Stabilizing Nonuniformly Quantized Compressed Sensing with Scalar Companders

L. Jacques, D. K. Hammond, M. J. Fadili

Abstract

his paper addresses the problem of stably recovering sparse or compressible signals from compressed sensing measurements that have undergone optimal non-uniform scalar quantization, *i.e.*, minimizing the common ℓ_2-norm distortion. Generally, this Quantized Compressed Sensing (QCS) problem is solved by minimizing the ℓ_1-norm constrained by the ℓ_2-norm distortion. In such cases, re-measurement and quantization of the reconstructed signal do not necessarily match the initial observations, showing that the whole QCS model is not *consistent*. Our approach considers instead that quantization distortion more closely resembles heteroscedastic uniform noise, with variance depending on the observed quantization bin. Generalizing our previous work on uniform quantization, we show that for non-uniform quantizers described by the "compander" formalism, quantization distortion may be better characterized as having bounded weighted ℓ_p-norm ($p \geqslant 2$), for a particular weighting. We develop a new reconstruction approach, termed Generalized Basis Pursuit DeNoise (GBPDN), which minimizes the ℓ_1-norm of the signal to reconstruct constrained by this weighted ℓ_p-norm fidelity. We prove that, for standard Gaussian sensing matrices and K sparse or compressible signals in $\mathbb{R}^N$ with at least $\Omega((K \log N/K)^{p/2})$ measurements, *i.e.*, under strongly oversampled QCS scenario, GBPDN is $\ell_2 - \ell_1$ instance optimal and stable recovers all such sparse or compressible signals. The reconstruction error decreases as $O(2^{-B}/\sqrt{p+1})$ given a budget of B bits per measurement. This yields a reduction by a factor $\sqrt{p+1}$ of the reconstruction error compared to the one produced by ℓ_2-norm constrained decoders. We also propose an primal-dual proximal splitting scheme to solve the GBPDN program which is efficient for large-scale problems. Interestingly, extensive simulations testing the GBPDN effectiveness confirm the trend predicted by the

LJ is with the ICTEAM institute, ELEN Department, Université catholique de Louvain (UCL), Belgium. LJ is a Postdoctoral Researcher of the Belgian National Science Foundation (F.R.S.-FNRS).

DKH is with the Neuroinformatics Center, University of Oregon, USA.

MJF is with the GREYC, CNRS-ENSICAEN-Université de Caen, France.

Parts of a preliminary version of this work have been presented in SPARS11 Workshop (June 27-30, 2011 - Edinburgh, Scotland, UK), in IEEE ICIP 2011 (Sept. 11-14, 2011 - Brussels, Belgium) and in iTWIST Workshop (May 9-11, 2012 - Marseille, France).

theory, that the reconstruction error can indeed be reduced by increasing p, but this is achieved at a much less stringent oversampling regime than the one expected by the theoretical bounds. Besides the QCS scenario, we also show that GBPDN applies straightforwardly to the related case of CS measurements corrupted by heteroscedastic Generalized Gaussian noise with provable reconstruction error reduction.his paper addresses the problem of stably recovering sparse or compressible signals from compressed sensing measurements that have undergone optimal non-uniform scalar quantization, *i.e.*, minimizing the common ℓ_2-norm distortion. Generally, this Quantized Compressed Sensing (QCS) problem is solved by minimizing the ℓ_1-norm constrained by the ℓ_2-norm distortion. In such cases, re-measurement and quantization of the reconstructed signal do not necessarily match the initial observations, showing that the whole QCS model is not *consistent*. Our approach considers instead that quantization distortion more closely resembles heteroscedastic uniform noise, with variance depending on the observed quantization bin. Generalizing our previous work on uniform quantization, we show that for non-uniform quantizers described by the "compander" formalism, quantization distortion may be better characterized as having bounded weighted ℓ_p-norm ($p \geqslant 2$), for a particular weighting. We develop a new reconstruction approach, termed Generalized Basis Pursuit DeNoise (GBPDN), which minimizes the ℓ_1-norm of the signal to reconstruct constrained by this weighted ℓ_p-norm fidelity. We prove that, for standard Gaussian sensing matrices and K sparse or compressible signals in $\mathbb{R}^N$ with at least $\Omega((K \log N/K)^{p/2})$ measurements, *i.e.*, under strongly oversampled QCS scenario, GBPDN is $\ell_2 - \ell_1$ instance optimal and stable recovers all such sparse or compressible signals. The reconstruction error decreases as $O(2^{-B}/\sqrt{p+1})$ given a budget of B bits per measurement. This yields a reduction by a factor $\sqrt{p+1}$ of the reconstruction error compared to the one produced by ℓ_2-norm constrained decoders. We also propose an primal-dual proximal splitting scheme to solve the GBPDN program which is efficient for large-scale problems. Interestingly, extensive simulations testing the GBPDN effectiveness confirm the trend predicted by the theory, that the reconstruction error can indeed be reduced by increasing p, but this is achieved at a much less stringent oversampling regime than the one expected by the theoretical bounds. Besides the QCS scenario, we also show that GBPDN applies straightforwardly to the related case of CS measurements corrupted by heteroscedastic Generalized Gaussian noise with provable reconstruction error reduction.T

I. Introduction

A. Problem statement

Measurement quantization is a critical step in the design and in the dissemination of new technologies implementing the Compressed Sensing (CS) paradigm. Quantization is indeed mandatory for transmitting, storing and even processing any data sensed by a CS device.

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]. However, the distortion introduced by any quantization step is often still crudely modeled as a noise with bounded ℓ_2-norm.

Such an approach results in reconstruction methods aiming at finding a sparse signal estimate for which the sensing is close, in a ℓ_2-sense, to the available quantized signal observations. However, earlier works have pointed out that this method is not optimal. For instance, [11] analyses the error achieved when a signal is reconstructed from its quantized coefficients in some overcomplete expansion. Translated to our context, this amounts to the ideal CS scenario where some *oracle* provides us the true signal support knowledge. In this context, a linear *least square* (LS) reconstruction minimizing the ℓ_2-distance in the coefficient domain is inconsistent and has a *mean square error* (MSE) decaying, at best, as the inverse of the frame redundancy factor. Interestingly, any *consistent* reconstruction method, *i.e.*, for which the quantized coefficients of the reconstructed signal match those of the original signal, shows a much better behavior since its MSE is in general lower-bounded by the inverse of the *squared* frame redundancy; this lower bound being attained for specific overcomplete Fourier frames.

A few other works in the Compressed Sensing literature have also considered the quantization distortion differently. In [3], an adaptation of both Basis Pursuit DeNoise (BPDN) program and the Subspace Pursuit algorithm integrates an explicit constraint enforcing consistency. In [5], nonuniform quantization noise and Gaussian noise in the measurements before quantization are properly dealt with using an ℓ_1-penalized maximum likelihood decoder.

Finally, in [4, 6, 7], the extreme case of 1-bit CS is studied, *i.e.*, when only the signs of the measurements are sent to the decoder. These works have shown that consistency with the 1-bit quantized measurements is of paramount importance for reconstructing the signal where straightforward methods relying on ℓ_2 fidelity constraints reach poor estimate quality.

B. Contributions

The present work addresses the problem of recovering sparse or compressive signals in a *given* non-uniform Quantized Compressed Sensing (QCS) scenario. In particular, we assume that the signal measurements have undergone an optimal non-uniform scalar quantization process, *i.e.*, optimized a priori according to a common minimal distortion standpoint with respect to a source with known probability

density function (pdf). This *post-quantization* reconstruction strategy, where only increasing the number of measurements can improve the signal reconstruction, is inspired by other works targeting consistent reconstruction approaches in comparison with methods advocating solutions of minimal ℓ_2-distortion [3, 8, 11]. Our work is therefore distinct from approaches where other quantization schemes (*e.g.*, $\Sigma\Delta$-quantization [13]) are tuned to the global CS formalism or to specific CS decoding schemes (*e.g.*, Message Passing Reconstruction [12]). These techniques often lead to signal reconstruction MSE rapidly decaying with the measurement number M – for instance, a r-order $\Sigma\Delta$-quantization of CS measurements combined with a particular reconstruction procedure has a MSE decaying nearly as $O\big(M^{-r+\frac{1}{2}}\big)$ [13] – but their application involves generally more involved quantization strategies at the CS encoding stage.

This paper also generalizes the results provided in [8] to cover the case of non-uniform scalar quantization of CS measurements. We show that the theory of "Companders" [9] provides an elegant framework for stabilizing the reconstruction of a sparse (or compressible) signal from non-uniformly quantized CS measurements. Under the *High Resolution Assumption* (HRA), *i.e.*, when the bit budget of the quantizer is high and the quantization bins are narrow, the compander theory provides an equivalent description of the action of a quantizer through sequential application of a *comp*ressor, a uniform quantization, then an ex*pander* (see Section II-A for details). As will be clearer later, this equivalence allows us to define new distortion constraints for the signal reconstruction which are more faithful to the non-uniform quantization process given a certain QCS measurement regime.

Algorithms for reconstructing from quantized measurements commonly rely on mathematically describing the noise induced by quantization as bounded in some particular norm. A data fidelity constraint reflecting this fact is then incorporated in the reconstruction method. Two natural examples of such constraints are that the ℓ_2-norm be bounded, or that the quantization error be such that the unquantized values lie in specified, known quantization bins. In this paper, guided by the compander theory, we show that these two constraints can be viewed as special (extreme) cases of a particular *weighted* ℓ_p-norm, which forms the basis for our reconstruction method. The weights are determined from a set of p-optimal quantizer levels, that are computed from the observed quantized values. We draw the reader attention to the fact these weights do not depend on the original signal which is of course unknown. They are used only for signal reconstruction purposes, and are optimized with respect to the weighted norm. In the QCS framework, and owing to the particular weighting of the norm, each quantization bin contributes equally to the related global distortion. Thanks to a new estimator of the weighted ℓ_p-norm of the quantization distortion associated to these particular levels (see Lemma 3), and with the proviso that the

sensing matrix obeys a generalized Restricted Isometry Property (RIP) expressed in the same norm (see (14)), we show that solving a General Basis Pursuit DeNoising program (GBPDN) – an ℓ_1-minimization problem constrained by a weighted ℓ_p-norm whose radius is appropriately estimated – stably recovers strictly sparse or compressible signals (see Theorem 1).

We also quantify precisely the reconstruction error of GBPDN as a function of the quantizer bit rate (under the HRA) for any value of p in the weighted ℓ_p constraint. These results reveal a set of conflicting considerations for setting the optimal p. On the one hand, given a budget of B bits per measurement and for a high number of measurements M, the error decays as $O(2^{-B}/\sqrt{p+1})$ when p increases (see Proposition 3), *i.e.*, a favorable situation since then GBPDN tends also to a consistent reconstruction method. On the other hand, the larger p, the greater the number of measurements required to ensure that the generalized RIP is fulfilled. In particular, one needs $\Omega((K\log N/K)^{p/2})$ measurements compared to a ℓ_2-based CS bound of $\Omega(K\log N/K)$ measurements (see Proposition 1). Put differently, given a certain number of measurements, the range of theoretically admissible p is upper bounded, an effect which is expected since the error due to quantization cannot be eliminated in the reconstruction.

In fact, the stability of GBPDN in the context of QCS is a consequence of a an even more general stability result that holds for a broader class additive heteroscedastic measurement noise having a bounded weighted ℓ_p norm. This for instance covers the case of heteroscedastic Generalized Gaussian noise where the constraint of GBPDN can be interpreted as a (variance) stabilization of the measurement distortion, see Section III-C).

C. Relation to prior work

Our work is novel in several respects. For instance, as stated above, the quantization distortion in the literature is often modeled as a mere Gaussian noise with bounded variance [3]. In [8], only uniform quantization is handled and theoretically investigated. In [5], nonuniform quantization noise and Gaussian noise are handled but theoretical guarantees are lacking. To the best of our knowledge, this is the first work thoroughly investigating the theoretical guarantees of ℓ_1 sparse recovery from non-uniformly quantized CS measurements, by introducing a new class of convex ℓ_1 decoders. The way we bring the compander theory in the picture to compute the optimal weights from the quantized measurements is also an additional originality of this work.

D. Paper organization

The paper is organized as follows. In Section II, we recall the theory of optimal scalar quantization seen through the compander formalism. We then explain how this point of view can help us in understanding the intrinsic constraints that quantized CS measurements must satisfy, and we introduce a new distortion measure, the p-Distortion Consistency, expressed in terms of a weighted ℓ_p-norm. Section III introduces the GBPDN CS class of decoders integrating weighted ℓ_p-constraints, and describes sufficient conditions for guaranteeing reconstruction stability. This section shows also the generality of this procedure for stabilizing additive heteroscedastic GGD measurement noise during the signal reconstruction. In Section IV, we explain how GBPDN can be used for reconstructing a signal in QCS when its fidelity constraint is adjusted to the parameters defined in Section II-C. We show that this specific choice leads to a (variance) stabilization of the quantization distortion forcing each quantization bin to contribute equally to the overall distortion error. In Section V, we describe a provably convergent primal-dual proximal splitting algorithm to solve the GBPDN program, and demonstrate the power of the proposed approach with several numerical experiments on sparse signals.

E. Notation

All finite space dimensions are denoted by capital letters (*e.g.*, $K, M, N, D \in \mathbb{N}$), vectors (resp. matrices) are written in small (resp. capital) bold symbols. For any vector $\boldsymbol{u}$, the ℓ_p-norm for $1 \leqslant p < \infty$ is $\|\boldsymbol{u}\|_p = (\sum_i |u_i|^p)^{1/p}$, as usual $\|\boldsymbol{u}\|_\infty = \max_i |u_i|$ and we write $\|\boldsymbol{u}\| = \|\boldsymbol{u}\|_2$. We write $\|\boldsymbol{u}\|_0 = \#\{i : u_i \neq 0\}$, which counts the number of non-zero components. We denote the set of K-sparse vectors in the canonical basis by $\Sigma_K = \{\boldsymbol{u} \in \mathbb{R}^N : \|\boldsymbol{u}\|_0 \leqslant K\}$. When necessary, we write ℓ_p^D as the normed vector space $(\mathbb{R}^D, \|\cdot\|_p)$. The identity matrix in $\mathbb{R}^D$ is written $\mathbb{1}_D$ (or simply $\mathbb{1}$ if the D is clear from the context). $\boldsymbol{U} = \mathrm{diag}(\boldsymbol{u})$ is the diagonal matrix with diagonal entries from $\boldsymbol{u}$, *i.e.*, $\boldsymbol{U}_{ij} = u_i \delta_{ij}$. Given the N-dimensional signal space $\mathbb{R}^N$, the index set is $[N] = \{1, \cdots, N\}$, and $\boldsymbol{\Phi}_I \in \mathbb{R}^{M \times \#I}$ is the restriction of the columns of $\boldsymbol{\Phi}$ to those indexed in the subset $I \subset [N]$, whose cardinality is $\#I$. Given $\boldsymbol{x} \in \mathbb{R}^N$, $\boldsymbol{x}_K^{\boldsymbol{\Psi}}$ stands for the best K-term ℓ_2-approximation of $\boldsymbol{x}$ in the orthonormal basis $\boldsymbol{\Psi} \in \mathbb{R}^{N \times N}$, that is, $\boldsymbol{x}_K^{\boldsymbol{\Psi}} = \boldsymbol{\Psi}\big(\operatorname{argmin}\{\|\boldsymbol{x} - \boldsymbol{\Psi}\boldsymbol{\zeta}\| : \boldsymbol{\zeta} \in \mathbb{R}^N, \|\boldsymbol{\zeta}\|_0 \leqslant K\}\big)$. When $\boldsymbol{\Psi} = \mathbb{1}$, we write $\boldsymbol{x}_K = \boldsymbol{x}_K^{\mathbb{1}}$ with $\|\boldsymbol{x}_K\|_0 \leqslant K$. A random matrix $\boldsymbol{\Phi} \sim \mathcal{N}^{M \times N}(0, 1)$ is a $M \times N$ matrix with entries $\boldsymbol{\Phi}_{ij} \sim_{\text{iid}} \mathcal{N}(0, 1)$. The 1-D Gaussian pdf of mean $\mu \in \mathbb{R}$ and variance $\sigma^2 \in \mathbb{R}_+^*$ is denoted $\gamma_{\mu,\sigma}(t) := (2\pi\sigma^2)^{-1/2} \exp(-\frac{(t-\mu)^2}{2\sigma^2})$. For a function $f : \mathbb{R} \to \mathbb{R}$, we write $\|\!|f|\!\|_q := (\int_{\mathbb{R}} \mathrm{d}t\, |f(t)|^q)^{1/q}$, with $\|\!|f|\!\|_\infty := \sup_{t \in \mathbb{R}} |f(t)|$.

In order to state many results which hold asymptotically as a dimension $D \in \mathbb{R}$ increases, we will use the common Landau family of notations, *i.e.*, the symbols O, Ω, Θ, o, and ω (their exact definition can be found in [14]). Additionally, for $f, g \in C^1(\mathbb{R}_+)$, we write $f(D) \simeq_D g(D)$ when $f(D) = g(D)(1+o(1))$. We also introduce two new asymmetric notations dealing with asymptotic quantity ordering, *i.e.*,

$$f(D) \lesssim_D g(D) \quad \Leftrightarrow \ \exists\, \delta : \mathbb{R} \to \mathbb{R}_+ : \ f(D) + \delta(D) \simeq_D g(D)$$
$$f(D) \gtrsim_D g(D) \quad \Leftrightarrow \ -f(D) \lesssim_D -g(D).$$

If any of the asymptotic relations above hold with respect to several large dimensions $D_1, D_2, \cdots$, we write $\simeq_{D_1,D_2,\ldots}$ and correspondingly for $\lesssim$ and $\gtrsim$.

II. Non-Uniform Quantization in Compressed Sensing

Let us consider a signal $\boldsymbol{x} \in \mathbb{R}^N$ to be measured. We assume that it is either strictly sparse or compressible, in a prescribed orthonormal basis $\boldsymbol{\Psi} = \big(\boldsymbol{\Psi}_1, \cdots, \boldsymbol{\Psi}_N\big) \in \mathbb{R}^{N\times N}$. This means that the signal $\boldsymbol{x} = \boldsymbol{\Psi}\boldsymbol{\zeta} = \sum_j \boldsymbol{\Psi}_j \zeta_j$ is such that the ℓ_2^N-approximation error $\|\boldsymbol{\zeta} - \boldsymbol{\zeta}_K\| = \|\boldsymbol{x} - \boldsymbol{x}_K^{\boldsymbol{\Psi}}\|$ quickly decreases (or vanishes) as K increases. For the sake of simplicity, and without loss of generality, the sparsity basis is taken in the sequel as the standard basis, *i.e.*, $\boldsymbol{\Psi} = \mathbb{1}$, and $\boldsymbol{\zeta}$ is identified with $\boldsymbol{x}$. All the results can be readily extended to other orthonormal bases $\boldsymbol{\Psi} \neq \mathbb{1}$.

