

Dequantizing Compressed Sensing:

When Oversampling and Non-Gaussian Constraints Combine

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

Abstract

In this paper we study the problem of recovering sparse or compressible signals from uniformly quantized measurements. We present a new class of convex optimization programs, or decoders, coined Basis Pursuit DeQuantizer of moment p (BPDQ_p), that model the quantization distortion more faithfully than the commonly used Basis Pursuit DeNoise (BPDN) program. Our decoders proceed by minimizing the sparsity of the signal to be reconstructed subject to a data-fidelity constraint expressed in the ℓ_p -norm of the residual error for $2 \leq p \leq \infty$.

We show theoretically that, (i) the reconstruction error of these new decoders is bounded if the sensing matrix satisfies an extended Restricted Isometry Property involving the ℓ_p norm, and (ii), for Gaussian random matrices and uniformly quantized measurements, BPDQ_p performance exceeds that of BPDN by dividing the reconstruction error due to quantization by $\sqrt{p+1}$. This last effect happens with high probability when the number of measurements exceeds a value growing with p , *i.e.*, in an oversampled situation compared to what is commonly required by $\text{BPDN} = \text{BPDQ}_2$. To demonstrate the theoretical power of BPDQ_p , we report numerical simulations on signal and image reconstruction problems.

Index Terms

Compressed Sensing, Convex Optimization, Denoising, Optimality, Oversampling, Quantization, Sparsity.

LJ is with the Information and Communication Technologies, Electronics and Applied Mathematics (ICTEAM) Sector, 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.

A part of this work was presented at the IEEE Intl. Conf. Image Proc. (ICIP), Cairo, Egypt, 2009 [1].

I. INTRODUCTION

The theory of Compressed Sensing (CS) [2], [3] aims at reconstructing sparse or compressible signals from a small number of linear measurements compared to the dimensionality of the signal space. In short, the signal reconstruction is possible if the underlying sensing matrix is well behaved, *i.e.*, if it respects a Restricted Isometry Property (RIP) saying roughly that any small subset of its columns is “close” to an orthogonal basis. The signal recovery is then obtained using non-linear techniques based on convex optimization promoting signal sparsity, such as the Basis Pursuit program [3]. What makes CS more than merely an interesting theoretical concept is that some classes of randomly generated matrices (*e.g.*, Gaussian, Bernoulli, partial Fourier ensemble, etc) satisfy the RIP with overwhelming probability. This happens as soon as their number of rows, *i.e.*, the number of CS measurements, is higher than a few multiples of the assumed signal sparsity.

In a realistic acquisition system, quantization of these measurements is a natural process that Compressed Sensing theory has to handle conveniently. One commonly used technique is to simply treat the quantization distortion as Gaussian noise, which leads to reconstruction based on solving the Basis Pursuit DeNoising (BPDN) program (either in its constrained or augmented Lagrangian forms) [4]. While this approach can give acceptable results, it is theoretically unsatisfactory as the measurement error created by quantization is highly non-Gaussian, being essentially uniform and bounded by the quantization bin width.

An appealing requirement for the design of better reconstruction methods is the Quantization Consistency (QC) constraint, *i.e.*, that the requantized measurements of the reconstructed signal equal the original quantized measurements. This idea, in some form, has appeared previously in the literature. Near the beginning of the development of CS theory, Candès et al. mentioned that the ℓ_2 -norm of BPDN should be replaced by the ℓ_∞ -norm to handle more naturally the quantization distortion of the measurements [4]. More recently, in [5], the extreme case of 1-bit CS is studied, *i.e.*, when only the signs of the measurements are sent to the decoder. Authors tackle the reconstruction problem by adding a sign consistency constraint in a modified BPDN program working on the sphere of unit-norm signals. In [6], an adaptation of both BPDN and the Subspace Pursuit integrates an explicit QC constraint. In [7], a model integrating additional Gaussian noise on the measurements before their quantization is analyzed and solved with a ℓ_1 -regularized maximum likelihood program. However, in spite of interesting experimental results, no theoretical guarantees are given about the approximation error reached by these solutions.

