(\mathbf{d}_1^k, \ldots, \mathbf{d}_N^k\right)$ *converges to a point in* $D$.

## IV. Experiment

In this section, we give a practical example of a simple problem that can be solved via Algorithm 2. Consider the constrained problem

$$\underset{\mathbf{x}\in[c,d]^n}{\text{minimize}} \quad \|\mathbf{b} - \mathbf{H}\mathbf{x}\|^2 + \mu\|\mathbf{x}\|_1, \tag{16}$$

where $\mathbf{b} \in \mathbb{R}^n$, $c \in \mathbb{R}$, $d \in \mathbb{R}$, $\mu > 0$, $\mathbf{H} = {}^1\!/_n\widehat{\mathbf{H}}$, and $\widehat{\mathbf{H}} \in \mathbb{R}^{n\times n}$ is a lower-triangular matrix of ones. Griesse and Lorenz studied a non-constrained, and therefore simpler, version of this problem in the context of inverse integration [14, Section 4.1]. Problem (16) can be solved via Algorithm 2 if we let $\gamma > 0$, $\tau > 0$ and make $m = n$, $N = 1$, $\mathbf{L}_1 = \mathrm{Id}$, $\mathbf{r}_1 = \mathbf{0}$, $\mathbf{z} = \mathbf{0}$, and, for all $k$, $\mathbf{U}_1^k = \gamma\mathrm{Id}$, $\mathbf{U}^k = \tau\mathrm{Id}$, $\mathbf{e}_1^k = \mathbf{0}$, $\mathbf{b}_1^k = \mathbf{0}$, $\mathbf{\Lambda}_1^k = \mathrm{Id}$, $\mathbf{c}^k = \mathbf{0}$, $\mathbf{a}^k = \mathbf{0}$, $f = \|\mathbf{b} - \mathbf{H}\cdot\|^2$, $g = \mu\|\cdot\|_1$, $h = \delta_{[c,d]^n}(\cdot)$, $l_1 : \mathbf{u} \to 0$ if $\mathbf{u} = 0$, $l_1 : \mathbf{u} \to +\infty$ otherwise.

If we take $\mathbf{\Lambda}^k$ to be a sequence of scalars, we recover a version of [18, Example 6.4]. However, inspired by the fast convergence properties of the methods discussed in Section II and following a similar reasoning to [14, Proposition 3.7], we consider the B-differential for the operator $\mathrm{prox}_{\mu\|\cdot\|_1}$ given in (10) and take $\mathbf{\Lambda}^k$ to be the inverse of

$$\left(\mathbf{P}^k\right)^{-1}\begin{bmatrix}\tau[\mathbf{H}]_{:I^k}^*[\mathbf{H}]_{:I^k} & \tau[\mathbf{H}]_{:I^k}^*[\mathbf{H}]_{:A^k} \\ \mathbf{0} & \mathrm{Id}\end{bmatrix}\mathbf{P}^k,$$

where

$$A^k \triangleq \{i \in \mathbb{N} : \left|\left[\mathbf{x}^k - 2\tau\left(\mathbf{H}^*\left(\mathbf{H}\mathbf{x}^k - \mathbf{b}\right) + \mathbf{y}_1^k\right)\right]_i\right| \leq \tau\mu\},$$
$$I^k \triangleq \{i \in \mathbb{N} : \left|\left[\mathbf{x}^k - 2\tau\left(\mathbf{H}^*\left(\mathbf{H}\mathbf{x}^k - \mathbf{b}\right) + \mathbf{y}_1^k\right)\right]_i\right| > \tau\mu\},$$

and $\{\mathbf{P}^k\}$ is a sequence of appropriate permutation matrices such that, given a vector $\mathbf{x}$, the first elements of the vector $\mathbf{P}^k\mathbf{x}$ correspond to the indices in $I^k$ and the last elements to the indices in $A^k$, for all $k$. By again following a similar reasoning to the one of [14, Section 3.3], it can be shown that Line 12 of Algorithm 2 can be rewritten in such a way that this algorithm is easily seen to be equivalent to an active-set method. In fact, that line is given by