In this paper, we are interested in compressively sensing $\boldsymbol{x} \in \mathbb{R}^N$ with a given measurement matrix $\boldsymbol{\Phi} \in \mathbb{R}^{M\times N}$. Each CS measurement, *i.e.*, each entry of $\boldsymbol{z} = \boldsymbol{\Phi}\boldsymbol{x}$, undergoes a general scalar *quantization*. We will assume this quantization to be optimal relative to a known distribution of each entry z_i. For simplicity, we only consider matrices $\boldsymbol{\Phi}$ that yield z_i to be i.i.d. $\mathcal{N}(0, \sigma_0^2)$ Gaussian, with pdf $\varphi_0 := \gamma_{0,\sigma_0}$. This is satisfied, for instance, if $\boldsymbol{\Phi} \sim \mathcal{N}^{M\times N}(0,1)$, with $\sigma_0 = \|\boldsymbol{x}\|_2$. When $\boldsymbol{\Phi} = [\boldsymbol{\varphi}_1^T, \cdots, \boldsymbol{\varphi}_M^T]^T$ is a (fixed) realization of $\mathcal{N}^{M\times N}(0,1)$, the entries $z_j = \langle \boldsymbol{\varphi}_j, \boldsymbol{x}\rangle$ of the vector $\boldsymbol{z} = \boldsymbol{\Phi}\boldsymbol{x}$ are M (fixed) realizations of the same Gaussian distribution $\mathcal{N}(0, \|\boldsymbol{x}\|^2)$. It is therefore legitimate to quantize these values optimally using the normality of the source.[1].

Our quantization scenario uses a B-bit quantizer $\mathcal{Q}$ which has been optimized with respect to the measurement pdf φ_0 for $\mathcal{B} = 2^B = \#\Omega$ levels $\Omega = \{\omega_k : 1 \leqslant k \leqslant \mathcal{B}\}$ and thresholds $\{t_k : 1 \leqslant k \leqslant \mathcal{B}+1\}$ with $-t_1 = t_{\mathcal{B}+1} = +\infty$. Unlike the framework developed in [5], our sensing scenario considers that

[1]Avoiding pathological situations where $\boldsymbol{x}$ is adversarially forged knowing $\boldsymbol{\Phi}$ for breaking this assumption.

any noise corrupting the measurements before quantization is negligible compared to the quantization distortion.

Consequently, given a measurement matrix $\mathbf{\Phi} \in \mathbb{R}^{M\times N}$, our quantized sensing model is

$$\boldsymbol{y} \;=\; \mathcal{Q}[\mathbf{\Phi}\boldsymbol{x}] \;=\; \mathcal{Q}[\boldsymbol{z}] \;\in \Omega^M. \tag{1}$$

Following recent studies [3, 8, 15] in the CS literature, this work is interested in optimizing the signal reconstruction stability from $\boldsymbol{y}$ under different sensing conditions, for instance, when the *oversampling ratio* M/K is allowed to be large. Before going further into this signal sensing model, let us describe first the selected quantization framework. The latter is based on a scalar quantization of each component of the signal measurement vector.

A. Quantization, Companders and Distortion

A scalar quantizer $\mathcal{Q}$ is defined from $\mathcal{B} = 2^B$ *levels* ω_k (coded by $B = \log_2 \mathcal{B}$ bits) and $\mathcal{B}+1$ *thresholds* $t_k \in \mathbb{R} \cup \{\pm\infty\} = \overline{\mathbb{R}}$, with $\omega_k < \omega_{k+1}$ and $t_k \leqslant \omega_k < t_{k+1}$ for all $1 \leqslant k \leqslant \mathcal{B}$. The k^{th} quantizer *bin* (or *region*) is $\mathcal{R}_k = [t_k, t_{k+1})$, with bin width $\tau_k = t_{k+1} - t_k$. The quantizer $\mathcal{Q}$ is a map: $\mathbb{R} \to \Omega = \{\omega_k : 1 \leqslant k \leqslant \mathcal{B}\}$, $t \mapsto \mathcal{Q}[t] = \omega_k \iff t \in \mathcal{R}_k$. An optimal scalar quantizer $\mathcal{Q}$ with respect to a random source $\mathcal{Z}$ with pdf $\varphi_{\mathcal{Z}}$ is such that the distortion $\mathbb{E}|\mathcal{Z} - \mathcal{Q}[\mathcal{Z}]|^2$ is minimized. Optimal levels and thresholds can be calculated for a fixed number of quantization bins by the Lloyd-Max Algorithm [16, 17], or by an asymptotic (with respect to B) *companding* approach [9].

Throughout this paper, we work under the HRA. This means that, given the source pdf $\varphi_{\mathcal{Z}}$, the number of bits B is sufficient to validate the approximation

$$\varphi_{\mathcal{Z}}(t) \simeq_B \varphi_{\mathcal{Z}}(\omega_k), \quad \forall t \in \mathcal{R}_k. \tag{**HRA**}$$

A common argument in quantization theory [9] states that under the HRA, every optimal regular quantizer can be described by a compander (a portemanteau for "**comp**ressor" and "exp**ander**"). More precisely, we have

$$\mathcal{Q} = \mathcal{G}^{-1} \circ \mathcal{Q}_\alpha \circ \mathcal{G},$$

with $\mathcal{G} : \mathbb{R} \to [0,1]$ a bijective function called the *compressor*, $\mathcal{Q}_\alpha$ a uniform quantizer of the interval $[0,1]$ of bin width $\alpha = 2^{-B}$, and the inverse mapping $\mathcal{G}^{-1} : [0,1] \to \mathbb{R}$ called the *expander*.

For optimal quantizers the compressor $\mathcal{G}$ maps the thresholds $\{t_k : 1 \leqslant k \leqslant \mathcal{B}\}$ and the levels $\{\omega_k\}$ into the values

$$t'_k := \mathcal{G}(t_k) = (k-1)\alpha, \qquad \omega'_k := \mathcal{G}(\omega_k) = (k-1/2)\alpha, \tag{2}$$

and under the HRA the optimal $\mathcal{G}$ satisfies

$$\mathcal{G}' := \tfrac{\mathrm{d}}{\mathrm{d}\lambda}\mathcal{G}(\lambda) \;=\; \left[\int_{\mathbb{R}} \varphi_{\mathcal{Z}}^{1/3}(t)\,\mathrm{d}t\right]^{-1} \varphi_{\mathcal{Z}}^{1/3}(\lambda). \tag{3}$$

Intuitively, the function $\mathcal{G}'$, also called *quantizer point density function* (qpdf) [9], relates the quantizer bin widths before and after domain compression by $\mathcal{G}$. Indeed, under HRA, we can show that $\mathcal{G}'(\lambda) \simeq \alpha/\tau_k$ if $\lambda \in \mathcal{R}_k$. We will see later that this function is the key to conveniently weight some new quantizer distortion measures.

We note that, for $\varphi_{\mathcal{Z}}(t) = \gamma_{0,\sigma}(t)$ with cumulative distribution function $\phi_{\mathcal{Z}}(\lambda;\sigma^2) = \frac{1}{2}\mathrm{erfc}(-\frac{\lambda}{2\sigma})$ so that $\phi_{\mathcal{Z}}^{-1}(\lambda';\sigma^2) = \sigma\sqrt{2}\,\mathrm{erf}^{-1}(2\lambda'-1)$, we have $\mathcal{G}(\lambda) = \phi_{\mathcal{Z}}(\lambda;3\sigma^2)$ and $\mathcal{G}^{-1}(\lambda') = \phi_{\mathcal{Z}}^{-1}(\lambda';3\sigma^2)$.

The application of $\mathcal{G}$ modifies the source $\mathcal{Z}$ such that $\mathcal{G}(\mathcal{Z}) - \mathcal{G}(\mathcal{Q}[\mathcal{Z}])$ behaves more like a uniformly distributed random variable over $[-\alpha/2, \alpha/2]$. The compander formalism predicts the distortion of optimal scalar quantizer under HRA. For high bit rate B, the Panter and Dite formula [18] states that

$$\mathbb{E}|\mathcal{Z} - \mathcal{Q}[\mathcal{Z}]|^2 \;\underset{B}{\simeq}\; \tfrac{2^{-2B}}{12}\int_{\mathbb{R}} \mathcal{G}'(t)^{-2}\,\varphi_{\mathcal{Z}}(t)\,\mathrm{d}t \;=\; \tfrac{2^{-2B}}{12}\left(\int_{\mathbb{R}} \varphi_{\mathcal{Z}}^{1/3}(t)\,\mathrm{d}t\right)^3 \;=\; \tfrac{2^{-2B}}{12}\,\|\!|\varphi_{\mathcal{Z}}|\!\|_{1/3}. \tag{4}$$

Finally, we note that by the construction defined in (2), the quantized values $\mathcal{Q}[\lambda]$ satisfy

$$|\mathcal{G}(\lambda) - \mathcal{G}(\mathcal{Q}[\lambda])| \leqslant \alpha/2, \quad \forall \lambda \in \mathbb{R}. \tag{5}$$

We describe in the next sections how (5) and (4) may be viewed as two extreme cases of a general class of constraints satisfied by a quantized source $\mathcal{Z}$.

B. Distortion and Quantization Consistency

Let us consider the sensing model (1), for which the scalar quantizer $\mathcal{Q}$ and associated compressor $\mathcal{G}$ are optimal relative to the measurements $\boldsymbol{z} = \boldsymbol{\Phi}\boldsymbol{x}$ whose entries z_i are iid realizations of $\mathcal{N}(0,\sigma_0^2)$. In the compressor domain we may write

$$\mathcal{G}(\boldsymbol{y}) \;=\; \mathcal{G}(\boldsymbol{z}) + (\mathcal{G}(\mathcal{Q}[\boldsymbol{z}]) - \mathcal{G}(\boldsymbol{z})) = \mathcal{G}(\boldsymbol{z}) + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon}$ represents the quantization distortion. (5) then shows that

$$\|\boldsymbol{\varepsilon}\|_\infty = \|\mathcal{G}(\mathcal{Q}[\boldsymbol{z}]) - \mathcal{G}(\boldsymbol{z})\|_\infty \leqslant \alpha/2.$$

Naively, one may expect any reasonable estimate $\boldsymbol{x}^*$ of $\boldsymbol{x}$ (obtained by some reconstruction method) to reproduce the same quantized measurements as originally observed. Inspired by the terminology introduced in [10, 11], we say that $\boldsymbol{x}^*$ satisfies the *quantization consistency* (QC) if $\mathcal{Q}[\boldsymbol{\Phi}\boldsymbol{x}^*] = \boldsymbol{y}$. From the previous reasoning this is equivalent to

$$\|\mathcal{G}(\boldsymbol{\Phi}\boldsymbol{x}^*) - \mathcal{G}(\boldsymbol{y})\|_\infty \ \leqslant \ \epsilon_{\mathrm{QC}} := \alpha/2. \tag{$\mathbf{QC}$}$$

At first glance, it is tempting to try to impose directly QC in the data fidelity constraint. However, as will be revealed by our analysis, directly imposing QC does *not* lead to an effective QCS reconstruction algorithm. This counterintuitive effect, already observed in the case of signal recovery from uniformly quantized CS [8], is due to the specific requirements that the sensing matrix should respect to make such a consistent reconstruction method stable.

In contrast the Basis Pursuit DeNoise (BPDN) program [19] enforces a constraint on the ℓ_2 norm of the reconstruction quantization error, which we will call distortion consistency. For BPDN, the estimate $\boldsymbol{x}^*$ is provided by

$$\boldsymbol{x}^* \ \in \ \underset{\boldsymbol{u}\in\mathbb{R}^N}{\mathrm{Argmin}} \ \|\boldsymbol{u}\|_1 \ \text{s.t.} \ \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{u}\| \leqslant \epsilon_{\mathrm{DC}},$$

where the bound $\epsilon^2_{\mathrm{DC}} := M \, \frac{\sqrt{3}\,\pi}{2} \, \sigma_0^2 \, 2^{-2B}$ is dictated by the Panter-Dite formula. According to the Strong Law of Large Numbers (SLLN) obeyed by the HRA, and since z_i are iid realizations of $Z \sim \mathcal{N}(0, \sigma_0^2)$, the following holds almost surely

$$\tfrac{1}{M}\|\boldsymbol{z} - \mathcal{Q}[\boldsymbol{z}]\|^2 \ \underset{M}{\simeq} \ \mathbb{E}|\mathcal{Z} - \mathcal{Q}[\mathcal{Z}]|^2 \ \underset{B}{\simeq} \ \tfrac{2^{-2B}}{12}\|\!|\varphi_0|\!\|_{1/3} = \ \tfrac{\sqrt{3}\,\pi}{2} \, \sigma_0^2 \, 2^{-2B}. \tag{6}$$

Accordingly, we say that any estimate $\boldsymbol{x}^*$ satisfies *distortion consistency* (DC) if

$$\|\boldsymbol{\Phi}\boldsymbol{x}^* - \boldsymbol{y}\| \leqslant \epsilon_{\mathrm{DC}}. \tag{$\mathbf{DC}$}$$

However, as stated for the uniform quantization case in [8], DC and QC do not imply each other. In particular, the output $\boldsymbol{x}^*$ of BPDN needs not satisfy quantization consistency. A major motivation for the present work is the desire to develop provably stable QCS recovery methods based on measures of quantization distortion that are as close as possible to QC.

C. p-Distortion Consistency

This section shows that the QC and DC constraints may be seen as limit cases of a weighted ℓ_p-norm description of the quantization distortion. The expression of the appropriate weights in the weighted ℓ_p

norm will depend both on the p-optimal quantizer levels, described below, and of the quantizer point density function $\mathcal{G}'$ introduced in Section II-A.

For the Gaussian pdf $\varphi_0 = \gamma_{0,\sigma_0}$, given a set of thresholds $\{t_k : 1 \leqslant k \leqslant \mathcal{B}\}$, we define the *$p$-optimal quantizer levels* $\omega_{k,p} \in \overline{\mathbb{R}}$ as

$$\omega_{k,p} := \operatorname*{argmin}_{\lambda \in \mathcal{R}_k} \int_{\mathcal{R}_k} |t-\lambda|^p \ \varphi_0(t)\,\mathrm{d}t, \tag{7}$$

for $2 \leqslant p < \infty$, and $\omega_{k,\infty} := \frac{1}{2}(t_k + t_{k+1})$. These generalized levels were for instance already defined by Max in his minimal distortion study [17], and their definition (7) is also related to the concept of minimal p^{th}-power distortion [9]. For $p = 2$, we find the definition of the initial quantizer levels, *i.e.*, $\omega_{k,2} = \omega_k$. In this paper, we always assume that p is a positive integer but all our analysis can be extended to the positive real case. As proved in Appendix B, the p-optimal levels are well-defined.

Lemma 1 (**p-optimal Level Well-Definiteness).** *The p-optimal levels $\omega_{k,p}$ are uniquely defined. Moreover, for $\sigma_0 > 0$, $\lim_{p\to+\infty} \omega_{k,p} = \omega_{k,\infty}$, with $|\omega_{k,p}| = \Omega(\sqrt{p})$ for $k \in \{1, \mathcal{B}\}$.*

Using these new levels, we define the (suboptimal) quantizers $\mathcal{Q}_p$ (with $\mathcal{Q}_2 = \mathcal{Q}$) such that

$$\mathcal{Q}_p[t] = \omega_{k,p} \ \Leftrightarrow \ t \in \mathcal{R}_k = \mathcal{Q}_p^{-1}[\omega_{k,p}] = \mathcal{Q}^{-1}[\omega_k]. \tag{8}$$

Two important points must be explained regarding the definition of $\mathcal{Q}_p$. First, the *(re)quantization* of any source $\mathcal{Z}$ with $\mathcal{Q}_p$ is possible from the knowledge of the quantized value $\mathcal{Q}[\mathcal{Z}]$, as $\mathcal{Q}_p[\mathcal{Z}] = \mathcal{Q}_p[\mathcal{Q}[\mathcal{Z}]]$ since both quantizers share the same decision thresholds. Second, despite the sub-optimality of $\mathcal{Q}_p$ relative to the untouched thresholds $\{t_k : 1 \leqslant k \leqslant \mathcal{B}\}$, we will see later that introducing this quantizer provides improvement in the modeling of $\mathcal{Q}_p[\mathcal{Z}] - \mathcal{Z}$ by a Generalized Gaussian Distribution (GGD) in each quantization bin.

Remark 1. *Unfortunately, there is no closed form formula for computing $\omega_{k,p}$. However, as detailed in Appendix H, they can be computed up to numerical precision using Newton's method combined with simple numerical quadrature for the integral in* (7)*.*

Given $p \geqslant 2$ and for high B, the asymptotic behaviour of a quantizer $\mathcal{Q}_p$ and of its p^{th} *power distortion* $\int_{\mathcal{R}_k} |t-\omega_{k,p}|^p \ \varphi_0(t)\,\mathrm{d}t$ in each bin $\mathcal{R}_k$ follows two very different regimes in $\mathbb{R}$ governed by a particular transition value $T = \Theta(\sqrt{B})$. This is described in the following lemma (proved in Appendix C), which, to the best of our knowledge, provides new results and may be of independent interest for characterizing

Gaussian source quantization (even for the standard case $p = 2$).

Lemma 2 (**Asymptotic p-Quantization Characterization).** *Given the Gaussian pdf φ_0 and its associated compressor $\mathcal{G}$ function, choose $0 < \beta < 1$ and $p \in \mathbb{N}$, and define the transition value*

$$T = T(B) = (6\,\sigma_0^2(\log 2^\beta)\,B)^{1/2}.$$

T defines two specific asymptotic regimes for the quantizer $\mathcal{Q}_p$:

1) *The* vanishing bin regime $\mathcal{T} = [-T, T]$*: for all $\mathcal{R}_k \subset \mathcal{T}$ and any $c \in \mathcal{R}_k$, the bin widths decay as $\tau_k = O(2^{-(1-\beta)B})$, and the the related p^{th}-power distortion and* qpdf *asymptotically obey*

$$\int_{\mathcal{R}_k} |t - \omega_{k,p}|^p\ \varphi_0(t)\,\mathrm{d}t\ \simeq_B\ \tfrac{\tau_k^{p+1}}{(p+1)\,2^p}\,\varphi_0(c), \tag{9}$$

$$\mathcal{G}'(c)\ \simeq_B\ \tfrac{\alpha}{\tau_k}. \tag{10}$$

2) *The* vanishing distortion regime $\mathcal{T}^c$*: we have $\mathcal{G}'(t) \leqslant \mathcal{G}'(T(B)) = \Theta(2^{-\beta B})$ for all $t \in \mathcal{T}^c$. Moreover, the number of bins in $\mathcal{T}^c$ and their p^{th}-power distortion decay, respectively, as*

$$\#\{k : \mathcal{R}_k \subset \mathcal{T}^c\} = \Theta\big(B^{-1/2}\,2^{(1-\beta)B}\big), \tag{11}$$

$$\int_{\mathcal{R}_k} |t - \omega_{k,p}|^p\ \varphi_0(t)\,\mathrm{d}t = O\big(B^{-(p+1)/2}\,2^{-3\beta B}\big), \quad \forall \mathcal{R}_k \subset \mathcal{T}^c. \tag{12}$$

We now state an important result, proved in Appendix D from the statements of Lemma 2, which, together with the SLLN, estimates the quantization distortion of $\mathcal{Q}_p$ on a random Gaussian vector. Given $p \geqslant 1$ and some positive weights $\boldsymbol{w} = (w_1, \cdots, w_M)^T \in \mathbb{R}_+^M$, this distortion is measured by a weighted ℓ_p-norm defined as[2] $\|\boldsymbol{v}\|_{p,\boldsymbol{w}} := \|\operatorname{diag}(\boldsymbol{w})\,\boldsymbol{v}\|_p$ for any $\boldsymbol{v} \in \mathbb{R}^M$.

