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 ith 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 $\|\boldsymbol{u}\|_0 = \#\mathrm{supp}\,\boldsymbol{u}$, where $\#$ is the cardinality operator and $\mathrm{supp}\,\boldsymbol{u} = \{i : u_i \neq 0\} \subseteq [D]$. 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 $\mathcal{I} \subset \mathbb{R}$, $\mathrm{sign}\,\lambda$ 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

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

$$\min_{\boldsymbol{u} \in \mathbb{R}^N} \tfrac{1}{2}\|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{u}\|^2 \text{ s.t. } \|\boldsymbol{u}\|_0 \le K. \tag{1}$$

It proceeds by computing the following recursion

$$\boldsymbol{x}^{(n+1)} = \mathcal{H}_K\big[\boldsymbol{x}^{(n)} + \mu\boldsymbol{\Phi}^*(\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x}^{(n)})\big], \qquad \textbf{(IHT)}$$

where $\boldsymbol{x}^{(0)} = \boldsymbol{0}$, 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 $\mathcal{E}(\boldsymbol{u}) := \tfrac{1}{2}\|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{u}\|^2$ 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 $\boldsymbol{u} \in \Sigma_{3K}$, $(1 - \delta_{3K})\|\boldsymbol{u}\|^2 \le \|\boldsymbol{\Phi}\boldsymbol{u}\|^2 \le (1 + \delta_{3K})\|\boldsymbol{u}\|^2$, then, at iteration $n^* = \lceil \log_2 \|\boldsymbol{x}\|/\varepsilon \rceil$, the reconstruction error satisfies $\|\boldsymbol{x} - \boldsymbol{x}^{(n^*)}\| \le 5\varepsilon$.

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 $\{\tau_i \in \bar{\mathbb{R}} : 1 \le i \le 2^b + 1\}$ (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 $\{q_i \in \mathcal{R}_i : 1 \le i \le 2^b\}$ 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.
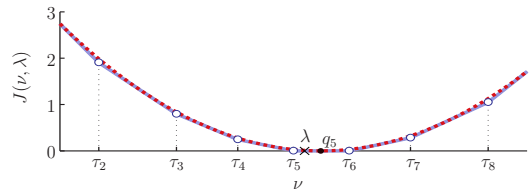

Fig. 1: (plain curve) Plot of J as a function of $\nu \in \mathbb{R}$ for $b = 3$ ($\tau_5 = 0$) and $\lambda \in \mathcal{R}_5$. (dashed curve) Plot of $\tfrac{1}{2}(\nu - q_5)^2$.

Given $\nu, \lambda \in \mathbb{R}$ and using the levels and thresholds associated to $\mathcal{Q}_b$, we first define

$$J(\nu, \lambda) = \sum_{j=2}^{2^b} w_j \left| \left(\operatorname{sign}(\lambda - \tau_j)(\nu - \tau_j) \right)_- \right|, \tag{3}$$

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 $\operatorname{sign}(\lambda - \tau_j) \ne \operatorname{sign}(\nu - \tau_j)$. 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 $\tfrac{1}{2}\|\boldsymbol{u} - \boldsymbol{v}\|^2$. Indeed, in this case $w_j \ll 1$ and, the sum in (3) tends to

$$J(\nu, \lambda) \simeq \big| \int_\nu^\lambda (\nu - t)\,\mathrm{d}t \big| = \tfrac{1}{2}(\nu - \lambda)^2.$$

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):

$$\min_{\boldsymbol{u} \in \mathbb{R}^N} \mathcal{E}_b(\boldsymbol{u}) \text{ s.t. } \|\boldsymbol{u}\|_0 \le K, \tag{5}$$

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

$$\sum_{j=k_-+1}^{k_+} \tfrac{w_j}{2}\left(\operatorname{sign}(\nu - \tau_j) - \operatorname{sign}(\lambda - \tau_j) \right), \tag{6}$$

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 $\mathcal{J}(\boldsymbol{\Phi}\boldsymbol{u}, \boldsymbol{y})$ with respect to $\boldsymbol{u}$ corresponds to $\boldsymbol{\Phi}^*(\mathcal{Q}_b(\boldsymbol{\Phi}\boldsymbol{u}) - \boldsymbol{y})$.

Therefore, from this last ingredient, we define the *quantized iterative hard thresholding algorithm* (QIHT) by the recursion

$$\boldsymbol{x}^{(n+1)} = \mathcal{H}_K\big[\boldsymbol{x}^{(n)} + \mu\boldsymbol{\Phi}^*\big(\boldsymbol{y} - \mathcal{Q}_b(\boldsymbol{\Phi}\boldsymbol{x}^{(n)})\big)\big], \quad \textbf{(QIHT)}$$