$$\mathbf{x}^{k+1} \leftarrow \left(\mathbf{P}^k\right)^{-1}\begin{bmatrix}\left([\mathbf{H}]_{:I^k}^*[\mathbf{H}]_{:I^k}\right)^{-1}\left[\mathbf{H}^*\mathbf{b} - \mathbf{y}_1^k + \tau\mathbf{e}_{\pm}^k\right]_{I^k} \\ \mathbf{0}\end{bmatrix},$$

where $\mathbf{e}_{\pm}^k \triangleq \mathrm{sgn}\left[\mathbf{x}^k - 2\tau\left(\mathbf{H}^*\left(\mathbf{H}\mathbf{x}^k - \mathbf{b}\right) + \mathbf{y}_1^k\right)\right]$, for every $k$. The dimension of the problem to solve at each iteration is given by the cardinality of the set $I^k$. Naturally, the sparser the solution is estimated to be, the smaller the dimension of this problem is. In contrast, methods such as the alternating-direction method of multipliers (ADMM) [19] require the solution of a problem involving the full matrix $\mathbf{H}^*\mathbf{H}$. This is the reason why semimooth Newton methods are able to achieve faster convergence rates in practice than others.

We simulate an example similar to the one studied by Griesse and Lorenz [14, Section 4.1] but consider the noise to be Gaussian with a signal-to-noise ratio (SNR) of 30 dB. We have set $\mu = 3 \times 10^{-3}$, $c = -80$, and $d = 52$. We

ISBN 978-90-827970-1-5 © EURASIP 2018

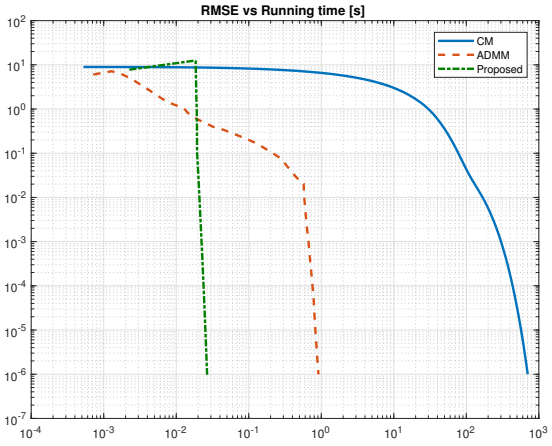


Figure 1. RMSE, as a function of time, between the estimates of each iteration and the representative solution, for the three methods.

compared Algorithm 2 (denoted in what follows as *Proposed*) with ADMM and with the method by Condat [20] (CM) to solve (16). We manually tuned the different parameters of the three methods in order to achieve the fastest convergence results in practice. We arbitrarily chose the result of ADMM after $10^7$ iterations as representative of the solution given by the three methods. Fig. 1 illustrates the behavior of the three methods by showing the root-mean-squared error (RMSE) between the estimates of each method and the representative solution, as a function of time. The three methods were initialized with the zero vector. The experiments were performed using MATLAB on an Intel Core i7 CPU running at 3.20 GHz, with 32 GB of RAM.

In this example, we did not enforce assumptions (15) but verified in practice that they were satisfied. However, in more complex examples, it may be necessary to devise a strategy that generates a sequence $\left\{\boldsymbol{\Lambda}^k\right\}$ satisfying these assumptions. This is akin to the necessity of devising globalization strategies in other Newton-like methods [13, Chapter 8].

### A. Appraisal

It is clear that, for this example, the proposed method has a much faster convergence than either CM or ADMM. This improvement in convergence is similar to the one observed in the methods discussed in Section II. In general, the sparser the solution is, the faster the method is as well. In order to benefit from this property, we must be able to solve the lower-dimensional linear system faster than the full system. This may not always be possible: for example, in problems that involve computations with the fast Fourier Transform (FFT) of a signal, we usually have only modest improvements in speed if we wish to compute only selected elements of the FFT.[2] However, for large-scale problems and for highly-sparse signals, methods known as sparse FFTs [21] may be useful. We verified in other experiments not detailed here that the proposed method has a comparable convergence speed to ADMM in problems whose solutions are not sparse or where we cannot take advantage of their sparsity.

[2]See `http://www.fftw.org/pruned.html` for details.

## V. Conclusions

In this work, we defined operator-weighted averaged operators, and showed that they can be used to construct a number of algorithms with good convergence properties. These algorithms have very broad applications, and seem to be particularly suitable to address problems with sparsity-inducing regularizers, as suggested by a simple experiment. Possible future directions to be explored are the possibility of relaxing the assumptions on $\boldsymbol{\Lambda}^k$, and the study of which problems are most suitable to be tackled by these methods.