Lemma 3 (**Asymptotic Weighted ℓ_p-Distortion).** *Let $\boldsymbol{z} \in \mathbb{R}^M$ be a random vector where each component $z_i \sim_{\text{iid}} \varphi_0$. Given the optimal compressor function $\mathcal{G}$ associated to φ_0 and the weights $\boldsymbol{w} = \boldsymbol{w}(p)$ such that $w_i(p) = \mathcal{G}'\big(\mathcal{Q}_p[z_i]\big)^{(p-2)/p}$ for $p \geqslant 2$, the following holds almost surely*

$$\|\mathcal{Q}_p[\boldsymbol{z}] - \boldsymbol{z}\|_{p,\boldsymbol{w}}^p\ \underset{B,M}{\simeq}\ M\,\tfrac{2^{-Bp}}{(p+1)\,2^p}\,\|\!|\varphi_0|\!\|_{1/3}\ =:\ \epsilon_p^p, \tag{13}$$

with $\|\!|\varphi_0|\!\|_{1/3} = 2\pi\,\sigma_0^2\,3^{3/2}$.

This lemma provides a tight estimation for $p = 2$ and $p \to +\infty$. Indeed, in the first case $\boldsymbol{w} = \mathbf{1}$ and the bound matches the Panter-Dite estimation (6). For $p \to \infty$, we observe that $\epsilon_\infty = 2^{-(B+1)} = \alpha/2 = \epsilon_{\text{QC}}$.

[2]A more standard weighted ℓ_p-norm definition reads $(\sum_i w_i |v_i|^p)^{1/p}$. Our definition choice, which is strictly equivalent, offers useful writing simplifications, *e.g.*, when observing that $\|\boldsymbol{\Phi}\boldsymbol{x}\|_{p,\boldsymbol{w}} = \|\boldsymbol{\Phi}'\boldsymbol{x}\|_p$ with $\boldsymbol{\Phi}' = \operatorname{diag}(\boldsymbol{w})\boldsymbol{\Phi}$.

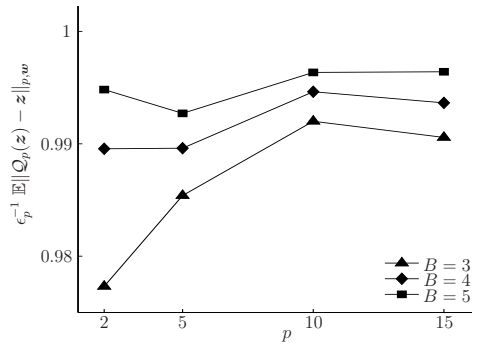


Fig. 1: Comparing the theoretical bound ϵ_p to the empirical mean estimate of $\mathbb{E}\|\mathcal{Q}_p[\boldsymbol{z}]-\boldsymbol{z}\|_{p,\boldsymbol{w}}$ using 1000 trials of Monte-Carlo simulations, for each $B=3,4,5$).

Fig. 1 shows how well the ϵ_p estimates the distortion $\|\mathcal{Q}_p[\boldsymbol{z}]-\boldsymbol{z}\|_{p,\boldsymbol{w}}$ for the weights and the p-optimal levels given in Lemma 2. This has been measured by averaging this quantization distortion for 1000 realizations of a Gaussian random vector $\sim\mathcal{N}^M(0,1)$ with $M=2^{10}$, $p\in\{2,\cdots,15\}$ and $B=3,4$ and 5. We observe that the bias of ϵ_p, as reflected here by the ratio $\epsilon_p^{-1}\,\mathbb{E}\|\mathcal{Q}_p[\boldsymbol{z}]-\boldsymbol{z}\|_{p,\boldsymbol{w}}$, is rather limited and decreases when p and B increase with a maximum relative error of about 2.5% between the true and estimated distortion at $B=3$ and $p=2$.

Inspired by relation (13), we say that an estimate $\boldsymbol{x}^*\in\mathbb{R}^N$ of $\boldsymbol{x}$ sensed by the model (1) satisfies the *p-Distortion Consistency* (or D$_p$C) if

$$\|\boldsymbol{\Phi x}^*-\mathcal{Q}_p[\boldsymbol{y}]\|_{p,\boldsymbol{w}}\ \leqslant\ \epsilon_p, \qquad \textbf{(D}_p\textbf{C)}$$

with the weights $w_i(p)=\mathcal{G}'\big(\mathcal{Q}_p[y_i]\big)^{(p-2)/p}$.

The class of D$_p$C constraints has QC and DC as its limit cases.

Lemma 4. *Given $\boldsymbol{y}=\mathcal{Q}[\boldsymbol{\Phi x}]$, we have asymptotically in B*

$$\mathrm{D}_2\mathrm{C}\ \equiv\ \mathrm{DC}\qquad \textit{and}\qquad \mathrm{D}_\infty\mathrm{C}\ \equiv\ \mathrm{QC}.$$

Proof: Let $\boldsymbol{x}^*\in\mathbb{R}^N$ be a vector to be tested with the DC, QC or D$_p$C constraints. The first equivalence for $p=2$ is straightforward since $\boldsymbol{w}(2)=\mathbf{1}$, $\|\boldsymbol{\Phi x}^*-\mathcal{Q}_p[\boldsymbol{y}]\|_{p,\boldsymbol{w}}=\|\boldsymbol{\Phi x}^*-\mathcal{Q}[\boldsymbol{y}]\|_2$ and $\epsilon_2^2=\epsilon_{\mathrm{DC}}^2=\frac{2^{-2B}}{12}\,\|\varphi_0\|_{1/3}$ from (6).

For the second, we use the fact that $\boldsymbol{y}=\mathcal{Q}[\boldsymbol{\Phi x}]$ is fixed by the sensing model (1). Let us denote by $k(i)$ the index of the bin to which $\mathcal{Q}_p[y_i]$ belongs for $1\leqslant i\leqslant M$. Since $\|\boldsymbol{\Phi x}\|_\infty$ is fixed, and

because relation (11) in Lemma 2 implies that the amplitude of the first or of the last $\Theta(B^{-1/2}2^{(1-\beta)B})$ thresholds grow faster than $T = \Theta(\sqrt{\beta B})$ for $0 < \beta < 1$, there exists necessarily a $B_0 \geqslant 0$ such that $-T(B) \leqslant t_{k(i)} \leqslant t_{k(i)+1} \leqslant T(B)$ for all $B \geqslant B_0$ and all $1 \leqslant i \leqslant M$.

Writing $\boldsymbol{W}_p = \mathrm{diag}(\boldsymbol{w}(p))$, we can use the equivalence $\|\cdot\|_\infty \leqslant \|\cdot\|_p \leqslant M^{1/p}\|\cdot\|_\infty$ and the squeeze theorem on the following limit:

$$\lim_{p\to\infty} \|\boldsymbol{\Phi x}^* - \mathcal{Q}_p[\boldsymbol{y}]\|_{p,\boldsymbol{w}(p)} = \lim_{p\to\infty} \|\boldsymbol{W}_p\big(\Phi \boldsymbol{x}^* - \mathcal{Q}_p[\boldsymbol{y}]\big)\|_p = \lim_{p\to\infty} \|\boldsymbol{W}_p\big(\Phi \boldsymbol{x}^* - \mathcal{Q}_p[\boldsymbol{y}]\big)\|_\infty.$$

Moreover, since for $B \geqslant B_0$ and for all $1 \leqslant i \leqslant M$ the bin $\mathcal{R}_{k(i)}$ is finite, the limit

$$\lim_{p\to\infty} \mathcal{G}'(\mathcal{Q}_p[y_i])^{(p-2)/p} \left|(\boldsymbol{\Phi x}^*)_i - \mathcal{Q}_p[y_i]\right|$$

exists and is finite. Therefore, from the continuity of the $\max$ function applied on the M components of vectors in $\mathbb{R}^M$, we find

$$\begin{aligned}
\lim_{p\to\infty} \|\boldsymbol{\Phi x}^* - \mathcal{Q}_p[\boldsymbol{y}]\|_{p,\boldsymbol{w}(p)} &= \lim_{p\to\infty} \max_i \mathcal{G}'(\mathcal{Q}_p[y_i])^{(p-2)/p} \left|(\boldsymbol{\Phi x}^*)_i - \mathcal{Q}_p[y_i]\right| \\
&= \max_i \lim_{p\to\infty} \mathcal{G}'(\mathcal{Q}_p[y_i])^{(p-2)/p} \left|(\boldsymbol{\Phi x}^*)_i - \mathcal{Q}_p[y_i]\right| \\
&= \max_i \mathcal{G}'(\mathcal{Q}_\infty(y_i)) \left|(\boldsymbol{\Phi x}^*)_i - \mathcal{Q}_\infty(y_i)\right|.
\end{aligned}$$

For $B \geqslant B_0$, (10) provides $\mathcal{G}'(\mathcal{Q}_\infty(y_i)) \simeq_B \frac{\alpha}{\tau_{k(i)}}$, so that, if we impose $\lim_{p\to\infty} \|\boldsymbol{\Phi x}^* - \mathcal{Q}_p[\boldsymbol{y}]\|_{p,\boldsymbol{w}(p)} \leqslant \epsilon_{\mathrm{QC}} = \alpha/2$, we get asymptotically in B

$$\max_i \tfrac{1}{\tau_{k(i)}} \left|(\boldsymbol{\Phi x}^*)_i - \mathcal{Q}_\infty(y_i)\right| \underset{B}{\lesssim} \tfrac{1}{2},$$

which is equivalent to imposing $(\boldsymbol{\Phi x}^*)_i \in \mathcal{R}_{k(i)}$, *i.e.*, the Quantization Constraint. ■

III. Weighted ℓ_p Fidelities in Compressed Sensing and General Reconstruction Guarantees

The last section has provided us some weighted $\ell_{p,\boldsymbol{w}}$ constraints, with appropriate weights $\boldsymbol{w}$, that can be used for stabilizing the reconstruction of a signal observed through the quantized sensing model (1). We now turn to studying the stability of ℓ_1-based decoders integrating these weighted $\ell_{p,\boldsymbol{w}}$-constraints as data fidelity. We will highlight also the requirements that the sensing matrix must fulfill to ensure this stability. We then then apply this general stability result to additive heteroscedastic GGD noise, where weighing can be view as a variance stabilization transform. Section IV will later instantiate the outcome of this section to the particular case of QCS.

A. Generalized Basis Pursuit DeNoise

Given some positive weights $\boldsymbol{w} \in \mathbb{R}^M$ and $p \geqslant 2$, we study the following general minimization program, coined General Basis Pursuit DeNoise (GBPDN),

$$\Delta_{p,\boldsymbol{w}}(\boldsymbol{y}, \boldsymbol{\Phi}, \epsilon) \;=\; \underset{\boldsymbol{u} \in \mathbb{R}^N}{\operatorname{Argmin}} \|\boldsymbol{u}\|_1 \text{ s.t. } \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{u}\|_{p,\boldsymbol{w}} \leqslant \epsilon, \qquad (\textbf{GBPDN}(\ell_{p,\boldsymbol{w}}))$$

where $\|\cdot\|_{p,\boldsymbol{w}}$ is the weighted ℓ_p-norm defined in the previous section. Note that BPDN is special case of GBPDN corresponding to $p = 2$ and $\boldsymbol{w} = \mathbf{1}$. The Basis Pursuit DeQuantizers (BPDQ) introduced in [8] are associated to $p \geqslant 1$ and $\boldsymbol{w} = \mathbf{1}$, while the case $p = 1$ and $\boldsymbol{w} = \mathbf{1}$ has also been covered in [20].

We are going to see that the stability of GBPDN($\ell_{p,\boldsymbol{w}}$) is guaranteed if $\boldsymbol{\Phi}$ satisfies a particular instance of the following general isometry property.

Definition 1. *Given two normed spaces $\mathcal{X} = (\mathbb{R}^M, \|\cdot\|_{\mathcal{X}})$ and $\mathcal{Y} = (\mathbb{R}^N, \|\cdot\|_{\mathcal{Y}})$ (with $M < N$), a matrix $\boldsymbol{\Phi} \in \mathbb{R}^{M \times N}$ satisfies the Restricted Isometry Property from $\mathcal{X}$ to $\mathcal{Y}$ at order $K \in \mathbb{N}$, radius $0 \leqslant \delta < 1$ and for a normalization $\mu > 0$, if for all $\boldsymbol{x} \in \Sigma_K$,*

$$(1-\delta)^{1/\kappa} \|\boldsymbol{x}\|_{\mathcal{Y}} \leqslant \tfrac{1}{\mu}\|\boldsymbol{\Phi}\boldsymbol{x}\|_{\mathcal{X}} \leqslant (1+\delta)^{1/\kappa} \|\boldsymbol{x}\|_{\mathcal{Y}}, \qquad (14)$$

κ being an exponent function of the geometries of $\mathcal{X}, \mathcal{Y}$. To lighten notation, we will write that $\boldsymbol{\Phi}$ is RIP$_{\mathcal{X},\mathcal{Y}}(K, \delta, \mu)$.

We may notice that the common RIP is equivalent to[3] RIP$_{\ell_2^M,\ell_2^N}(K, \delta, 1)$ with $\kappa = 1$, while the RIP$_{p,q}$ introduced earlier in [8] is equivalent to RIP$_{\ell_p^M,\ell_q^N}(K, \delta, \mu)$ with $\kappa = q$ and μ depending only on M, p and q. Moreover, the RIP$_{p,K,\delta'}$ defined in [21] is equivalent to the RIP$_{\ell_p^M,\ell_p^N}(K, \delta, \mu)$ with $\kappa = 1$, $\delta' = 2\delta/(1-\delta)$ and $\mu = 1/(1-\delta)$. Finally, the Restricted p-Isometry Property proposed in [22] is also equivalent to the RIP$_{\ell_p^M,\ell_2^N}(K, \delta, 1)$ with $\kappa = p$.

In order to study the behavior of the GBPDN program, we are interested in the embedding induced by $\boldsymbol{\Phi}$ in (14) of $\mathcal{Y} = \ell_2^N$ into the normed space $\mathcal{X} = \ell_{p,\boldsymbol{w}}^M = (\mathbb{R}^M, \|\cdot\|_{p,\boldsymbol{w}})$, *i.e.*, we consider the RIP$_{\ell_{p,\boldsymbol{w}}^M,\ell_2^N}$ property that we write in the following as RIP$_{p,\boldsymbol{w}}$. The following theorem establishes that GBPDN provides stable recovery from distorted measurements, if the RIP$_{p,\boldsymbol{w}}$ holds.

[3]Assuming the columns of $\boldsymbol{\Phi}$ are normalized to unit-norm.

Theorem 1. *Let $K \geqslant 0$, $2 \leqslant p < \infty$ and $\boldsymbol{\Phi} \in \mathbb{R}^{M \times N}$ be a RIP$_{p,\boldsymbol{w}}(s, \delta_s, \mu)$ matrix for $s \in \{K, 2K, 3K\}$ such that*

$$\delta_{2K} + \sqrt{(1+\delta_K)(\delta_{2K}+\delta_{3K})(p-1)} < 1/3 \ . \tag{15}$$

Then, for any signal $\boldsymbol{x} \in \mathbb{R}^N$ observed according to the noisy sensing model $\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{x} + \boldsymbol{\varepsilon}$ with $\|\boldsymbol{\varepsilon}\|_{p,\boldsymbol{w}} \leqslant \epsilon$, the unique solution $\boldsymbol{x}^* = \Delta_{p,\boldsymbol{w}}(\boldsymbol{y}, \boldsymbol{\Phi}, \epsilon)$ obeys

$$\|\boldsymbol{x}^* - \boldsymbol{x}\| \ \leqslant \ 4\, e_0(K) \ + \ 8\, \epsilon/\mu, \tag{16}$$

where $e_0(K) = K^{-\frac{1}{2}} \|\boldsymbol{x} - \boldsymbol{x}_K\|_1$ is the K-term ℓ_1-approximation error.

Proof: If $\boldsymbol{\Phi}$ is RIP$_{p,\boldsymbol{w}}(s, \delta_s, \mu)$ for $s \in \{K, 2K, 3K\}$, then, by definition of the weighted $\ell_{p,\boldsymbol{w}}$-norm, $\mathrm{diag}(\boldsymbol{w})\boldsymbol{\Phi}$ is RIP$_{\ell_p^M, \ell_2^N}(s, \delta_s, \mu)$. Since $\Delta_{p,\boldsymbol{w}}(\boldsymbol{y}, \boldsymbol{\Phi}, \epsilon) \ = \Delta_p(\mathrm{diag}(\boldsymbol{w})\boldsymbol{y}, \mathrm{diag}(\boldsymbol{w})\boldsymbol{\Phi}, \epsilon)$, the stability results proved in [8, Theorem 2] for GBPDN(ℓ_p) [4] shows that

$$\|\boldsymbol{x} - \boldsymbol{x}^*\| \leqslant A_p\, e_0(K) + B_p\, \tfrac{\epsilon}{\mu},$$

with $A_p = \frac{2(1+C_p-\delta_{2K})}{1-\delta_{2K}-C_p}$, $B_p = \frac{4\sqrt{1+\delta_{2K}}}{1-\delta_{2K}-C_p}$ and $C_p \leqslant \sqrt{(1+\delta_K)(\delta_{2K}+\delta_{3K})(p-1)}$ [8]. It is easy to see that if (15) holds, then $A_p \leqslant 4$ and $B_p \leqslant 8$. ■

As we shall see shortly, this theorem may be used to characterize the impact of measurement corruption due to both additive heteroscedastic GGD noise (Section III-C) as well as those induced by a non-uniform scalar quantization (Section IV). Before detailing these two sensing scenarios, we first address the question of designing matrices satisfying the RIP$_{p,\boldsymbol{w}}$ for $2 \leqslant p < \infty$.

B. Weighted Isometric Mappings

We will describe a random matrix construction that will satisfy the RIP$_{p,\boldsymbol{w}}$ for $1 \leqslant p < \infty$. To quantify when this is possible, we introduce some properties on the positive weights $\boldsymbol{w}$.