The QC constraint has also been used previously for image and signal processing outside of the CS field. Examples include oversampled Analog to Digital Converters (ADC) [8], and in image restoration problems [9], [10].

In this paper, we propose a new class of convex optimization programs, or decoders, coined the Basis Pursuit DeQuantizer of moment p ($BPDQ_p$) that model the quantization distortion more faithfully. These proceed by minimizing the sparsity of the reconstructed signal (expressed in the ℓ_1 -norm) subject to a particular data-fidelity constraint. This constraint imposes that the difference between the original and the reproduced measurements have bounded ℓ_p -norm, for $2 \leq p \leq \infty$. As p approaches infinity, this fidelity term reproduces the QC constraint as promoted initially in [4]. However, our idea is to study, given a certain sparsity level and in function of the number of measurements available, which moment $2 \leq p \leq \infty$ provides the best reconstruction result.

Our overall result, which surprisingly does not favor $p = \infty$, may be expressed by the principle: *Given a certain sparsity level, if the number of measurements is higher than a minimal value growing with p , i.e., in oversampled situations, by using $BPDQ_p$ instead of $BPDN = BPDQ_2$ the reconstruction error due to quantization can be reduced by a factor of $\sqrt{p+1}$.*

At first glance, it could seem counterintuitive to oversample the “compressive sensing” of a signal. After all, many results in Compressed Sensing seek to limit the number of measurements required to encode a signal, while guaranteeing exact reconstruction with high probability. However, as analyzed for instance in [11], this way of thinking avoids to considering the actual amount of information needed to describe the measurement vector. In the case of noiseless observations of a sparse signal, Compressed Sensing guarantees perfect reconstruction only for real-valued measurements, i.e., for an infinite number of bits per measurements.

From a rate-distortion perspective, the analysis shown in [12], [13] demonstrates also that CS is suboptimal compared to transform coding. Under that point of view, the best CS encoding strategy is to use all the available bit-rate to obtain as few CS measurements as possible and quantize them as finely as possible.

However, in many practical situations the quantization bit-depth per measurement is pre-determined by the hardware, e.g., for real sensors embedding CS and a fixed A/D conversion of the measurements. In that case, the only way to improve the reconstruction quality is to gather more measurements,

i.e., to oversample the signal¹. This does not degrade one of the main interests of Compressed Sensing, *i.e.*, providing highly informative linear signal measurements at a very low computation cost.

The paper is structured as follows. In Section II, we review the principles of Compressed Sensing and previous approaches for accommodating the problem of measurement quantization. Section III introduces the BPDQ_p decoders. Their stability, *i.e.*, the $\ell_2 - \ell_1$ instance optimality, is deduced using an extended version of the Restricted Isometry Property involving the ℓ_p -norm. In Section IV, Standard Gaussian Random matrices, *i.e.*, whose entries are independent and identically distributed (iid) standard Gaussian, are shown to satisfy this property with high probability for a sufficiently large number of measurements. Section V explains the key result of this paper; that the approximation error of BPDQ_p scales inversely with $\sqrt{p+1}$. Section VI describes the convex optimization framework adopted to solve the BPDQ_p programs. Finally, Section VII provides experimental validation of the theoretical power of BPDQ_p on 1-D signals and on an image reconstruction example.

II. COMPRESSED SENSING AND QUANTIZATION OF MEASUREMENTS

In Compressed Sensing (CS) theory [2], [3], the signal $x \in \mathbb{R}^N$ to be acquired and subsequently reconstructed is typically assumed to be sparse or *compressible* in an orthogonal² basis $\Psi \in \mathbb{R}^{N \times N}$ (*e.g.*, wavelet basis, Fourier, etc.). In other words, the best K -term approximation x_K of x in Ψ gives an exact (for the sparse case) or accurate (for the compressible case) representation of x even for small $K < N$. For simplicity, only the canonical basis $\Psi = \text{Id}$ will be considered here.

At the acquisition stage, x is encoded by m linear measurements (with $K \leq m \leq N$) provided by a sensing matrix $\Phi \in \mathbb{R}^{m \times N}$, *i.e.*, all known information about x is contained in the m measurements $\langle \varphi_i, x \rangle = \sum_k \varphi_{ik}^* x_k$, where $\{\varphi_i\}_{i=0}^{m-1}$ are the rows of Φ .