## References

[1] M. Figueiredo and R. Nowak, "An EM algorithm for wavelet-based image restoration," *IEEE Trans. Image Process.*, vol. 12, no. 8, pp. 906–916, Aug 2003.

[2] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Comm. Pure Appl. Math.*, vol. 57, no. 11, pp. 1413–1457, 2004.

[3] J. Moreau, "Fonctions convexes duales et points proximaux dans un espace Hilbertien," *Comptes Rendus Acad. Sci.*, vol. A255, pp. 2897–2899, 1962.

[4] C. Byrne, "A unified treatment of some iterative algorithms in signal processing and image reconstruction," *Inverse Probl.*, vol. 20, no. 1, pp. 103–120, 2004.

[5] R. Rockafellar, *Convex Analysis*. New Jersey, USA: Princeton University Press, 1970.

[6] H. Bauschke and P. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. New York, NY, USA: Springer, 2011.

[7] M. Schmidt, D. Kim, and S. Sra, "Projected Newton-type methods in machine learning," in *Optimization for Machine Learning*. MIT Press, 2011, pp. 305–330.

[8] S. Becker and M. Fadili, "A quasi-Newton proximal splitting method," in *Proc. 25th Int. Conf. Neural Informat. Process. Systems*, Lake Tahoe, Nevada, 2012, pp. 2618–2626.

[9] J. Lee, Y. Sun, and M. Saunders, "Proximal Newton-type methods for minimizing composite functions," *SIAM J. Optim.*, vol. 24, no. 3, pp. 1420–1443, 2014.

[10] P. 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 Var. Anal.*, vol. 20, no. 2, pp. 307–330, 2011.

[11] M. Simões, *On some aspects of inverse problems in image processing*. Universidade de Lisboa, Instituto Superior Técnico, Portugal & Université Grenoble Alpes, France: PhD dissertation, 2017. [Online]. Available: `http://cascais.lx.it.pt/%7Emsimoes/dissertation/`

[12] M. Hintermüller, K. Ito, and K. Kunisch, "The primal–dual active set strategy as a semismooth Newton method," *SIAM J. Optim.*, vol. 13, no. 3, pp. 865–888, 2003.

[13] F. Facchinei and J.-S. Pang, *Finite-Dimensional Variational Inequalities and Complementarity Problems, Vols. I & II*. Springer-Verlag, 2003.

[14] R. Griesse and D. Lorenz, "A semismooth Newton method for Tikhonov functionals with sparsity constraints," *Inverse Probl.*, vol. 24, no. 3, p. 035007, 2008.

[15] L. Qi, "Convergence analysis of some algorithms for solving nonsmooth equations," *Math. Oper. Res.*, vol. 18, no. 1, pp. 227–244, 1993.

[16] M. Ulbrich, *Semismooth Newton Methods for Variational Inequalities and Constrained Optimization Problems in Function Spaces*. Philadelphia, PA: MOS-SIAM Ser. Optim., 2011.

[17] P. Combettes and B. Vũ, "Variable metric quasi-Fejér monotonicity," *Nonlinear Anal-Theor*, vol. 78, pp. 17–31, 2013.

[18] P. Combettes and B. Vũ, "Variable metric forward–backward splitting with applications to monotone inclusions in duality," *Optim.*, vol. 63, no. 9, pp. 1289–1318, 2014.

[19] M. Afonso, J. Bioucas-Dias, and M. Figueiredo, "Fast image recovery using variable splitting and constrained optimization," *IEEE Trans. Image Process.*, vol. 19, no. 9, pp. 2345–2356, Sept 2010.

[20] L. Condat, "A primal–dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms," *J. Optim. Theory Appl.*, vol. 158, no. 2, pp. 460–479, 2013.

[21] A. Gilbert, P. Indyk, M. Iwen, and L. Schmidt, "Recent developments in the sparse Fourier transform: A compressed Fourier transform for big data," *IEEE Signal Process. Mag.*, vol. 31, no. 5, pp. 91–100, Sept 2014.

ISBN 978-90-827970-1-5 © EURASIP 2018