Definition 2. *A weight generator $\mathcal{W}$ is a process (random or deterministic) that associates to $M \in \mathbb{N}$ a weight vector $\boldsymbol{w} = \mathcal{W}(M) \in \mathbb{R}^M$. This process is said to be of Converging Moments (CM) if for $p \geqslant 1$ and all $M \geqslant M_0$ for a certain $M_0 > 0$,*

$$\rho_p^{\min} \ \leqslant \ M^{-1/p}\, \|\mathcal{W}(M)\|_p \ \leqslant \ \rho_p^{\max}, \tag{17}$$

[4] Dubbed BPDQ in [8].

where $\rho_p^{\min} > 0$ *and* $\rho_p^{\max} > 0$ *are, respectively, the largest and the smallest values such that* (17) *holds. In other words, a CM generator* $\mathcal{W}$ *is such that* $\|\mathcal{W}(M)\|_p^p = \Theta(M)$*. By extension, we say that the weighting vector* $\boldsymbol{w}$ *has the CM property, if it is generated by some CM weight generator* $\mathcal{W}$*.*

The CM property can be ensured if $\lim_{M\to\infty} M^{-1/p}\,\|\boldsymbol{w}\|_p$ exists, bounded and nonzero. It is also ensured if the weights $\{w_i\}_{1\leqslant i\leqslant M}$ are taken (with repetition) from a finite set of positive values. More generally, if $\{w_i : 1 \leqslant i \leqslant M\}$ are iid random variables, we have $M^{-1}\,\|\boldsymbol{w}\|_p^p = \mathbb{E}|w_1|^p$ almost surely by the SLLN. Notice finally that $\rho_p^{\max} \leqslant \|\boldsymbol{w}\|_\infty = \rho_\infty^{\max}$ since $\|\boldsymbol{w}\|_p^p \leqslant M\|\boldsymbol{w}\|_\infty^p$, and $\rho_p^{\min} \geqslant \min_i |w_i|$.

For a weighting vector $\boldsymbol{w}$ having the CM property, we define also its *weighting dynamic* at moment p as the ratio

$$\theta_p = \big(\tfrac{\rho_\infty^{\max}}{\rho_p^{\min}}\big)^2.$$

We will see later that θ_p directly influences the number of measurements required to guarantee the existence of $\mathrm{RIP}_{p,\boldsymbol{w}}$ random Gaussian matrices.

Given a weight vector $\boldsymbol{w}$, the following lemma (proved in Appendix E) characterizes the expectation of the $\ell_{p,\boldsymbol{w}}$-norm of a random Gaussian vector.

Lemma 5 (Gaussian $\ell_{p,w}$-Norm Expectation). *If* $\boldsymbol{\xi} \sim \mathcal{N}^M(0,1)$ *and if the weights* $\boldsymbol{w}$ *have the CM property, then, for* $1 \leqslant p < \infty$ *and* $\mathcal{Z} \sim \mathcal{N}(0,1)$*,*

$$\big(1 \;+\; 2^{p+1}\,\theta_p^p\,M^{-1}\big)^{\frac{1}{p}-1}\,(\mathbb{E}\|\boldsymbol{\xi}\|_{p,\boldsymbol{w}}^p)^{\frac{1}{p}} \;\leqslant\; \mathbb{E}\|\boldsymbol{\xi}\|_{p,\boldsymbol{w}} \;\leqslant\; (\mathbb{E}\|\boldsymbol{\xi}\|_{p,\boldsymbol{w}}^p)^{\frac{1}{p}} = (\mathbb{E}|\mathcal{Z}|^p)^{1/p}\|\boldsymbol{w}\|_p.$$

In particular, $\mathbb{E}\|\boldsymbol{\xi}\|_{p,\boldsymbol{w}} \simeq_M \nu_p\,\|\boldsymbol{w}\|_p \geqslant \nu_p\,M^{1/p}\,\rho_p^{\min}$*, with* $\nu_p^p := \mathbb{E}|\mathcal{Z}|^p = 2^{p/2}\pi^{-1/2}\Gamma(\frac{p+1}{2})$*.*

With an appropriate modification of [8, Proposition 1], we can now prove the existence of random Gaussian $\mathrm{RIP}_{p,\boldsymbol{w}}$ matrices (see Appendix F).

Proposition 1 ($\mathrm{RIP}_{p,\boldsymbol{w}}$ Matrix Existence). *Let* $\boldsymbol{\Phi} \sim \mathcal{N}^{M\times N}(0,1)$ *and some CM weights* $\boldsymbol{w} \in \mathbb{R}^M$*. Given* $p \geqslant 1$ *and* $0 \leqslant \eta < 1$*, then there exists a constant* $c > 0$ *such that* $\boldsymbol{\Phi}$ *is* $\mathit{RIP}_{p,\boldsymbol{w}}(K,\delta,\mu)$ *with probability higher than* $1-\eta$ *when we have jointly* $M \geqslant 2\,(2\theta_p)^p$*, and*

$$M^{2/\max(2,p)} \;\;\geqslant\;\; c\,\delta^{-2}\,\theta_p\,\big(K\log[e\tfrac{N}{K}(1+12\delta^{-1})] \;+\; \log\tfrac{2}{\eta}\big). \tag{18}$$

Moreover, the value $\mu = \mu(\ell_{p,\boldsymbol{w}}^M, \ell_2^N)$ *in* (14) *is given by* $\mu = \mathbb{E}\|\boldsymbol{\xi}\|_{p,\boldsymbol{w}}$ *for a random vector* $\boldsymbol{\xi} \sim \mathcal{N}^M(0,1)$*.*

The RIP normalizing constant μ can be bounded owing to Lemma 5.

Remark 2. *In the light of Proposition 1, assumption* (15) *becomes reasonable since following the simple argument presented in [8, Appendix B] the saturation of requirement* (18) *implies that* δ_K *decays as* $O(\sqrt{K \log M}/M^{1/p})$ *for RIP*$_{p,\boldsymbol{w}}$ *Gaussian matrices. Therefore, for any value* p*, it is always possible to find a* M *such that* (15) *holds. However, this is only possible for high oversampling situation,* i.e.*, for* $\Omega((K \log N/K)^{p/2})$ *measurements.*

C. GBPDN stabilizes Heteroscedastic GGD Noise

Consider the following general signal sensing model

$$\boldsymbol{y} \;=\; \boldsymbol{\Phi x} \;+\; \boldsymbol{\varepsilon}\;, \tag{19}$$

where $\boldsymbol{\varepsilon} \in \mathbb{R}^M$ is the noise vector. For heteroscedastic GGD noise, each ε_i follows a zero-mean $\mathrm{GGD}(0, \alpha_i, p)$ distribution with pdf $\propto \exp(-|t/\alpha_i|^p)$, where $p > 0$ is the shape parameter (the same for all ε_i's), and $\alpha_i > 0$ the scale parameter [23]. It is obvious that

$$\mathbb{E}\boldsymbol{\varepsilon} = \mathbf{0} \quad \text{and} \quad \mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}) = \Gamma(3/p)(\Gamma(1/p))^{-1}\,\mathrm{diag}(\alpha_1^2, \cdots, \alpha_M^2).$$

If one sets the weights to $w_i = 1/\alpha_i$ in GBPDN($\ell_{p,\boldsymbol{w}}$), it can be seen that the associated constraint corresponds precisely to the negative log-likelihood of the joint pdf of $\boldsymbol{\varepsilon}$. As detailed below, introducing these non-uniform weights w_i leads to a reduction in the error of the reconstructed signal, relative to using constant weights. Without loss of genrality, we here restrict our analysis to strictly K-sparse $\boldsymbol{x} \in \Sigma_K$, and assume knowledge of bounds (estimators) for the ℓ_p and the $\ell_{p,\boldsymbol{w}}$ norms used for characterizing $\boldsymbol{\varepsilon}$, *i.e.*, we know that $\|\boldsymbol{\varepsilon}\|_p \simeq_M \epsilon$ and $\|\boldsymbol{\varepsilon}\|_{p,\boldsymbol{w}} \simeq_M \epsilon_{\mathrm{st}}$ for some $\epsilon, \epsilon_{\mathrm{st}} > 0$ to be detailed later.

In this case, if the random matrix $\boldsymbol{\Phi} \sim \mathcal{N}^{M\times N}(0,1)$ is RIP$_{p,\boldsymbol{w}}(K, \delta, \mu)$ for $p \geqslant 2$, with $\mu = \mathbb{E}\|\boldsymbol{\xi}\|_p$ for $\boldsymbol{\xi} \sim \mathcal{N}^M(0,1)$, Theorem 1 asserts that

$$\|\boldsymbol{x}^* - \boldsymbol{x}\| \;\leqslant\; B_p\,\epsilon/\mu,$$

for $\boldsymbol{x}^* = \Delta_{p,\mathbf{1}}(\boldsymbol{y}, \boldsymbol{\Phi}, \epsilon)$ and $B_p \simeq_M 8$. Conversely, for the weights to $w_i = 1/\alpha_i$, and assuming $\boldsymbol{\Phi}$ being RIP$_{p,\boldsymbol{w}}(K, \delta', \mu_{\mathrm{st}})$ with $\mu_{\mathrm{st}} = \mathbb{E}\|\boldsymbol{\xi}\|_{p,\boldsymbol{w}}$, we get

$$\|\boldsymbol{x}^*_{\mathrm{st}} - \boldsymbol{x}\| \;\leqslant\; B'_p\,\epsilon_{\mathrm{st}}/\mu_{\mathrm{st}},$$

for $\boldsymbol{x}^*_{\mathrm{st}} = \Delta_{p,\boldsymbol{w}}(\boldsymbol{y}, \boldsymbol{\Phi}, \epsilon)$ and $B'_p \simeq_M 8$.

When the number of measurements M is large, using classical GGD absolute moments formula, the two bounds ϵ and $\epsilon_{\rm st}$ can be set close to $\epsilon^p \simeq_M \sum_i \mathbb{E}|\varepsilon_i|^p = \|\boldsymbol{\alpha}\|_p^p/p$ and $\epsilon_{\rm st}^p \simeq_M \sum_i w_i^p\, \mathbb{E}|\varepsilon_i|^p = M/p$. Moreover, using Lemma 5, $\mu^p \simeq_M \sum_i \mathbb{E}|\xi_i|^p = M\mathbb{E}|\mathcal{Z}|^p$ and $\mu_{\rm st}^p \simeq_M \mathbb{E}|\mathcal{Z}|^p\, \|\boldsymbol{w}\|_p^p$, where $\mathcal{Z} \sim \mathcal{N}(0,1)$.

Proposition 2. *For an additive heteroscedastic noise $\boldsymbol{\varepsilon} \in \mathbb{R}^M$ such that $\varepsilon_i \sim_{\rm iid} \mathrm{GGD}(0, \alpha_i, p)$, setting $w_i = 1/\alpha_i$ provides $\epsilon_{\rm st}^p/\mu_{\rm st}^p \ \lesssim_M \ \epsilon^p/\mu^p$. Therefore, asymptotically in M, GBPDN($\ell_{p,\boldsymbol{w}}$) has a smaller reconstruction error compared to GBPDN(ℓ_p) when estimating $\boldsymbol{x}$ from the sensing model* (19)*.*

Proof: Let us observe that $\epsilon_{\rm st}^p/\mu_{\rm st}^p \simeq_M \ M(p\,\mathbb{E}|\mathcal{Z}|^p\, \|\boldsymbol{w}\|_p^p)^{-1} = (p\,\mathbb{E}|\mathcal{Z}|^p)^{-1}(\frac{1}{M}\sum_i \frac{1}{\alpha_i^p})^{-1}$. By the Jensen inequality, $(\frac{1}{M}\sum_i \frac{1}{\alpha_i^p})^{-1} \leqslant \frac{1}{M}\sum_i \alpha_i^p$, so that $\epsilon_{\rm st}^p/\mu_{\rm st}^p \lesssim_M \frac{1}{p}\,(\mathbb{E}|\mathcal{Z}|^p)^{-1}\|\boldsymbol{\alpha}\|_p^p/M = \epsilon^p/\mu^p$. ■

The price to pay for this stabilization is an increase of the weighting dynamic $\theta_p = (\frac{\rho_\infty^{\max}}{\rho_p^{\min}})^2$ defined in Proposition 1, which implies an increase in the number of measurements M needed to ensure that the $\mathrm{RIP}_{p,\boldsymbol{w}}(K,\delta,\mu)$ is satisfied.

Example. *Let us consider a simple situation where the α_i's take only two values,* i.e.*, $\alpha_i \in \{1, H\}$ for some $H \geqslant 1$. Let us assume also that the proportion of α_i's equal to H converges to $r \in [0,1]$ with M as $|\frac{1}{M}\#\{i : \alpha_i = H\} \ - \ r| = O(M^{-1})$. In this case, the stabilizing weights are $w_i = 1/\alpha_i \in \{1, 1/H\}$. An easy computation provides*

$$\mathsf{E} := \tfrac{\epsilon^p}{\mu^p} \ \underset{M}{\simeq} \ \tfrac{1}{p}\nu_p^{-p}\,\big(\,r\,H^p + (1-r)\,\big),$$
$$\mathsf{E}_{\rm st} := \tfrac{\epsilon_{\rm st}^p}{\mu_{\rm st}^p} \ \underset{M}{\simeq} \ \tfrac{1}{p}\nu_p^{-p}\,\big(\,r\,H^{-p} + (1-r)\,\big)^{-1},$$

so that, the "stabilization gain" with respect to an unstabilized setting can be quantified by the ratio

$$(\tfrac{\mathsf{E}}{\mathsf{E}_{\rm st}})^{\frac{1}{p}} \underset{M}{\simeq} \big(\,r\,H^{-p} + (1-r)\,\big)^{\frac{1}{p}}\big(\,r\,H^p + (1-r)\,\big)^{\frac{1}{p}} \underset{M,H}{\simeq} \big(r\,(1-r)\big)^{\frac{1}{p}}\,H.$$

We see that the stabilization provides a clear gain which increases as the measurements get very unevenly corrupted, i.e.*, when H is large. Interestingly, the higher p is, the less sensitive is this gain to r. We also observe that the overhead in the number of measurements between the stabilized and the unstabilized situations is related to*

$$\theta_p^{p/2} \ = \big(\tfrac{\rho_\infty^{\max}}{\rho_p^{\min}}\big)^p \underset{M}{\simeq} \big(\,r\,H^{-p} + (1-r)\,\big)^{-1} \underset{M,H}{\simeq} (1-r)^{-1}.$$

The limit case where $H \gg 1$ can be interpreted as ignoring r percent of the measurements in the data fidelity constraint, keeping only those for which the noise is not dominating. In that case, the sufficient condition (18) *in Proposition 1 for $\boldsymbol{\Phi}$ to be $\mathrm{RIP}_{p,\boldsymbol{w}}$ tends to $\theta_p^{-p/2}M = (1-r)M = \Omega\big((K\log N/K)^{p/2}\big)$*

which is consistent with the fact that on average only fraction $1-r$ of the M measurements significantly participate to the CS scheme, i.e.*, $M' = (1-r)M$ must satisfy the common RIP requirement. For $p = 2$, this is somehow related to the democratic property of RIP matrices [4],* i.e.*, the fact that a reasonable number of rows can be discarded from a matrix while preserving the RIP. This property was successfully used for discarding saturated CS measurements in the case of a limited dynamic quantizer [4].*

IV. Dequantizing with Generalized Basis Pursuit DeNoise

Let us now instantiate the use of GBPDN to the reconstruction of signals in the QCS scenario defined in SectionII. Under the quantization formalism defined in Lemma 3 and for Gaussian matrices $\mathbf{\Phi}$, the factor ϵ/μ in (16) can be shown to decrease as $1/\sqrt{p+1}$ asymptotically in M and B. This asymptotic and *almost sure* result which relies on the SLLN (see Appendix G) suggests increasing p to the highest value allowed by (15) in order to decrease the GBPDN reconstruction error.

Proposition 3 (Dequantizing Reconstruction Error). *Given $\boldsymbol{x} \in \mathbb{R}^N$ and $\mathbf{\Phi} \sim \mathcal{N}^{M\times N}(0,1)$, assume that the entries of $\boldsymbol{z} = \mathbf{\Phi}\boldsymbol{x}$ are* iid *realizations from $\mathcal{Z} \sim \mathcal{N}(0,\sigma_0^2)$. We take the corresponding optimal compressor function $\mathcal{G}$ defined in* (3) *and the p-optimal B-bits scalar quantizer $\mathcal{Q}_p$ as defined in* (8)*. Then, the ratio ϵ/μ given in* (16) *is asymptotically and almost surely bounded by*

$$\tfrac{\epsilon}{\mu} \lesssim_{B,M} c'\, 2^{-B}\, \tfrac{(p+1)^{-\frac{1}{2p}}}{\sqrt{p+1}} \leqslant c'\, \frac{2^{-B}}{\sqrt{p+1}}.$$

with $c' = (9/8)(e\pi/3)^{1/2}$.

Notice that, under HRA and for large M, it is possible to provide a rough estimation of the weighting dynamic θ_p when the weights are those provided by the D$_p$C constraints. Indeed, since $w_i(p) = \mathcal{G}'(\mathcal{Q}_p[y_i])^{(p-2)/p}$ and $\mathcal{G}' = \gamma_{0,\sqrt{3}\sigma_0}$, we find

$$\begin{aligned}
\|\boldsymbol{w}\|_p^p = \sum_i \mathcal{G}'(\mathcal{Q}_p[y_i])^{p-2} &\simeq_M M \sum_k \mathcal{G}'^{p-2}(\omega_{k,p})\, p_k \\
&\simeq_{B,M} M\,(2\pi 3\sigma_0^2)^{(2-p)/2}(2\pi\sigma_0^2)^{-1/2}\sum_k \tau_k\, \exp(-\tfrac{1}{2}\omega_{k,p}^2\, \tfrac{p+1}{3\sigma_0^2}) \\
&\simeq_{B,M} M\,(2\pi 3\sigma_0^2)^{(2-p)/2}(2\pi\sigma_0^2)^{-1/2}(2\pi\tfrac{3\sigma_0^2}{p+1})^{1/2} \\
&= M\,(2\pi\sigma_0^2)^{(2-p)/2}3^{(3-p)/2}(p+1)^{-1/2},
\end{aligned}$$

where we recall that $p_k = \int_{\mathcal{R}_k} \varphi_0(t)\mathrm{d}t \simeq_B \varphi_0(c')\tau_k$, for any $c' \in \mathcal{R}_k$ (see the proof of Lemma 9).