In this paper, we are interested in a particular non-ideal sensing model. Indeed, as measurement of continuous signals by digital devices always involves some form of quantization, in practice devices based on CS encoding must be able to accommodate the distortions in the linear measurements created by quantization. Therefore, we adopt the noiseless and uniformly quantized sensing (or coding) model:

$$y_q = Q_\alpha[\Phi x] = \Phi x + n, \quad (1)$$

¹Generally, it is also less expensive in hardware to oversample a signal than to quantize measurements more finely.

²A generalization for redundant basis, or dictionary, exists [14], [15].

where $y_q \in (\alpha\mathbb{Z} + \frac{\alpha}{2})^m$ is the quantized measurement vector, $(Q_\alpha[\cdot])_i = \alpha\lfloor(\cdot)_i/\alpha\rfloor + \frac{\alpha}{2}$ is the uniform quantization operator in \mathbb{R}^m of bin width α , and $n \triangleq Q_\alpha[\Phi x] - \Phi x$ is the *quantization distortion*.

The model (1) is a realistic description of systems where the quantization distortion dominates other secondary noise sources (*e.g.*, thermal noise), an assumption valid for many electronic measurement devices including ADC. In this paper we restrict our study to using this extremely simple uniform quantization model, in order to concentrate on the interaction with the CS theory. For instance, this quantization scenario does not take into account the possible *saturation* of the quantizer happening when the value to be digitized is outside the operating range of the quantizer, this range being determined by the number of bits available. For Compressed Sensing, this effect has been studied recently in [16]. Authors obtained better reconstruction methods by either imposing to reproduce saturated measurements (Saturation Consistency) or by discarding these thanks to the “democratic” property of most of the random sensing matrices. Their work however does not integrate the Quantization Consistency for all the unsaturated measurements. The study of more realistic non-uniform quantization is also deferred as a question for future research.

In much previous work in CS, the reconstruction of x from y_q is obtained by treating the quantization distortion n as a noise of bounded power (*i.e.*, ℓ_2 -norm) $\|n\|_2^2 = \sum_k |n_k|^2$. In this case, a robust reconstruction of the signal x from corrupted measurements $y = \Phi x + n$ is provided by the Basis Pursuit DeNoise (BPDN) program (or decoder) [17]:

$$\Delta(y, \epsilon) = \underset{u \in \mathbb{R}^N}{\operatorname{argmin}} \|u\|_1 \text{ s.t. } \|y - \Phi u\|_2 \leq \epsilon. \quad (\text{BPDN})$$

This convex optimization program can be solved numerically by methods like Second Order Cone Programming or by monotone operator splitting methods [18], [19] described in Section VI. Notice that the noiseless situation $\epsilon = 0$ leads to the Basis Pursuit (BP) program, which may also be solved by Linear Programming [20].

An important condition for BPDN to provide a good reconstruction is the *feasibility* of the initial signal x , *i.e.*, we must chose ϵ in the (*fidelity*) constraint of BPDN such that $\|n\|_2 = \|y - \Phi x\|_2 \leq \epsilon$. In [17], an estimator of ϵ for $y = y_q$ is obtained by considering n as a random vector $\xi \in \mathbb{R}^m$ distributed uniformly over the quantization bins, *i.e.*, $\xi_i \sim_{\text{iid}} U([- \frac{\alpha}{2}, \frac{\alpha}{2}])$.

An easy computation shows then that $\|\xi\|_2^2 \leq \epsilon_2^2(\alpha)$ with probability higher than $1 - e^{-c_0 \kappa^2}$ for a

certain constant $c_0 > 0$ (by the Chernoff-Hoeffding bound [21]), where

$$\epsilon_2^2(\alpha) \triangleq \mathbb{E}\|\xi\|_2^2 + \kappa \sqrt{\text{Var}\|\xi\|_2^2} = \frac{\alpha^2}{12}m + \kappa \frac{\alpha^2}{6\sqrt{5}}m^{\frac{1}{2}}.$$

Therefore, CS usually handles quantization distortion by setting $\epsilon = \epsilon_2(\alpha)$, typically for $\kappa = 2$.