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 $\mathcal{Q}_1(\boldsymbol{\Phi}\boldsymbol{a}) = \mathcal{Q}_1(\boldsymbol{\Phi}\boldsymbol{b})$ 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 $\mathcal{Q}_1(\boldsymbol{\Phi}\boldsymbol{a})$ and $\mathcal{Q}_1(\boldsymbol{\Phi}\boldsymbol{b})$, then, with high probability on $\boldsymbol{\Phi}$,

$$\|\frac{\boldsymbol{a}}{\|\boldsymbol{a}\|} - \frac{\boldsymbol{b}}{\|\boldsymbol{b}\|}\| \leq \big(\frac{K+r}{K}\big)\,\delta,$$

for $M = O(\delta^{-1}K\log MN)$. We understand then the beneficial impact of any increase of consistency between $\mathcal{Q}_1(\boldsymbol{\Phi}\boldsymbol{x}^{(n)})$ 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 $\boldsymbol{u} \to \text{sign}(\boldsymbol{\Phi}\boldsymbol{u})$ 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

$$\big|\mu^*\langle\text{sign}(\boldsymbol{\Phi}\boldsymbol{u}), \boldsymbol{\Phi}\boldsymbol{v}\rangle - \langle\boldsymbol{u}, \boldsymbol{v}\rangle\big| \leq \delta, \quad \forall \boldsymbol{u}, \boldsymbol{v} \in \Sigma_K^*, \quad (7)$$

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 $\text{SPE}(\Sigma_K^*, \delta)$. When $\boldsymbol{u}$ is fixed, we say that $\boldsymbol{\Phi}$ is locally $\text{SPE}(\Sigma_K^*, \delta)$ 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 $\text{SPE}(\Sigma_{2K}^*, \delta)$ on $\boldsymbol{x}$ for some $0 < \delta < 1$. Then, given $\boldsymbol{y} = \mathcal{Q}_1[\boldsymbol{\Phi}\boldsymbol{x}] = q_0\,\text{sign}(\boldsymbol{\Phi}\boldsymbol{x})$, 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 $\mathcal{T}_0 = \text{supp}\,\boldsymbol{x}$, $\mathcal{T} = \mathcal{T}_0 \cup \text{supp}\,\hat{\boldsymbol{x}}$, and $\boldsymbol{a} = \frac{1}{q_0^2 M}\boldsymbol{\Phi}^*\boldsymbol{y} = \mu^*\boldsymbol{\Phi}^*\text{sign}(\boldsymbol{\Phi}\boldsymbol{x})$ 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 $\|\boldsymbol{x} - \boldsymbol{a}_\mathcal{T}\| = \sup_{\boldsymbol{w} \in \Sigma_{2K}^*}\langle\boldsymbol{w}, \boldsymbol{x} - \boldsymbol{a}_\mathcal{T}\rangle$ and $\boldsymbol{\Phi}$ is $\text{SPE}(\Sigma_{2K}^*, \delta)$, $\|\boldsymbol{x} - \hat{\boldsymbol{x}}\| \leq 2\sup_{\boldsymbol{w} \in \Sigma_\mathcal{T}^*}\big(\langle\boldsymbol{w}, \boldsymbol{x}\rangle - \mu^*\langle\boldsymbol{\Phi}\boldsymbol{w}, \text{sign}(\boldsymbol{\Phi}\boldsymbol{x})\rangle\big) \leq 2\sup_{\boldsymbol{w} \in \Sigma_\mathcal{T}^*}\big(\langle\boldsymbol{w}, \boldsymbol{x}\rangle - \langle\boldsymbol{w}, \boldsymbol{x}\rangle + \delta\big) = 2\delta$, using $\text{supp}(\boldsymbol{x} - \boldsymbol{a}_\mathcal{T}) \subseteq \mathcal{T}$ 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 $\text{argmax}_{\boldsymbol{u}}\langle\boldsymbol{y}, \boldsymbol{\Phi}\boldsymbol{u}\rangle$ 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.
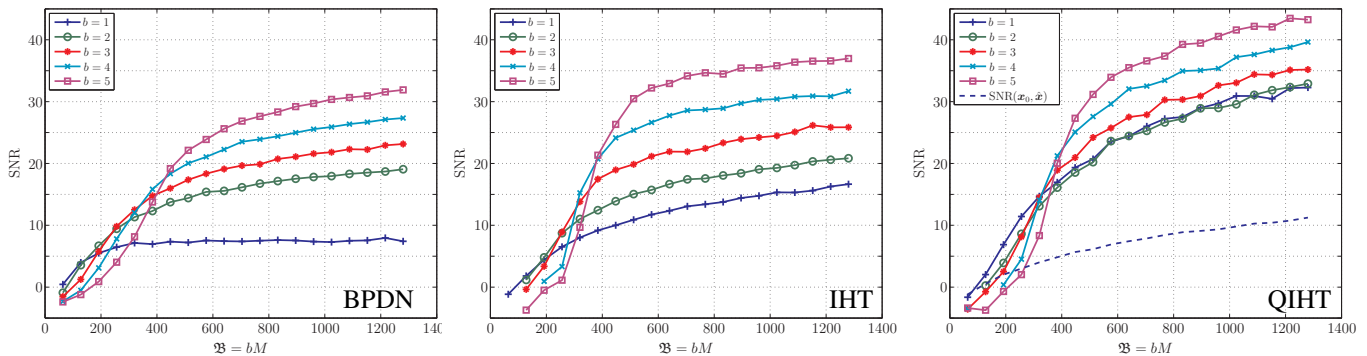
Fig. 2: Comparison between (from left to right) BPDN, IHT and QIHT for several quantization scenarios. The SNR is expressed in dB as a function of the bit budget $\mathfrak{B}$ and the number of bits b used to quantize each measurement.

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 $\mathrm{SNR}(\boldsymbol{x}_0, \boldsymbol{x}^*) = -20\log_{10}\|\boldsymbol{x}_0 - \|\boldsymbol{x}^*\|^{-1}\boldsymbol{x}^*\|$. 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 $b = 4 \to 5$.

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.