Moreover, using (10) and since one of the two smallest quantization bins is $\mathcal{R}_{\mathcal{B}/2} = [0, \tau_{\mathcal{B}/2})$,

$$\|\boldsymbol{w}\|_\infty^p \simeq_B (\alpha/\tau_{\mathcal{B}/2})^{p-2} = (\alpha/\mathcal{G}^{-1}(1/2+\alpha))^{p-2} \simeq_B (2\pi 3\sigma_0^2)^{(2-p)/2}.$$

Therefore, estimating θ_p^p with $M^2\|\boldsymbol{w}\|_\infty^{2p}/\|\boldsymbol{w}\|_p^{2p}$, we find

$$\theta_p^{p/2} \ \simeq_{B,M} \ \sqrt{(p+1)/3}.$$

Therefore, at a given $p \geqslant 2$, since (18) involves that M evolves like $\Omega(\theta_p^{p/2}(K\log N/K)^{p/2})$, using the weighting induced by GBPDN($\ell_{p,\boldsymbol{w}}$) requires collecting $\sqrt{(p+1)/3}$ times more measurements than GBPDN(ℓ_p) in order to ensure the appropriate RIP$_{p,\boldsymbol{w}}$ property. This represents part of the price to pay for guaranteeing bounded reconstruction error by adapting to non-uniform quantization.

Dequantizing is Stabilizing Quantization Distortion:

In connection with the procedure developed in Section III-C, the weights and the p-optimal levels introduced in Lemma 3 can be interpreted as a "stabilization" of the quantization distortion seen as a heteroscedastic noise. This means that, asymptotically in M, selecting these weights and levels, *all quantization regions $\mathcal{R}_k$ contribute equally to the $\ell_{p,\boldsymbol{w}}$ distortion measure*.

To understand this fact, we start by studying the following relation shown in the proof of Lemma 3 (see Appendix D):

$$\|\mathcal{Q}_p[\boldsymbol{z}] - \boldsymbol{z}\|_{p,\boldsymbol{w}}^p \ \underset{M}{\simeq} \ M \sum_k [\mathcal{G}'(\omega_{k,p})]^{p-2} \int_{\mathcal{R}_k} |t-\omega_{k,p}|^p \, \varphi_0(t)\,\mathrm{d}t. \tag{20}$$

Using the threshold $T(B) = \Theta(\sqrt{B})$ and $\mathcal{T} = [-T(B), T(B)]$ as defined in Lemma 2, the proof of Lemma 9 in Appendix D shows that

$$\|\mathcal{Q}_p[\boldsymbol{z}] - \boldsymbol{z}\|_{p,\boldsymbol{w}}^p \ \underset{M,B}{\simeq} \ M \sum_{k:\mathcal{R}_k\subset\mathcal{T}} [\mathcal{G}'(\omega_{k,p})]^{p-2} \int_{\mathcal{R}_k} |t-\omega_{k,p}|^p \, \varphi_0(t)\,\mathrm{d}t, \tag{21}$$

$$\underset{B}{\simeq} \ M \sum_{k:\mathcal{R}_k\subset\mathcal{T}} [\mathcal{G}'(\omega_{k,p})]^{p-2} \tfrac{\tau_k^{p+1}}{(p+1)\,2^p} \, \varphi_0(\omega_{k,p}), \tag{22}$$

using (9). However, using (10) and the relation $\mathcal{G}' = \varphi_0^{1/3}/\|\!|\varphi_0|\!\|_{1/3}^{1/3}$, we find $\tau_k^3\,\varphi_0(\omega_{k,p}) \simeq_B \alpha^3\,\|\!|\varphi_0|\!\|_{1/3}$. Therefore, each term of the sum in (21) provides a contribution

$$[\mathcal{G}'(\omega_{k,p})]^{p-2} \tfrac{\tau_k^{p+1}}{(p+1)\,2^p} \, \varphi_0(\omega_{k,p}) \ \underset{B,M}{\simeq} \ \|\!|\varphi_0|\!\|_{1/3} \, \tfrac{\alpha^{p+1}}{(p+1)2^p},$$

which is independent of k! This phenomenon is well known for $p=2$ and may actually serve for defining $\mathcal{G}'$ itself [9]. The fact that this effect is preserved for $p \geqslant 2$ is a surprise for us.

V. Numerical Experiments

We first describe how to numerically solve the GBPDN optimization problem using a primal-dual convex optimization scheme, then illustrate the use of GBPDN for stabilizing heteroscedastic Gaussian noise on the CS measurements. Finally, we apply GBPDN for reconstructing signals in the quantized CS scenario described in Section II.

A. Solving GBPDN

The optimization problem GBPDN($\ell_{p,\boldsymbol{w}}$) is a special instance of the general form

$$\min_{\boldsymbol{u}\in\mathbb{R}^N} f(\boldsymbol{u}) + g(\boldsymbol{L}\boldsymbol{u}) \ , \tag{23}$$

where f and g are closed convex functions that are not infinite everywhere (*i.e.*, proper functions), and $\boldsymbol{L} = \mathrm{diag}(\boldsymbol{w})\boldsymbol{\Phi}$ is a bounded linear operator, with $f(\boldsymbol{u}) := \|\boldsymbol{u}\|_1$, and $g(\boldsymbol{v}) := \imath_{\mathbb{B}_p^\epsilon}(\boldsymbol{v} - \boldsymbol{y})$ where $\imath_{\mathbb{B}_p^\epsilon}(\boldsymbol{v})$ is the indicator function of the ℓ_p-ball $\mathbb{B}_p^\epsilon$ centered at zero and of radius ϵ, *i.e.*, $\imath_{\mathbb{B}_p^\epsilon}(\boldsymbol{v}) = 0$ if $\boldsymbol{v} \in \mathbb{B}_p^\epsilon$ and $+\infty$ otherwise. For the case of GBPDN($\ell_{p,\boldsymbol{w}}$), both f and g are non-smooth but the associated proximity operators (to be defined shortly) can be computed easily. This will allow to minimize the GBPDN($\ell_{p,\boldsymbol{w}}$) objective by calling on proximal splitting algorithms.

Before delving into the details of the minimization splitting algorithm, we recall some results from convex analysis. The *proximity operator* [24] of a proper closed convex f is defined as the unique solution

$$\mathrm{prox}_f(\boldsymbol{u}) = \underset{\boldsymbol{z}}{\mathrm{argmin}}\ \tfrac{1}{2}\|\boldsymbol{z} - \boldsymbol{u}\|^2 + f(\boldsymbol{z}).$$

If $f = \imath_C$ for some closed convex set C, prox_f is equivalent to the orthogonal projector onto C, proj_C. f^* is the *Legendre-Fenchel conjugate* of f. For $\lambda > 0$, the proximity operator of λf^* can be deduced from that of f/λ through Moreau's identity

$$\mathrm{prox}_{\lambda f^*}(\boldsymbol{u}) = \boldsymbol{u} - \lambda\,\mathrm{prox}_{\lambda^{-1} f}(\boldsymbol{u}/\lambda) \ .$$

Solving (23) with an arbitrary bounded linear operator $\boldsymbol{L}$ can be achieved using primal-dual methods motivated by the classical Kuhn-Tucker theory. Starting from methods to solve saddle function problems such as the Arrow-Hurwicz method [25], this problem has received a lot of attention recently, *e.g.*, [26–28]. In this paper, we use the relaxed Arrow-Hurwicz algorithm as revitalized recently in [27]. Adapted to our problem, its steps are summarized in Algorithm 1.

Algorithm 1 Primal-dual scheme for solving GBPDN($\ell_{p,\boldsymbol{w}}$).

Inputs: Measurements $\boldsymbol{y}$, sensing matrix $\boldsymbol{\Phi}$, weights $\boldsymbol{w}$.

Parameters: Iteration number N_{iter}, $\theta \in [0,1]$, step-sizes $\sigma > 0$ and $\tau > 0$ with $\tau\sigma\|\boldsymbol{w}\|_\infty^2\|\boldsymbol{\Phi}\|^2 < 1$.

Main iteration:

for $k = 0$ **to** $N_{\mathrm{iter}} - 1$ **do**

- Update the dual variable:

$$\boldsymbol{v}_{k+1} = \mathrm{prox}_{\sigma g^*}(\boldsymbol{v}_k + \sigma \boldsymbol{L}\overline{\boldsymbol{u}}_k) \ .$$

- Update the primal variable:

$$\boldsymbol{u}_{k+1} = \mathrm{prox}_{\tau f}(\boldsymbol{u}_k - \tau \boldsymbol{L}^{\mathrm{T}}\boldsymbol{v}_{k+1}) \ .$$

- Approximate extragradient step:

$$\overline{\boldsymbol{u}}_{k+1} = \boldsymbol{u}_{k+1} + \theta(\boldsymbol{u}_{k+1} - \boldsymbol{u}_k) \ .$$

Output: Signal $\boldsymbol{u}_{N_{\mathrm{iter}}}$.

A sufficient condition for the sequences of Algorithm 1 to converge is to choose σ and τ such that $\tau\sigma\|\boldsymbol{w}\|_\infty^2\|\boldsymbol{\Phi}\|^2 < 1$. It has been shown in [27, Theorem 1] that under this condition and for $\theta = 1$, the primal sequence $(\boldsymbol{u}_k)_{k\in\mathbb{N}}$ converges to a (possibly strict) global minimizer of GBPDN($\ell_{p,\boldsymbol{w}}$), with the rate $O(1/k)$ in ergodic sense on the partial duality gap.

Proximity operator of f: For $f(\boldsymbol{u}) = \|\boldsymbol{u}\|_1$, $\mathrm{prox}_{\tau f}(\boldsymbol{u})$ is the popular component-wise soft-thresholding of $\boldsymbol{u}$ with threshold τ.

Proximity operator of g: Recall that $g(\boldsymbol{v}) = \imath_{\mathbb{B}_p^\epsilon}(\boldsymbol{v}-\boldsymbol{y})$. Using Moreau's identity above, and proximal calculus rules for translation and scaling, we have

$$\mathrm{prox}_{\sigma g^*}(\boldsymbol{v}) = \boldsymbol{v} - \sigma\boldsymbol{y} - \mathrm{proj}_{\imath_{\mathbb{B}_p^{\sigma\epsilon}}}(\boldsymbol{v} - \sigma\boldsymbol{y}) \ .$$

It remains to compute the orthogonal projection $\mathrm{proj}_{\mathbb{B}_p^1}$ to get $\mathrm{proj}_{\mathbb{B}_p^{\sigma\epsilon}} = \sigma\epsilon\,\mathrm{proj}_{\mathbb{B}_p^1}(\cdot/(\sigma\epsilon))$. For $p = 2$ and $p = +\infty$, this projector has an easy closed form. For $2 < p < +\infty$, we used the Newton method we proposed in [8] for solving the related Karush-Kuhn-Tucker system which is reminiscent of the strategy underlying sequential quadratic programming.

B. Gaussian Noise Stabilization Illustration

We explore numerically the impact of using non-uniform weights (*e.g.*, stabilizing the measurement noise) for signal reconstruction when the CS measurements are corrupted by heteroscedastic Gaussian noise, as discussed in Section III-C. This illustrates for $p = 2$ both the gain induced by stabilizing the sensing noise and the increase of measurements necessary for observing this gain.

In this illustration, we set the problem dimensions to $N = 1024$, $K = 16$, and let the oversampling factor be in $M/K \in \{5, 10, \cdots, 50\}$. The K-sparse unit norm signals were generated independently according to a Bernoulli-Gaussian mixture model with K-length support picked uniformly at random in $[N]$, and the non-zero signal entries drawn from $\mathcal{N}(0, \sigma_s^2)$ with $\sigma_s^2 \simeq 1/K$. Noisy measurements were simulated by setting $\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{x} + \boldsymbol{\varepsilon}$, with $\varepsilon_i \sim_{\mathrm{iid}} \mathcal{N}(0, \sigma_i^2)$ and $\boldsymbol{\Phi} \sim \mathcal{N}^{M\times N}(0, 1)$. The heteroscedastic behavior of $\boldsymbol{\varepsilon}$ has been designed so that $\sigma_i \sim_{\mathrm{iid}} \mathcal{U}([\sigma_0 - \delta_0, \sigma_0 + \delta_0])$ with $\sigma_0 = 0.1$ and $\delta_0 = 0.6\,\sigma_0$.

Two reconstruction methods were tested: one *with* and the other *without* stabilizing the noise variance. In the first case, the weights have been set to $w_i = 1/\sigma_i$, while in the second $\boldsymbol{w} = \mathbf{1}$. Since the purpose of this analysis is not focused on the design of efficient noise power estimators, ϵ and ϵ_{st} have been simply set by an oracle to $\epsilon_{\mathrm{st}} = \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x}\|_{2,\boldsymbol{w}}$ and $\epsilon = \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x}\|_2$.

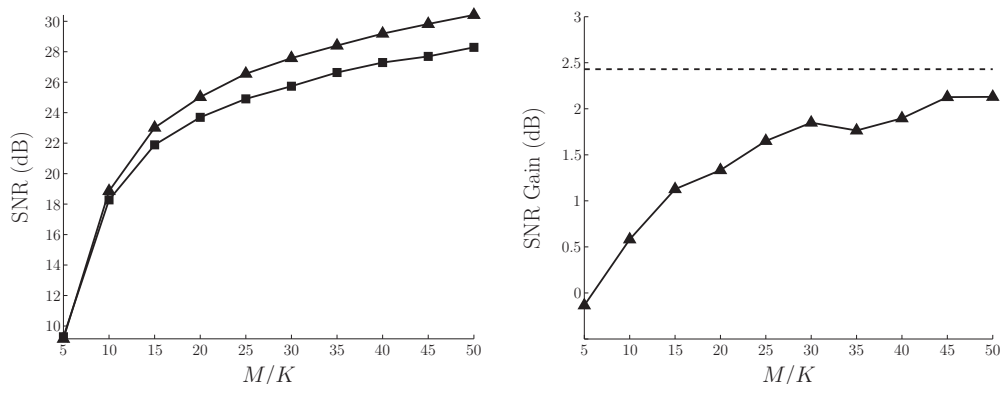
Given the parameters above, we compute the weighting dynamic $\theta_p \simeq_M \frac{M\,\mathbb{E}\|\boldsymbol{w}\|_\infty^2}{\mathbb{E}\|\boldsymbol{w}\|_2^2} = \frac{\sigma_0+\delta_0}{\sigma_0-\delta_0} = 4$, and the average stabilization gain should be (see Proposition 2)

$$20\log_{10}\|\boldsymbol{x} - \boldsymbol{x}^*\|/\|\boldsymbol{x} - \boldsymbol{x}^*_{\mathrm{st}}\| \;\simeq_M\; 20\log_{10}(\epsilon\|\boldsymbol{w}\|)/(\epsilon_{\mathrm{st}}\sqrt{M}) \;< 2.43\,\mathrm{dB}.$$

Numerically, GBPDN($\ell_{2,\boldsymbol{w}}$) and GBPDN(ℓ_2) $\equiv$ BPDN have been solved with the method described in Section V-B until the relative ℓ_2-change in the iterates was smaller than 10^{-6} (with a maximum of 2000 iterations). Reconstruction results were averaged over 50 experiments. In Fig. 2(a), the reconstruction signal-to-noise ratio (SNR) of the stabilized reconstruction is clearly superior to the unstabilized one and this gain increases with increasing oversampling ratio M/K. This SNR gain is displayed in Fig. 2(b). The dashed horizontal line represents the theoretical prediction of 2.43 dB which turns to be an upper-bound on the numerically observed gain.

C. Non-Uniform Quantization

We describe several simulations challenging the power of GBPDN for reconstructing sparse signals from non-uniformly quantized measurements when the weights and the p-optimal levels of Lemma 3 are



(a) (b)

Fig. 2: Stabilized versus unstabilized reconstruction using GBPDN($\ell_{2,\boldsymbol{w}}$) and BPDN respectively. (a) The reconstruction SNR using stabilized (triangles) and unstabilized (squares) methods. (b) Observed (triangles) and theoretically predicted (dashed) SNR gain brought by stabilization.

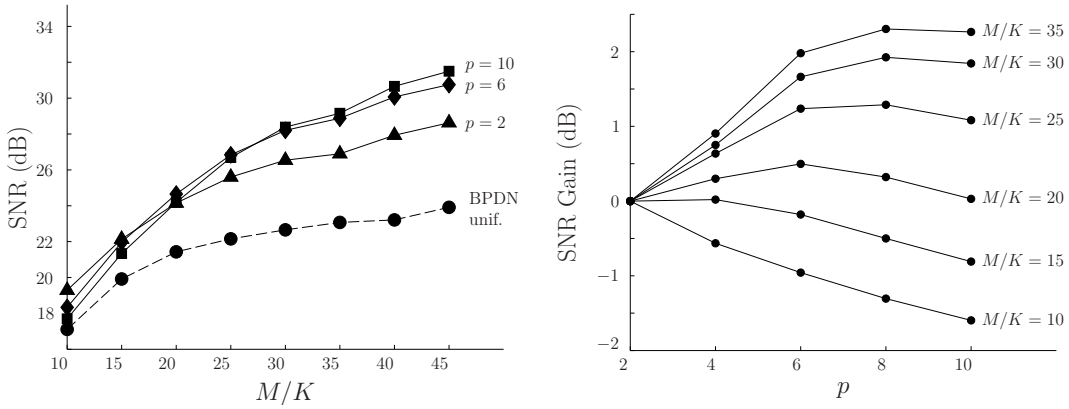
combined. Several configurations have been tested for different $p \geqslant 2$, oversampling ratio M/K, number of bits B and for non-uniform and uniform quantization.

For this experiment, we set the key dimensions to $N = 1024, K = 16, B = 4$, and the K-sparse unit norm signals have been generated as in the previous section. The oversampling ratio was taken as $M/K \in \{10, 15, \cdots, 45\}$, $p \in \{2, 4, \cdots, 10\}$ and the matrix $\boldsymbol{\Phi}$ has been drawn randomly as $\boldsymbol{\Phi} \sim \mathcal{N}^{M\times N}(0, 1)$. The non-uniform quantization of the measurements $\boldsymbol{\Phi x}$ was defined with a compressor $\mathcal{G}$ associated to γ_{0,σ_0} according to (3). The weights $\boldsymbol{w}$ were computed as in Lemma 3, and the p-optimal levels using the numerical method described in Appendix H.

For the sake of completeness, we also compared some results to those obtained for a uniformly quantized CS scenario. In this case, the measurements $\boldsymbol{z} = \boldsymbol{\Phi x}$ are quantized as $y_i = \alpha' \lfloor z_i/\alpha' \rfloor + \alpha'/2$, the quantization bin width $\alpha' = \alpha'(B)$ has been set by dividing regularly the interval $[-\|\boldsymbol{z}\|_\infty, \|\boldsymbol{z}\|_\infty]$ into the same number of bins as those used for the non-uniform quantization.

Again, GBPDN was solved with the primal-dual scheme described in Section V-B until either the relative ℓ_2-change in iterates was smaller than 10^{-6} or a maximum number of iterations of 2000 was reached. Finally, all the reconstruction results were averaged over 50 replications of sparse signals for each combination of parameters.