When the feasibility is satisfied, the stability of BPDN is guaranteed if the sensing matrix $\Phi \in \mathbb{R}^{m \times N}$ satisfies one instance of the following property:

Definition 1. A matrix $\Phi \in \mathbb{R}^{m \times N}$ satisfies the (extended) Restricted Isometry Property ($\text{RIP}_{p,q}$) (with $p, q > 0$) of order K and radius $\delta_K \in (0, 1)$, if there exists a constant $\mu_{p,q} > 0$ such that

$$\mu_{p,q} (1 - \delta_K)^{1/q} \|u\|_q \leq \|\Phi u\|_p \leq \mu_{p,q} (1 + \delta_K)^{1/q} \|u\|_q, \quad (2)$$

for all K -sparse signals $u \in \mathbb{R}^N$.

In other words, Φ , as a mapping from $\ell_p^m = (\mathbb{R}^m, \|\cdot\|_p)$ to $\ell_q^N = (\mathbb{R}^N, \|\cdot\|_q)$, acts as a (scaled) isometry on K -sparse signals of \mathbb{R}^N . This definition is more general than the common RIP [22]. This latter, which ensures the stability of BPDN (see Theorem 1 below), corresponds to $p = q = 2$ in (2). The original definition considers also normalized matrices $\bar{\Phi} = \Phi/\mu_{2,2}$ having unit-norm columns (in expectation) so that $\mu_{2,2}$ is absorbed in the normalizing constant.

We prefer to use this extended $\text{RIP}_{p,q}$ since, as it will become clear in Section V, the case $p \geq 2$ and $q = 2$ provides us the interesting embedding (2) for measurement vectors corrupted by generalized Gaussian and uniform noises. As explained below, this definition includes also other RIP generalizations [26], [28].

We note that there are several examples already described in the literature of classes of matrices which satisfy the $\text{RIP}_{p,q}$ for specific values of p and q . For instance, for $p = q = 2$, a matrix $\Phi \in \mathbb{R}^{m \times N}$ with each of its entries drawn independently from a (sub) Gaussian random variable satisfies this property with an overwhelming probability if $m \geq cK \log N/K$ for some value $c > 0$ independent of the involved dimensions [23], [24], [25]. This is the case of Standard Gaussian Random (SGR) matrices whose entries are iid $\Phi_{ij} \sim \mathcal{N}(0, 1)$, and of the Bernoulli matrices with $\Phi_{ij} = \pm 1$ with equal probability, both cases having $\mu_{2,2} = \sqrt{m}$ [23]. Other random constructions satisfying the $\text{RIP}_{2,2}$ are known (e.g., partial Fourier ensemble) [2], [17]. For the case $p = q = 1 + O(1)/\log N$, it is proved in [26], [27] that sparse matrices obtained from an adjacency matrix of a high-quality unbalanced expander graph are $\text{RIP}_{p,p}$ (with $\mu_{p,p}^2 = 1/(1 - \delta_K)$). In the context of non-convex signal reconstruction, the authors in [28] show also that

Gaussian random matrices satisfy the Restricted p -Isometry, *i.e.*, $\text{RIP}_{p,q}$ for $q = 2$, $0 < p < 1$, $\mu_{p,2} = 1$ and appropriate redefinition of δ_K .

The following theorem expresses the announced stability result, *i.e.*, the $\ell_2 - \ell_1$ instance optimality³ of BPDN, as a consequence of the $\text{RIP}_{2,2}$.