Fig. 3(a) displays the evolution of the signal reconstruction quality, as measured by the SNR, as a

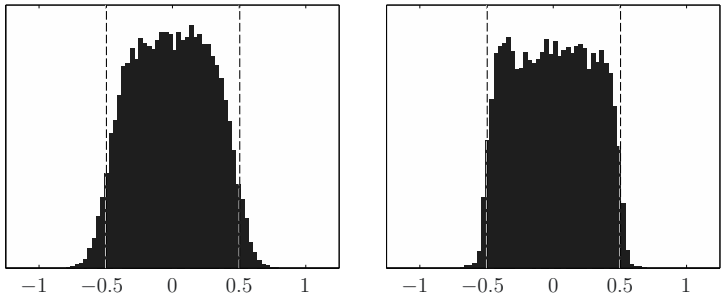


(a) (b)

Fig. 3: Reconstruction SNR of GBPDN($\ell_{p,\boldsymbol{w}}$). (a): the dashed line represents the reconstruction quality achieved from uniformly quantized CS and BPDN. (b) SNR gain versus p for each tested oversampling ratio M/K.

function of the oversampling factor M/K. We clearly see a reconstruction quality improvement with respect to both the uniformly quantized CS scheme (dashed curve) and to increasing values of p and M/K. This last effect is better analyzed in Fig. 3(b) where the SNR gain with respect to $p = 2$ for various values of p is shown. As predicted by Proposition 3, we clearly see that, as soon as the ratio M/K is large, taking higher p value leads to a higher reconstruction quality than the one obtained for $p = 2$ (BPDN). Moreover, Fig. 3(b) confirms that when p increases, the minimal measurement number inducing a positive SNR gain increases. For instance, to achieve a positive gain at $p = 4$, we must have $M/K \geqslant 15$, while at $p = 10$, M/K must be higher than 20. At p fixed, the reconstruction quality increased also monotonically with M/K.

We observe that, given the oversampling ratio, these experimental results allow to increase p to a greater extent than would be allowed by our theory deployed in Section IV. In particular, the sufficient condition (18) dictated by Proposition 1 requires the number of measurements M to scale as $K^{p/2}$ (ignoring times the usual logarithmic terms) in order to ensure the RIP$_{p,\boldsymbol{w}}$. This would imply an exponential increase in the number of measurements needed as p increases. However, from Fig. 3(b), one can see that for $M/K = 15$, $p = 4$ was the largest value before performance starts degrading. With $M/K = 20$, p could be increased to 6 before degradation, and to 8 before degradation with $M/K = 30$. At least for this example, we do not observe such a severe exponential dependence in the needed oversampling in order to benefit from error decrease when increasing p.



(a) (b)

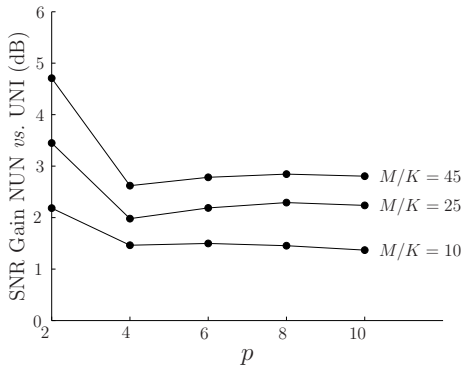
Fig. 4: Testing the Quantization Consistency (QC). (a) Histogram of the components of $\alpha^{-1}\left(\mathcal{G}(\boldsymbol{\Phi}\boldsymbol{x}^*) - \mathcal{G}(\boldsymbol{y})\right)$ for $p = 2$ and $M/K = 40$ (averaged over 100 trials). (b) Same histogram for $p = 10$. The QC is better respected in this case.

In Fig. 4, the quantization consistency of the reconstructed signals is tested by looking at the histogram of $\alpha^{-1}(\mathcal{G}(\boldsymbol{\Phi}\boldsymbol{x}^*) - \mathcal{G}(\boldsymbol{y}))$. We do observe that this histogram is closer to a uniform distribution for $p = 10$ than for $p = 2$, in good agreement with the "companded" quantizer definition $\mathcal{Q} = \mathcal{G}^{-1} \circ \mathcal{Q}_\alpha \circ \mathcal{G}$ showing that in the domain compressed by $\mathcal{G}$, this quantizer is similar to a uniform one.

As a last test, we have more thoroughly compared a uniform quantization scenario described in the experimental setup above with the BPDQ$_p$ decoder developed in [8] to the non-uniform case studied in this paper. More precisely, Fig. 5(a) shows the reconstruction SNR gain between non-uniform and uniform quantization at various p, *i.e.*, SNR(GBPDN($\ell_{p,\boldsymbol{w}}$)) − SNR(BPDQ$_p$). We see that, at a given p, this gain improves with M/K, and the highest SNR improvement values are obtained for $p = 2$. This points the fact that for $p \neq 2$, the quantization scheme is not optimized for reducing the $\ell_{p,\boldsymbol{w}}$-norm distortion. This would require us to change the quantization scenario by not only optimizing the p-optimal levels but also the thresholds. This will be be left to a future research.

VI. Conclusion

In this paper, we have shown that, when the compressive measurements of a sparse or compressible signal are non-uniformly quantized, there is a clear interest in modifying the reconstruction procedure by adapting the way it imposes the reconstructed signal to "match" the observed data. In particular, we have proved that in an oversampled scenario, replacing the common BPDN ℓ_2-norm constraint by a weighted ℓ_p-norm adjusted to the non-uniform nature of the quantizer reduces the reconstruction error by a factor of $\sqrt{p+1}$. Moreover, we showed that this improvement stems from a stabilization of the quantization



(a)

Fig. 5: Reconstruction gain (in dB) between non-uniform or uniform quantization at the same p.

distortion seen as an additive heteroscedastic GGD noise on the measurements.

In future work, we will investigate if the quantization scheme can also be optimized with respect to the proposed reconstruction procedure, *i.e.*, by adjusting the thresholds for minimizing the weighted ℓ_p-distortion at a fixed bit budget.

APPENDIX A
PREPARATORY LEMMATA

This appendix contains several key lemmata that are useful for the subsequent proofs developed in the other appendices.

The first lemma will serve later to evaluate asymptotically the contribution of each quantization bin to the global quantizer distortion measured with $\ell_{p,\boldsymbol{w}}$-norm when a Gaussian source (with pdf φ_0) is quantized.

Lemma 6. *Given $a, b \in \mathbb{R}$ with $a < b$, $n \in \mathbb{N} \setminus \{0\}$ and a Gaussian pdf $\varphi_0 = \gamma_{0,\sigma_0}$. Let λ_n be the (unique) minimizer of $\min_{\lambda\in[a,b]} \int_a^b |t-\lambda|^n \, \varphi_0(t)\,\mathrm{d}t$. Then,*

$$\int_a^b |t-\lambda_n|^n \, \varphi_0(t)\,\mathrm{d}t \;\geqslant\; \tfrac{(b-a)^{n+1}}{(n+1)\,2^{n+1}} \left(1 + \left(\tfrac{\mathsf{D}}{\mathsf{C}}\right)^{-(n+1)/n}\right) \mathsf{C}, \tag{24}$$

$$\int_a^b |t-\lambda_n|^n \, \varphi_0(t)\,\mathrm{d}t \;\leqslant\; \tfrac{(b-a)^{n+1}}{(n+1)\,2^{n+1}} \left(1 + \left(\tfrac{\mathsf{D}}{\mathsf{C}}\right)^{(n+1)/n}\right) \mathsf{D}, \tag{25}$$

$$\tfrac{1}{1+\mathsf{S}^{1/n}}(\mathsf{S}^{1/n}a+b) \;\leqslant\; \lambda_n \;\leqslant\; \tfrac{1}{1+\mathsf{S}^{1/n}}(a+\mathsf{S}^{1/n}b), \tag{26}$$

with $\mathsf{C} := \min_{t\in[a,b]} \varphi_0(t)$, $\mathsf{D} := \max_{t\in[a,b]} \varphi_0(t)$ *and* $\mathsf{S} = \mathsf{D}/\mathsf{C}$.

Proof: Let us first show the upper bound (25). In Lemma 1 and its proof, it was show that λ_n exists and is unique, *i.e.*, the minimization problem is well-posed. Furthermore, λ_n satisfies

$$\mathsf{A} := \int_a^{\lambda_n} (\lambda_n - t)^{n-1}\ \varphi_0(t)\,\mathrm{d}t \;=\; \int_{\lambda_n}^{b} (t-\lambda_n)^{n-1}\ \varphi_0(t)\,\mathrm{d}t.$$

Since $\varphi_0(t) \in [\mathsf{C},\mathsf{D}]$ for $t\in[a,b]$, we have $(\lambda_n - a)^n\,\mathsf{C} \leqslant n\mathsf{A} \leqslant (\lambda_n-a)^n\,\mathsf{D}$ and $(b-\lambda_n)^n\,\mathsf{C} \leqslant n\mathsf{A} \leqslant \frac{1}{n}(b-\lambda_n)^n\,\mathsf{D}$. This implies $(\lambda_n-a)^n \geqslant \big(\frac{\mathsf{C}}{\mathsf{D}}\big)(b-\lambda_n)^n$ and $(b-\lambda_n)^n \geqslant \big(\frac{\mathsf{C}}{\mathsf{D}}\big)(\lambda_n-a)^n$, from which we easily deduce (26).

Since $\int_a^b |t-\lambda_n|^n\ \varphi_0(t)\,\mathrm{d}t \;=\; \int_a^{\lambda_n}(\lambda_n-t)^n\ \varphi_0(t)\,\mathrm{d}t \;+\int_{\lambda_n}^b (t-\lambda_n)^n\ \varphi_0(t)\,\mathrm{d}t$, we find $\int_a^b |t-\lambda_n|^n\ \varphi_0(t)\,\mathrm{d}t \leqslant \frac{1}{n+1}\,[(\lambda_n-a)^{n+1}+(b-\lambda_n)^{n+1}]\,\mathsf{D}$. From $((\lambda_n-a)/(b-\lambda_n))^n \in [\mathsf{C}/\mathsf{D},\mathsf{D}/\mathsf{C}]$, we find that $\int_a^b |t-\lambda_n|^n\ \varphi_0(t)\,\mathrm{d}t$ is smaller than

$$\tfrac{1}{n+1}\min\big(\,(\lambda_n-a)^{n+1},(b-\lambda_n)^{n+1}\big)\left[1+\big(\tfrac{\mathsf{D}}{\mathsf{C}}\big)^{(n+1)/n}\right]\,\mathsf{D}.$$

This provides (25) since $\min(\lambda_n-a, b-\lambda_n) \leqslant (b-a)/2$. The bound (24) is obtained similarly. ■

The following lemma presents a generalization of "Q-function like" bounds for lower partial moments of a Gaussian pdf.

Lemma 7. *Let* $\lambda > 0$, $n\in\mathbb{N}$ *and* $\varphi = \gamma_{0,1}$. *Let us define* $Q_n(\lambda) := \int_\lambda^{+\infty}\ (t-\lambda)^n\varphi(t)\ \mathrm{d}t$. *Then,* $Q_n(\lambda) = \Theta(\lambda^{-(n+1)}\varphi(\lambda))$. *More precisely,* $\frac{n!\,\lambda^{n+1}}{\Pi_{k=1}^{n+1}(\lambda^2+k)}\,\varphi(\lambda) \;\leqslant\; Q_n(\lambda) \;\leqslant\; \frac{n!}{\lambda^{n+1}}\,\varphi(\lambda)$.

This lemma generalizes the well known bound on $Q = Q_0$, namely $\frac{\lambda}{\lambda^2+1}\,\varphi(\lambda) \leqslant Q(\lambda) \leqslant \frac{1}{\lambda}\,\varphi(\lambda)$.

Proof: The proof involves integration by parts, the identities $-\varphi'(u) = u\varphi(u)$ and $(\varphi(u)/u^n)' = (1+\frac{n}{u^2})\frac{\varphi(u)}{u^{n-1}}$. Therefore, the upper bound is a simple consequence of

$$Q_n(\lambda) \leqslant \tfrac{1}{\lambda}\int_\lambda^{+\infty}(t-\lambda)^n\ t\varphi(t)\ \mathrm{d}t = \tfrac{n}{\lambda}\,Q_{n-1}(\lambda) \leqslant \cdots \leqslant \tfrac{n!}{\lambda^n}Q(\lambda) \leqslant \tfrac{n!}{\lambda^{n+1}}\varphi(\lambda).$$

To get the lower bound, observe first that, defining $Q_{n,k}(\lambda) := \int_\lambda^{+\infty}\ (t-\lambda)^n t^{-k}\varphi(t)\ \mathrm{d}t$, we find

$$(1+\tfrac{k+1}{\lambda^2})\,Q_{n,k}(\lambda) \geqslant \int_\lambda^{+\infty}\ (t-\lambda)^n(1+\tfrac{k+1}{t^2})\,t^{-k}\varphi(t)\ \mathrm{d}t = n\,Q_{n-1,k+1}(\lambda).$$

Therefore, $Q_n(\lambda) \geqslant \frac{n\lambda^2}{\lambda^2+1}\,Q_{n-1,1}(\lambda) \geqslant \cdots \geqslant \frac{n!\lambda^{2n}}{\Pi_{k=1}^n(\lambda^2+k)}\,Q_{0,n}(\lambda)$. But $(1+\frac{n+1}{\lambda^2})\,Q_{0,n}(\lambda) \geqslant \varphi(\lambda)/\lambda^{n+1}$, so that $Q_n(\lambda) \geqslant \frac{n!\lambda^{2n+2}}{\Pi_{k=1}^{n+1}(\lambda^2+k)}\,\frac{\varphi(\lambda)}{\lambda^{n+1}}$, which concludes the proof. ■

APPENDIX B
PROOF OF LEMMA 1: "p-OPTIMAL LEVEL DEFINITENESS"

Proof: For $2 \leqslant p < \infty$, $|t-\lambda|^p$ is a continuous, coercive and strictly convex function of λ over $\mathbb{R}$, and therefore so is $\int_{\mathcal{R}_k} |t-\lambda|^p \, \varphi_0(t) \, \mathrm{d}t$ since $\varphi_0(t) > 0$. It follows that the function $\int_{\mathcal{R}_k} |t-\lambda|^p \, \varphi_0(t) \, \mathrm{d}t$ has a unique minimizer on $\mathbb{R}$. Moreover, this minimizer is necessarily located in $\mathcal{R}_k$ since $\int_{\mathcal{R}_k} |t-\lambda|^p \, \varphi_0(t) \, \mathrm{d}t$ is monotonically decreasing (resp. increasing) on $(-\infty, t_k)$ (resp. $(t_{k+1}, +\infty)$)[5]. Consequently, $\omega_{k,n}$ exists and is unique.

For proving the limit case $p \to \infty$, for finite bins $\mathcal{R}_k$ ($k \notin \{1, \mathcal{B}\}$) and without loss of generality for $t_k \geqslant 0$, relation (26) in Lemma 6 with $a = t_k$ and $b = t_{k+1}$, together with the squeeze theorem shows that

$$\lim_{p \to +\infty} \omega_{k,p} = \lim_{p \to +\infty} \tfrac{1}{1+\mathsf{S}^{1/p}}(\mathsf{S}^{1/p} t_k + t_{k+1}) = \lim_{p \to +\infty} \tfrac{1}{1+\mathsf{S}^{1/p}}(t_k + \mathsf{S}^{1/p} t_{k+1}) = \omega_{k,\infty} \ ,$$

where $\mathsf{S} = \varphi_0(t_k)/\varphi_0(t_{k+1})$.

For infinite bins (*i.e.*, $k \in \{1, \mathcal{B}\}$) and assuming again $t_k \geqslant 0$, it follows from the beginning of the proof that $\omega_{k,p}$ is the unique root on $[t_k, +\infty)$ of $\mathcal{E}_p(\lambda) := \int_{t_k}^{\lambda} (\lambda - t)^{p-1} \varphi_0(t) \, \mathrm{d}t - \int_{\lambda}^{\infty} (t - \lambda)^{p-1} \varphi_0(t) \, \mathrm{d}t$. Let $\tilde{\omega}_{k,p} \in [t_k, L]$ be the root of $\tilde{\mathcal{E}}_p(\lambda, L) := \int_{t_k}^{\lambda} (\lambda - t)^{p-1} \varphi_0(t) \, \mathrm{d}t - \int_{\lambda}^{L} (t - \lambda)^{p-1} \varphi_0(t) \, \mathrm{d}t$ for some $L \geqslant t_k$. We then have $\mathcal{E}_p(\tilde{\omega}_{k,p}) = \int_{t_k}^{\tilde{\omega}_{k,p}} (\tilde{\omega}_{k,p} - t)^{p-1} \varphi_0(t) \, \mathrm{d}t - \int_{\tilde{\omega}_{k,p}}^{\infty} (t - \tilde{\omega}_{k,p})^{p-1} \varphi_0(t) \, \mathrm{d}t = -\int_{L}^{\infty} (t - \tilde{\omega}_{k,p})^{p-1} \varphi_0(t) \leqslant 0 = \mathcal{E}_p(\omega_{k,p})$, which implies $\tilde{\omega}_{k,p} \leqslant \omega_{k,p}$ since $\mathcal{E}_p$ is non-decreasing for $p \geqslant 1$. However, since $\tilde{\omega}_{k,p}$ is optimal on $[t_k, L]$, taking $L = L(p) = c\sqrt{p}$, for $c > 0$, we have by Lemma 6 with $a = t_k$ and $b = L(p)$, $\lim_{p \to +\infty} \tilde{\omega}_{k,p} \geqslant \lim_{p \to +\infty} \frac{1}{1+\mathsf{S}^{1/p}}(\mathsf{S}^{1/p} t_k + c\sqrt{p}) = +\infty$ since $\mathsf{S}^{1/p} = \exp(-t_k^2/2p\sigma_0^2) \exp(c^2/2\sigma_0^2) = \Theta(1)$. This proves $\lim_{p \to +\infty} |\omega_{k,p}| = +\infty = \omega_{k,\infty}$ and $|\omega_{k,p}| = \Omega(\sqrt{p})$ for $k \in \{1, \mathcal{B}\}$. ■

APPENDIX C
PROOF OF LEMMA 2: "ASYMPTOTIC p-QUANTIZATION CHARACTERIZATION"

The content of Lemma 2 is derived from this larger set of results which constitutes a *toolbox* lemma for other developments given in these appendices.

Lemma 8 (Extended Asymptotic p-Quantization Characterization). *Given the Gaussian pdf φ_0 and its associated compressor $\mathcal{G}$ function, choose $0 < \beta < 1$ and $p \in \mathbb{N}$, and define $T = T(B) =$*

[5]Where we used the Lebesgue dominated convergence theorem to interchange the integration and derivation signs.

$\sqrt{6\,\sigma_0^2(\log 2^\beta)\,B}$, $\mathcal{T} = [-T, T]$ *and* $\mathcal{T}^c = \mathbb{R} \setminus \mathcal{T}$. *We have the following asymptotic properties (relative to* B*):*

$$\mathcal{G}'(T(B)) = \Theta(2^{-\beta B}), \tag{27}$$

$$\#\{k : \mathcal{R}_k \subset \mathcal{T}^c\} = \Theta\big(B^{-1/2}\, 2^{(1-\beta)B}\big), \tag{28}$$

$$\int_{\mathcal{R}_k} |t - \omega_{k,p}|^p \;\varphi_0(t)\,\mathrm{d}t = O\big(B^{-(p+1)/2}\, 2^{-3\beta B}\big), \quad \forall \mathcal{R}_k \subset \mathcal{T}^c. \tag{29}$$

Moreover, for all k *such that* $\mathcal{R}_k \subset \mathcal{T}$ *and any* $c \in \mathcal{R}_k$

$$\tau_k := t_{k+1} - t_k = \; O(2^{-(1-\beta)B}), \tag{30}$$

$$1 \;\leqslant\; \tfrac{\max(\varphi_0(t_k),\,\varphi_0(t_{k+1}))}{\min(\varphi_0(t_k),\,\varphi_0(t_{k+1}))} = \; \exp\big(O(B^{1/2}\, 2^{-(1-\beta)B})\big) = 1 + O(B^{1/2}\, 2^{-(1-\beta)B}), \tag{31}$$

$$\textstyle\int_{\mathcal{R}_k} |t - \omega_{k,p}|^p \;\varphi_0(t)\,\mathrm{d}t \;\simeq_B\; \tfrac{\tau_k^{p+1}}{(p+1)\,2^p}\,\varphi_0(c), \tag{32}$$

$$\mathcal{G}'(c) \simeq_B \tfrac{\alpha}{\tau_k}. \tag{33}$$

Finally, if k *is such that* $T(B) \in \mathcal{R}_k$*, then, writing the interval length/measure* $\mathcal{L}(\mathcal{A}) = \int_{\mathcal{A}} \mathrm{d}t$ *for* $\mathcal{A} \subset \mathbb{R}$*,*

$$\mathcal{L}(\mathcal{R}_k \cap \mathcal{T}) = \; O(2^{-(1-\beta)B}), \tag{34}$$

$$\mathcal{G}'(\omega_{k,p}) \leqslant \max(\mathcal{G}'(t_k), \mathcal{G}'(t_{k+1})) = \; O(2^{-\beta B}), \tag{35}$$

$$\int_{\mathcal{R}_k} |t - \omega_{k,p}|^p \;\varphi_0(t)\,\mathrm{d}t \;= O\big(B^{-(p+1)/2}\, 2^{-3\beta B}\big). \tag{36}$$

Proof: In this proof we use the quantizer symmetry to restrict the analysis to the half (positive) real line $\mathbb{R}_+$, on which φ_0 is decreasing.

Relation (27) comes from the definition of $T(B)$ and that of $\mathcal{G}' = \gamma_{0,\sqrt{3}\,\sigma_0}$. For proving (28), we can observe that $\mathcal{G}(\lambda) = \|\varphi_0\|_{1/3}^{-1/3} \int_{-\infty}^{\lambda} \varphi_0^{1/3}(t)\,\mathrm{d}t = 1 - Q(\lambda/\sqrt{3}\sigma_0)$ where $Q(t) = \frac{1}{\sqrt{2\pi}} \int_t^{+\infty} \gamma_{0,1}(u)\,\mathrm{d}u$. Since $\frac{\lambda}{1+\lambda^2}\gamma_{0,1}(\lambda) \leqslant Q(\lambda) \leqslant \frac{1}{\lambda}\gamma_{0,1}(\lambda)$, we obtain

$$\tfrac{3\sigma_0^2\lambda}{3\sigma_0^2+\lambda^2}\,\mathcal{G}'(\lambda) \;\leqslant\; 1 - \mathcal{G}(\lambda) \;\leqslant\; \tfrac{3\sigma_0^2}{\lambda}\,\mathcal{G}'(\lambda).$$

Taking $\lambda = T(B)$ in the last inequalities and using (27), we deduce from the quantizer definition

$$\#\{k : \mathcal{R}_k \subset \mathcal{T}^c\} \;=\; 2\,\#\{k : t_k \geqslant T(B)\} \;=\; 2\,\alpha^{-1}\,(1 - \mathcal{G}(T)) \;= \Theta\big(B^{-1/2}\, 2^{(1-\beta)B}\big).$$

Relation (29) is proved by noting that, if $t_k \geqslant T(B)$,

$$\int_{\mathcal{R}_k} |t - \omega_{k,p}|^p \;\varphi_0(t)\,\mathrm{d}t \leqslant \int_{\mathcal{R}_k} (t - t_k)^p \;\varphi_0(t)\,\mathrm{d}t \leqslant \int_{t_k}^{\infty} (t - t_k)^p \;\varphi_0(t)\,\mathrm{d}t,$$

where the first inequality follows from the p-optimality of $\omega_{k,p} \in \mathcal{R}_k$. However, from Lemma 7, we know that, for $\lambda \in \mathbb{R}_+$

$$\tfrac{p!\,\lambda^{p+1}\sigma_0^{2p+2}}{\Pi_{k=1}^{p+1}(\lambda^2+k\sigma_0^2)}\,\varphi_0(\lambda) \;\leqslant\; \sigma_0^p\,Q_p(\tfrac{\lambda}{\sigma_0}) \;\leqslant\; \tfrac{p!\,\sigma_0^{2p+2}}{\lambda^{p+1}}\,\varphi_0(\lambda),$$

with $Q_p(\lambda) := \int_\lambda^\infty (t-\lambda)^p\,\gamma_{0,1}(t)\;\mathrm{d}t$ and $\sigma_0^p\,Q_p(\frac{\lambda}{\sigma_0}) = \int_\lambda^\infty (t-\lambda)^p\,\varphi_0(t)\;\mathrm{d}t$.

Therefore, since $\varphi_0 \propto (\mathcal{G}')^3$,

$$\int_{t_k}^\infty (t-t_k)^p\;\varphi_0(t)\,\mathrm{d}t \leqslant\; \tfrac{p!\,\sigma_0^{2(p+1)}}{t_k^{p+1}}\,\varphi_0(t_k) \leqslant \tfrac{p!\,\sigma_0^{2(p+1)}}{T^{p+1}}\,\varphi_0(T) = O\big(B^{-(p+1)/2}\,2^{-3\beta B}\big).$$

Relation (30) is obtained by observing that $\mathcal{G}$ is concave on $\mathbb{R}_+$. This implies $\tau_k \leqslant \alpha/\mathcal{G}'(t_{k+1})$ and if k is such that $0 \leqslant t_{k+1} \leqslant T(B)$, $\tau_k = O(2^{-(1-\beta)B})$. For (31), keeping the same k, we note that $1 \leqslant \frac{\varphi_0(t_k)}{\varphi_0(t_{k+1})} = \exp(\frac{1}{6\sigma_0^2}\tau_k(t_k+t_{k+1})) \leqslant \exp(\frac{1}{3\sigma_0^2}\tau_k t_{k+1}) = \exp\big(O(B^{1/2}\,2^{-(1-\beta)B})\big)$ which is then arbitrarily close to 1.

For proving (32), we assume first $p \geqslant 1$. Let us consider (24) and (25) with $a = t_k$, $b = t_{k+1}$, $\mathsf{C} = \varphi_0(t_{k+1})$ and $\mathsf{D} = \varphi_0(t_k)$ with $0 \leqslant t_{k+1} \leqslant T(B)$. From (31) we see that $1 \leqslant \frac{\mathsf{D}}{\mathsf{C}} = 1 + o(1)$. We show easily that this involves the equivalent relations $\mathsf{C} \simeq_B \mathsf{D}$, $\mathsf{C}/\mathsf{D} \simeq_B 1$ and $\mathsf{D}/\mathsf{C} \simeq_B 1$. Therefore, $(1 + (\mathsf{D}/\mathsf{C})^{(p+1)/p}) \simeq_B 2$ and $(1 + (\mathsf{C}/\mathsf{D})^{(p+1)/p}) \simeq_B 2$. Moreover, $\mathsf{C} \simeq_B \varphi_0(c)$ and $\mathsf{D} \simeq_B \varphi_0(c)$ for any $c \in \mathcal{R}_k$, so that (24) and (25)) show finally $\int_{\mathcal{R}_k} |t-\omega_{k,p}|^p\;\varphi_0(t)\,\mathrm{d}t \lesssim_B \frac{\tau_k^{p+1}}{(p+1)\,2^p}\,\varphi_0(c)$ and $\int_{\mathcal{R}_k} |t-\omega_{k,p}|^p\;\varphi_0(t)\,\mathrm{d}t \gtrsim_B \frac{\tau_k^{p+1}}{(p+1)\,2^p}\,\varphi_0(c)$, which proves the relation. The case $p = 0$ is demonstrated similarly by observing that $\varphi_0(t_{k+1})\tau_k \leqslant p_k := \int_{\mathcal{R}_k}\varphi_0(t)\,\mathrm{d}t \leqslant \varphi_0(t_k)\tau_k$.

Let's now turn to showing (33). From (31) and since $\mathcal{G}' \propto \varphi_0^{1/3}$, $1 \leqslant \mathcal{G}'(t_k)/\mathcal{G}'(t_{k+1}) = 1+o(1)$ so that $\mathcal{G}'(t_k)/\mathcal{G}'(t_{k+1}) \simeq_B 1$. By concavity of $\mathcal{G}$ on $\mathbb{R}_+$, we know that $\mathcal{G}'(t_{k+1}) \leqslant \alpha/\tau_k \leqslant \mathcal{G}'(t_k)$. Therefore, $1 \leqslant (\mathcal{G}'(t_{k+1}))^{-1}\alpha/\tau_k = 1+o(1)$ which yields $\mathcal{G}'(t_{k+1}) \simeq_B \alpha/\tau_k$. By the concavity argument again, we have $\mathcal{G}'(t_k) \geqslant \mathcal{G}'(c) \geqslant \mathcal{G}'(t_{k+1})$ for any $c \in \mathcal{R}_k$, and thus $1+o(1) = \mathcal{G}'(t_k)/\mathcal{G}'(t_{k+1}) \geqslant \mathcal{G}'(c)/\mathcal{G}'(t_{k+1}) \geqslant 1$. This implies $\mathcal{G}'(c) \simeq_B \mathcal{G}'(t_{k+1}) \simeq_B \alpha/\tau_k$.

If k is such that $0 \leqslant t_k \leqslant T(B) \leqslant t_{k+1}$, using again the concavity of $\mathcal{G}$ on $\mathbb{R}_+$, we find $\mathcal{L}(\mathcal{R}_k \cap \mathcal{T}) = T(B) - t_k \leqslant (\mathcal{G}(T(B)) - k\alpha)/\mathcal{G}'(T(B)) \leqslant \alpha/\mathcal{G}'(T(B)) = O(2^{-(1-\beta)B})$, which proves (34).

For showing (35), we note that $\mathcal{G}'(t_k) = \mathcal{G}'(T)(\mathcal{G}'(t_k)/\mathcal{G}'(T))$. Since $\mathcal{G}'(t_k)/\mathcal{G}'(T) = \exp(\frac{1}{6\sigma_0^2}(T-t_k)(T+t_k)) \leqslant \exp(\frac{1}{3\sigma_0^2}(T-t_k)T) = \exp(O(B^{1/2}\,2^{-(1-\beta)B}))$ which is arbitrarily close to 1 (*i.e.*, it is $e^{o(1)}$), we find $\mathcal{G}'(t_k) = O(2^{-\beta B})$, *i.e.*, it inherits the behavior of $\mathcal{G}'(T)$.

The last relation (36) is proved similarly to (29) by appealing again to Lemma 7,

$$\int_{\mathcal{R}_k} (t-t_k)^p \ \varphi_0(t)\,\mathrm{d}t \ \leqslant \ \int_{t_k}^{\infty} (t-t_k)^p \ \varphi_0(t)\,\mathrm{d}t \leqslant \ \tfrac{p!\,\sigma_0^{2p+2}}{t_k^{p+1}}\,\varphi_0(t_k) = O\big(B^{-(p+1)/2}\,2^{-3\beta B}\big),$$

where the asymptotic relation is obtained by seeing that, as soon as $T - t_k \leqslant 1/2$ (which is always possible to meet thanks to (34)),

$$\tfrac{1}{t_k} = \tfrac{1}{T}(1-\tfrac{T-t_k}{T})^{-1} \leqslant \tfrac{1}{T}\,(1+2\tfrac{T-t_k}{T}),$$

and $\varphi_0(t_k) = O(2^{-3\beta B})$ since $\varphi_0 \propto (\mathcal{G}')^3$. ∎

APPENDIX D
PROOF OF LEMMA 3: "ASYMPTOTIC WEIGHTED ℓ_p-DISTORTION"

Before proving Lemma 3, let us show the following asymptotic equivalence.

Lemma 9. *Let $p \in \mathbb{N} \setminus \{0\}$ and $\gamma > p-3$.*

$$\sum_{k=1}^{\mathcal{B}} [\mathcal{G}'(\omega_{k,p})]^\gamma \int_{\mathcal{R}_k} |t-\omega_{k,p}|^p\,\varphi_0(t)\,\mathrm{d}t \ \simeq_B \ \tfrac{2^{-pB}}{(p+1)\,2^p} \int_{\mathbb{R}} [\mathcal{G}'(t)]^{\gamma-p}\varphi_0(t)\,\mathrm{d}t, \tag{37}$$

Proof: Let us use the threshold $T(B)$ defined in Lemma 8 for splitting the sum (37) in two parts, *i.e.*, using the quantizer symmetry,

$$\sum_{k=1}^{\mathcal{B}} [\mathcal{G}'(\omega_{k,p})]^\gamma \int_{\mathcal{R}_k} |t-\omega_{k,p}|^p\,\varphi_0(t)\,\mathrm{d}t = \ 2 \sum_{k:\ 0\leqslant t_{k+1}<T(B)} [\mathcal{G}'(\omega_{k,p})]^\gamma \int_{\mathcal{R}_k} |t-\omega_{k,p}|^p\,\varphi_0(t)\,\mathrm{d}t \quad + \quad \mathsf{R},$$

where the residual R reads

$$\begin{aligned}\mathsf{R} &:= \ 2 \sum_{k:\ t_{k+1}\geqslant T(B)} [\mathcal{G}'(\omega_{k,p})]^\gamma \int_{\mathcal{R}_k} |t-\omega_{k,p}|^p\,\varphi_0(t)\,\mathrm{d}t,\\ &= \ 2\,[\mathcal{G}'(\omega_{k',p})]^\gamma \int_{\mathcal{R}_{k'}} |t-\omega_{k',p}|^p\,\varphi_0(t)\,\mathrm{d}t \ + \ 2 \sum_{k:\ t_k\geqslant T(B)} [\mathcal{G}'(\omega_{k,p})]^\gamma \int_{\mathcal{R}_k} |t-\omega_{k,p}|^p\,\varphi_0(t)\,\mathrm{d}t,\end{aligned}$$

where k' is such that $t_{k'} < T(B) \leqslant t_{k'+1}$.

From Lemma 8, we can easily bound this residual. We know from (27), (29), (35) and (36) that, for all $k \in \ \{j : \omega_{j,p} \geqslant t_j \geqslant T(B)\} \cup \{k'\}$,

$$[\mathcal{G}'(\omega_{k,p})]^\gamma \int_{\mathcal{R}_k} |t-\omega_{k,p}|^p\,\varphi_0(t)\,\mathrm{d}t = O(2^{-\beta(\gamma+3)B} B^{-(p+1)/2}).$$

However, (28) tells us that the sum in R is made of no more than $1+O\big(B^{-1/2}\,2^{(1-\beta)B}\big) = O\big(B^{-1/2}\,2^{(1-\beta)B}\big)$ terms, so that

$$\mathsf{R} = O\big(B^{-(p+2)/2}\,2^{-(\beta(\gamma+4)-1)B}\big).$$

Let us now study the terms for which $0 \leqslant t_{k+1} \leqslant T(B)$. Using (32) and (33) provides

$$\begin{aligned}
&\sum_{k=1}^{\mathcal{B}} [\mathcal{G}'(\omega_{k,p})]^\gamma \int_{\mathcal{R}_k} |t-\omega_{k,p}|^p \, \varphi_0(t)\,\mathrm{d}t \\
&\underset{B}{\simeq} \; 2 \sum_{k:\, 0 \leqslant t_{k+1} \leqslant T(B)} [\mathcal{G}'(\omega_{k,p})]^\gamma \tfrac{\tau_k^{p+1}}{(p+1)\,2^p} \, \varphi_0(\omega_{k,p}) \;+\; \mathsf{R} \\
&\underset{B}{\simeq} \; 2 \tfrac{\alpha^p}{(p+1)\,2^p} \sum_{k:\, 0 \leqslant t_{k+1} \leqslant T(B)} [\mathcal{G}'(\omega_{k,p})]^{\gamma-p} \, \varphi_0(\omega_{k,p})\, \tau_k \;+\; \mathsf{R} \\
&\underset{B}{\simeq} \; 2 \tfrac{2^{-pB}}{(p+1)\,2^p} \int_0^{T(B)} [\mathcal{G}'(t)]^{\gamma-p} \varphi_0(t)\,\mathrm{d}t \;+\; \mathsf{R},
\end{aligned}$$

where, knowing that $0 \leqslant t_{k+1} \leqslant T(B)$, we have also used (32) with $p = 0$ to see that $p_k = \int_{\mathcal{R}_k} \varphi_0(t)\,\mathrm{d}t \simeq_B \varphi_0(c')\tau_k$ for any $c' \in \mathcal{R}_k$.

Therefore, provided that $\beta(\gamma+4) \geqslant p+1$, which means that $\gamma > p-3$ since $\beta < 1$, the residual R decreases faster than the first term in the right-hand side of last of the last equivalence relation, so that

$$\sum_{k=1}^{\mathcal{B}} [\mathcal{G}'(\omega_{k,p})]^\gamma \int_{\mathcal{R}_k} |t-\omega_{k,p}|^p \, \varphi_0(t)\,\mathrm{d}t \;\underset{B}{\simeq}\; \tfrac{2^{-pB}}{(p+1)\,2^p} \int_{\mathbb{R}} [\mathcal{G}'(t)]^{\gamma-p} \varphi_0(t)\,\mathrm{d}t,$$

since $T(B) = \Theta(B^{1/2})$ by definition. ∎

With the three previous lemmata under our belts, we are now ready to prove Lemma 3.

Proof of Lemma 3: For $z_i \sim_{\text{iid}} \mathcal{N}(0,\sigma_0^2)$ with pdf φ_0, using the SLLN applied to z_i conditionally on each quantization bin, we have

$$\begin{aligned}
\|\mathcal{Q}_p[\boldsymbol{z}] - \boldsymbol{z}\|_{p,\boldsymbol{w}}^p &:= \sum_{i=1}^{M} [\mathcal{G}'(\mathcal{Q}_p[z_i])]^{p-2} \, |z_i - \mathcal{Q}_p[z_i]|^p, \\
&\underset{M}{\simeq} \; M \sum_{k=1}^{\mathcal{B}} [\mathcal{G}'(\omega_{k,p})]^{p-2} \int_{\mathcal{R}_k} |t-\omega_{k,p}|^p \, \varphi_0(t)\,\mathrm{d}t,
\end{aligned}$$

where we used implicitly the quantizer symmetry in the last relation. This last relation is characterized by Lemma 9 by taking $n = p$ and $\gamma = p-2 > p-3$, so that

$$\begin{aligned}
\|\mathcal{Q}_p[\boldsymbol{z}] - \boldsymbol{z}\|_{p,\boldsymbol{w}}^p &\underset{M,B}{\simeq} \; M \, \tfrac{2^{-pB}}{(p+1)\,2^p} \int_{\mathbb{R}} [\mathcal{G}'(t)]^{-2} \varphi_0(t)\,\mathrm{d}t, \\
&\underset{M,B}{\simeq} \; M \, \tfrac{2^{-pB}}{(p+1)2^p} \, |\!|\!| \varphi_0 |\!|\!|_{1/3}.
\end{aligned}$$

∎

APPENDIX E
PROOF OF LEMMA 5: "GAUSSIAN $\ell_{p,w}$-NORM EXPECTATION"

First, the inequality $\mathbb{E}\|\boldsymbol{\xi}\|_{p,\boldsymbol{w}} \leqslant (\mathbb{E}\|\boldsymbol{\xi}\|_{p,\boldsymbol{w}}^p)^{1/p}$ follows from the Jensen inequality applied on the convex function $(\cdot)^p$ on $\mathbb{R}_+$. Second, from our result in [8, Appendix C] it is easy to show that

$$\mathbb{E}\|\boldsymbol{\xi}\|_{p,\boldsymbol{w}} \geqslant (\mathbb{E}\|\boldsymbol{\xi}\|_{p,\boldsymbol{w}}^p)^{1/p} \left(1 \ + \ (\mathbb{E}\|\boldsymbol{\xi}\|_{p,\boldsymbol{w}}^p)^{-2} \operatorname{Var}\|\boldsymbol{\xi}\|_{p,\boldsymbol{w}}^p\right)^{\frac{1}{p}-1}.$$

Moreover, $\mathbb{E}\|\boldsymbol{\xi}\|_{p,\boldsymbol{w}}^p = \|\boldsymbol{w}\|_p^p \, \mathbb{E}|\mathcal{Z}|^p$, while

$$\operatorname{Var}\|\boldsymbol{\xi}\|_{p,\boldsymbol{w}}^p \ = \ \sum_i \operatorname{Var}|w_i\mathcal{Z}|^p = \|\boldsymbol{w}\|_{2p}^{2p} \operatorname{Var}|\mathcal{Z}|^p.$$

Therefore, assuming CM weights,

$$\begin{aligned}\mathbb{E}\|\boldsymbol{\xi}\|_{p,\boldsymbol{w}}/(\mathbb{E}\|\boldsymbol{\xi}\|_{p,\boldsymbol{w}}^p)^{1/p} &\geqslant \ \left(1 \ + \ (\rho_{2p}^{\max}/\rho_p^{\min})^{2p} M^{-1} (\mathbb{E}|\mathcal{Z}|^p)^{-2} \operatorname{Var}|\mathcal{Z}|^p\right)^{\frac{1}{p}-1} \\ &\geqslant \ \left(1 \ + \ 2^{p+1}\, \theta_p^p \, M^{-1}\right)^{\frac{1}{p}-1},\end{aligned}$$

since $\rho_{2p}^{\max} \leqslant \rho_\infty^{\max}$, and $(\mathbb{E}|\mathcal{Z}|^p)^{-2} \operatorname{Var}|\mathcal{Z}|^{2p} < 2^{p+1}$ [8].

APPENDIX F
PROOF OF PROPOSITION 1: "RIP$_{p,w}$ MATRIX EXISTENCE"

The proof proceeds simply by considering the Lipschitz function $F(\boldsymbol{u}) = \|\boldsymbol{u}\|_{p,\boldsymbol{w}}$ and the expected value $\mu = F(\boldsymbol{\xi})$ for a random vector $\boldsymbol{\xi} \sim \mathcal{N}^M(0,1)$ in [8, Appendix A]. The Lipschitz constant of F is

$$\lim_{\boldsymbol{u} \to_{\neq} \boldsymbol{v}} \left|F(\boldsymbol{u}) - F(\boldsymbol{v})\right| / \|\boldsymbol{u} - \boldsymbol{v}\| \ = \ \|\boldsymbol{w}\|_\infty \, \lambda_p,$$

with $\lambda_p = \max(M^{(2-p)/2p}, 1)$ for $p \geqslant 1$. The value $\mu = \mathbb{E}\|\boldsymbol{\xi}\|_{p,\boldsymbol{w}}$ can be estimated thanks to Lemma 5. Indeed, it tells us that if $M \geqslant 2(2\theta_p)^p$,

$$\mu \ \geqslant \ \tfrac{1}{2}(\mathbb{E}\|\boldsymbol{\xi}\|_{p,\boldsymbol{w}}^p)^{1/p} \geqslant \ \tfrac{1}{2}\, \rho_p^{\min} \nu_p \, M^{1/p},$$

with $\nu_p^p = \mathbb{E}|\mathcal{Z}|^p = 2^{p/2}\pi^{-1/2}\Gamma(\frac{p+1}{2})$.

Inserting these results in [8, Appendix A], it is easy to show that a matrix $\boldsymbol{\Phi} \sim \mathcal{N}^{M\times N}(0,1)$ is RIP$_{p,\boldsymbol{w}}(K,\delta,\mu)$ with a probability higher than $1-\eta$ if

$$M^{2/\max(2,p)} \ \geqslant \ c\left(\tfrac{\rho_\infty^{\max}}{\delta\, \rho_p^{\min}}\right)^2 \left(K \log[e\tfrac{N}{K}(1+12\delta^{-1})] \ + \ \log\tfrac{2}{\eta}\right),$$

for some constant $c > 0$.

APPENDIX G
PROOF OF PROPOSITION 3: DEQUANTIZING RECONSTRUCTION ERROR

Proof: We have to bound $\epsilon_p / \mathbb{E}\|\boldsymbol{\xi}\|_{p,\boldsymbol{w}}$, with $\boldsymbol{\xi} \sim \mathcal{N}^M(0,1)$, when M is large and under the HRA. First, according to Lemma 5, using the SLLN and using the same decomposition than in the proof of Lemma 3 with the threshold $T(B)$ (with $\beta = (p+1)/(p+2)$) and the bounds provided by Lemma 8, we find almost surely

$$\begin{aligned}\mu^p \;:=\; (\mathbb{E}\|\boldsymbol{\xi}\|_{p,\boldsymbol{w}})^p \;&\underset{M}{\simeq}\; \sum_{i=1}^{M} [\mathcal{G}'(\mathcal{Q}_p[z_i])]^{p-2} \mathbb{E}|\mathcal{Z}|^p \\ &\underset{M}{\simeq}\; M\,\mathbb{E}|\mathcal{Z}|^p \sum_{k:\, t_k \geqslant 0} p_k\, [\mathcal{G}'(\omega_{k,p})]^{p-2}.\end{aligned}$$

The sum in the last expression is characterized by Lemma 9 by setting inside (37) $n = 0$ and $\gamma = p-2$. This provides

$$\begin{aligned}\mu^p \;&\underset{M,B}{\simeq}\; M\,\mathbb{E}|\mathcal{Z}|^p \int_{\mathbb{R}} [\mathcal{G}'(t)]^{p-2} \varphi_0(t)\,\mathrm{d}t \\ &\underset{M,B}{\simeq}\; M\,\mathbb{E}|\mathcal{Z}|^p \,\big[\int_{\mathbb{R}} \varphi_0^{1/3}(t)\big]^{2-p} \,\big[\int_{\mathbb{R}} \varphi_0^{(p+1)/3}(t)\,\mathrm{d}t\big].\end{aligned}$$

Therefore, using the value ϵ_p defined in Lemma 3,

$$\tfrac{\epsilon^p}{\mu^p} \underset{B,M}{\simeq} \tfrac{2^{-p(B+1)}}{(p+1)\,\mathbb{E}|\mathcal{Z}|^p} \,\|\!|\varphi_0|\!\|_{1/3}^{(p+1)/3} \,\|\!|\varphi_0|\!\|_{(p+1)/3}^{-(p+1)/3}$$

However, for $\alpha > 0$,

$$\|\!|\varphi_0|\!\|_\alpha^\alpha := \int_{\mathbb{R}} \varphi_0^\alpha(t)\,\mathrm{d}t = (2\pi\sigma_0^2)^{-\alpha/2}\,(2\pi\sigma_0^2/\alpha)^{1/2} \int_{\mathbb{R}} \gamma_{0,\sigma_0/\sqrt{\alpha}}(t)\,\mathrm{d}t \;=\; (2\pi\sigma_0^2)^{(1-\alpha)/2}/\sqrt{\alpha}.$$

Consequently, $\|\!|\varphi_0|\!\|_{1/3}^{(p+1)/3} = 3^{(p+1)/2}\,(2\pi\sigma_0^2)^{(p+1)/3}$ and $\|\!|\varphi_0|\!\|_{(p+1)/3}^{(p+1)/3} = (2\pi\sigma_0^2)^{(2-p)/6}/\sqrt{(p+1)/3}$, so that

$$\tfrac{\epsilon^p}{\mu^p} \underset{B,M}{\simeq} \tfrac{2^{-p(B+1)}}{\sqrt{p+1}\,\mathbb{E}|\mathcal{Z}|^p}\,(6\pi\sigma_0^2)^{p/2}$$

Knowing that $(\mathbb{E}|\mathcal{Z}|^p)^{1/p} \geqslant c\sqrt{p+1}$ with $c = 8\sqrt{2}/(9\sqrt{e})$ [8], we get

$$\tfrac{\epsilon}{\mu} \;\lesssim_{B,M}\; c'\,2^{-B}\,\tfrac{(p+1)^{-\frac{1}{2p}}}{\sqrt{p+1}} \leqslant c'\,\frac{2^{-B}}{\sqrt{p+1}}.$$

with $c' = (9/8)(e\pi/3)^{1/2}$. ∎

APPENDIX H
COMPUTATION OF THE $\omega_{k,p}$

This section describes a numerical procedure for efficiently computing the p-optimal levels $\omega_{k,p}$ of a Gaussian source $\mathcal{N}(0,1)$ for integer $p \geqslant 2$, defined by $\omega_{k,p} := \operatorname{argmin}_{\lambda \in \mathcal{R}_k} \mathcal{E}_{k,p}(\lambda)$, where $\mathcal{E}_{k,p}(\lambda) = \int_{t_k}^{t_{k+1}} |t-\lambda|^p \ \gamma_{0,1}(t)\,\mathrm{d}t$. As $\mathcal{E}_{k,p}(\lambda)$ is strictly convex and differentiable, the desired $\omega_{k,p}$ are the unique stationary points satisfying $\mathcal{E}'_{k,p}(\omega_{k,p}) = 0$.

We compute the $\omega_{k,p}$ by Newton method, using standard numerical quadrature for $\mathcal{E}'_{k,p}$ and $\mathcal{E}''_{k,p}$. We handle the semi-infinite bins by replacing $t_1 = -\infty$ and $t_{\mathcal{B}} = \infty$ by -39 and +39, respectively (chosen as the smallest integer x so that $\gamma_{0,1}(x) = 0$ when evaluated in double precision floating point arithmetic). Given quadrature weights c_i, we approximate $\mathcal{E}_{k,p}$ by $\tilde{\mathcal{E}}_{k,p}(\lambda) = \sum_{i=1}^{N} c_i \gamma_{0,1}(x_i)|x_i-\lambda|^p$ with $x_i = t_k + (i-1)\Delta x$, where $\Delta x = (t_{k+1}-t_k)/(N-1)$. We then have $\tilde{\mathcal{E}}'_{k,p}(\lambda) = \sum_{i=1}^{N} c_i \gamma_{0,1}(x_i) p|x_i - \lambda|^{p-1}\mathrm{sign}\,(x_i-\lambda)$ and $\tilde{\mathcal{E}}''_{k,p}(\lambda) = \sum_{i=1}^{N} c_i \gamma_{0,1}(x_i) p(p-1)|x_i-\lambda|^{p-2}$. We initialize with the midpoint for each of the finite bins, *i.e.*, set $\lambda_k^{(0)} = (t_k + t_{k+1})/2$ for $2 \leqslant k \leqslant \mathcal{B}-1$, and $\lambda_1^{(0)} = t_2$, $\lambda_{\mathcal{B}}^{(0)} = t_{\mathcal{B}-1}$ for the semi-infinite bins. For each k we then iterate the Newton step $\lambda_k^{(n)} = \lambda_k^{(n-1)} - \tilde{\mathcal{E}}'_{k,p}(\lambda_k^{(n-1)})/\tilde{\mathcal{E}}''_{k,p}(\lambda_k^{(n-1)})$ until the convergence criterion $|(\lambda_k^n - \lambda_k^{n-1})/\lambda_k^n| < 10^{-15}$ is met. We used c_i given by the fourth-order accurate Simpson's rule, *e.g.*, $\boldsymbol{c} = (1,4,2,4\ldots2,4,1)\Delta x/3$, which yielded empirically observed $O(N^{-4})$ convergence of the calculated $w_{k,p}$. Results in this paper employed $N = 10^4 + 1$ quadrature points, sufficient to yield $w_{k,p}$ accurate to machine precision.

REFERENCES

[1] D. L. Donoho, "Compressed Sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[2] E. J. Candès, "The restricted isometry property and its implications for compressed sensing," *Compte Rendus Acad. Sc., Paris, Serie I*, vol. 346, pp. 589–592, 2008.

[3] W. Dai, H. V. Pham, and O. Milenkovic, "Information theoretical and algorithmic approaches to quantized compressive sensing," *IEEE Trans. Comm.*, vol. 59, no. 7, pp. 1857–1866, 2011.

[4] J. Laska, P. Boufounos, M. Davenport, and R. Baraniuk, "Democracy in action: Quantization, saturation, and compressive sensing," *App. Comp. and Harm. Anal.*, vol. 31, no. 3, pp. 429–443, November 2011.

[5] A. Zymnis, S. Boyd, and E. Candès, "Compressed sensing with quantized measurements," *IEEE Sig. Proc. Letters*, vol. 17, no. 2, pp. 149–152, Feb. 2010.

[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. Inf. Theory*, vol. 59, no. 4, pp. 2082-2102, Apr. 2013.

[7] Y. Plan and R. Vershynin, "One-bit compressed sensing by linear programming," *Comm. Pure App. Math.*, Feb. 2013.

[8] L. Jacques, D. K. Hammond, and M. J. Fadili, "Dequantizing Compressed Sensing: When Oversampling and Non-Gaussian Constraints Combine.," *IEEE Trans. Inf. Theory*, vol. 57, no. 1, pp. 559–571, Jan. 2011.

[9] R. M. Gray and D. L. Neuhoff, "Quantization," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2325–2383, 1998.

[10] N. T. Thao and M. Vetterli, "Reduction of the MSE in R-times oversampled A/D conversion $O(1/R)$ to $O(1/R^2)$". *IEEE Trans. Sig. Proc.*, vol. 42, no. 1, pp. 200-203, 1994.

[11] V. K. Goyal, M. Vetterli, N. T. Thao, "Quantized Overcomplete Expansions in $\mathbb{R}^N$: Analysis, Synthesis, and Algorithms", *IEEE Trans. Inf. Theory*, vol. 44, no. 1, pp. 16–31, 1998.

[12] U. Kamilov, V.K. Goyal, and S. Rangan, "Optimal quantization for compressive sensing under message passing reconstruction," in *IEEE Int. Symp. Inf. Theory Proc. (ISIT)*, 2011, pp. 459–463.

[13] S. Güntürk, A. Powell, R. Saab, and Ö. Yılmaz, "Sobolev duals for random frames and sigma-delta quantization of compressed sensing measurements," *Found. Comp. Math.*, vol. 13, no. 1, pp. 1–36, 2013.

[14] D. E. Knuth, "Big omicron and big omega and big theta," *ACM Sigact News*, vol. 8, no. 2, pp. 18–24, 1976.

[15] J. N. Laska, P. Boufounos, and R. G. Baraniuk, "Finite-range scalar quantization for compressive sensing," in *Conf. Sampling Th. Appl. (SampTA)*, 2009.

[16] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. 28, no. 2, pp. 129–137, Mar. 1982.

[17] J. Max, "Quantizing for minimum distortion," *IEEE Trans. Inf. Theory*, vol. 6, no. 1, pp. 7–12, Mar. 1960.

[18] P. F. Panter and W. Dite, "Quantization distortion in pulse-count modulation with nonuniform spacing of levels," *Proc. IRE*, vol. 39, no. 1, pp. 44–48, 1951.

[19] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic Decomposition by Basis Pursuit," *SIAM J. Sc. Comp.*, vol. 20, no. 1, pp. 33–61, 1998.

[20] J.-J. Fuchs, "Fast implementation of a ℓ_1-ℓ_1 regularized sparse representations algorithm.," in *Proc. IEEE Int. Conf. Acoustics, Sp. Sig. Proc.*, 2009, pp. 3329–3332.

[21] R. Berinde, A. C. Gilbert, P. Indyk, H. Karloff, and M. J. Strauss, "Combining geometry and combinatorics: A unified approach to sparse signal recovery," in *Allerton Conf. Comm., Control & Comp.*. IEEE, 2008, pp. 798–805.

[22] R. Chartrand and V. Staneva, "Restricted isometry properties and nonconvex compressive sensing," *Inv. Prob.*, vol. 24, no. 3, pp. 1–14, 2008.

[23] M. K. Varanasi and B. Aazhang, "Parametric generalized Gaussian density estimation," *J. Acoustical Soc. Am.*, vol. 86, pp. 1404–1415, 1989.

[24] J. J. Moreau, "Fonctions convexes duales et points proximaux dans un espace hilbertien," *CR Acad. Sci. Paris Ser. A Math*, vol. 255, pp. 2897–2899, 1962.

[25] K. J. Arrow, L. Hurwicz, and H. Uzawa, *Studies in linear and non-linear programming*, vol. 2, Stanford University Press Stanford, 1958,

[26] G. Chen and M. Teboulle, "A proximal-based decomposition method for convex minimization problems," *Math. Prog.*, vol. 64, no. 1, pp. 81–101, 1994.

[27] A. Chambolle and T. Pock, "A First-Order Primal-Dual Algorithm for Convex Problems with Applications to Imaging," *J. Math. Im. Vis.*, vol. 40, no. 1, pp. 120–145, Dec. 2010.

[28] L. M. Briceño-Arias and P. L. Combettes, "A monotone+skew splitting model for composite monotone inclusions in duality," *SIAM J. Optim.*, vol. 21, no. 4, pp. 1230–1250, Oct. 2011.

[29] R. L. Winkler, G. M. Roodman, and R. R. Britney, "The determination of partial moments," *Management Science*, pp. 290–296, 1972.