Theorem 1 ([22]). *Let $x \in \mathbb{R}^N$ be a signal whose compressibility is measured by the decreasing of the K -term ℓ_1 -approximation error $e_0(K) = K^{-\frac{1}{2}} \|x - x_K\|_1$, for $0 \leq K \leq N$, and x_K the best K -term ℓ_2 -approximation of x . Let Φ be a $\text{RIP}_{2,2}$ matrix of order $2K$ and radius $0 < \delta_{2K} < \sqrt{2} - 1$. Given a measurement vector $y = \Phi x + n$ corrupted by a noise n with power $\|n\|_2 \leq \epsilon$, the solution $x^* = \Delta(y, \epsilon)$ obeys*

$$\|x^* - x\|_2 \leq A e_0(K) + B \frac{\epsilon}{\mu_{2,2}}, \quad (3)$$

for $A(\Phi, K) = 2 \frac{1+(\sqrt{2}-1)\delta_{2K}}{1-(\sqrt{2}+1)\delta_{2K}}$ and $B(\Phi, K) = \frac{4\sqrt{1+\delta_{2K}}}{1-(\sqrt{2}+1)\delta_{2K}}$. For instance, for $\delta_{2K} = 0.2$, $A < 4.2$ and $B < 8.5$.

Let us precise that the theorem condition $\delta_{2K} < \sqrt{2} - 1$ on the RIP radius can be refined (like in [31]). We know nevertheless from Davies and Gribonval [32] that ℓ_1 -minimization will fail for at least one vector for $\delta_{2K} > 1/\sqrt{2}$. The room for improvement is then very small.

Using the BPDN decoder to account for quantization distortion is theoretically unsatisfying for several reasons. First, there is no guarantee that the BPDN solution x^* respects the Quantization Consistency, *i.e.*,

$$Q_\alpha[\Phi x^*] = y_q \Leftrightarrow \|y_q - \Phi x^*\|_\infty \leq \frac{\alpha}{2}, \quad (\text{QC})$$

which is not necessarily implied by the BPDN ℓ_2 fidelity constraint. The failure of BPDN to respect QC suggests that it may not be taking advantage of all of the available information about the noise structure in the measurements.

Second, from a Bayesian Maximum a Posteriori (MAP) standpoint, BPDN can be viewed as solving an ill-posed inverse problem where the ℓ_2 -norm used in the fidelity term corresponds to the conditional log-likelihood associated to an additive white Gaussian noise. However, the quantization distortion is not Gaussian, but rather uniformly distributed. This motivates the need for a new kind of CS decoder that more faithfully models the quantization distortion.

³Adopting the definition of mixed-norm instance optimality [29].

III. BASIS PURSUIT DEQUANTIZER (BPDQ_p)

The considerations of the previous section encourage the definition of a new class of optimization programs (or decoders) generalizing the fidelity term of the BPDN program.

Our approach is based on reconstructing a sparse approximation of x from its measurements $y = \Phi x + n$ under the assumption that ℓ_p -norm ($p \geq 1$) of the noise n is bounded, i.e., $\|n\|_p^p = \sum_k |n_k|^p \leq \epsilon^p$ for some $\epsilon > 0$. We introduce the novel programs

$$\Delta_p(y, \epsilon) = \underset{u \in \mathbb{R}^N}{\operatorname{argmin}} \|u\|_1 \text{ s.t. } \|y - \Phi u\|_p \leq \epsilon. \quad (\text{BPDQ}_p)$$

The fidelity constraint expressed in the ℓ_p -norm is now tuned to noises that follow a zero-mean Generalized Gaussian Distribution⁴ (GGD) of *shape parameter* p [30], with the uniform noise case corresponding to $p \rightarrow \infty$.

We dub this class of decoders *Basis Pursuit DeQuantizer of moment p* (or BPDQ_p) since, for reasons that will become clear in Section V, their approximation error when Φx is uniformly quantized has an interesting decreasing behavior when both the moment p and the oversampling factor m/K increase. Notice that the decoder corresponding to $p = 1$ has been previously analyzed in [33] for Laplacian noise.

One of the main results of this paper concerns the $\ell_2 - \ell_1$ instance optimality of the BPDQ_p decoders, i.e., their stability when the signal to be recovered is compressible, and when the measurements are contaminated by noise of bounded ℓ_p -norm. In the following theorem, we show that such an optimality happens when the sensing matrix respects the (extended) Restricted Isometry Property RIP_{p,2} for $2 \leq p < \infty$.

Theorem 2. *Let $x \in \mathbb{R}^N$ be a signal with a K -term ℓ_1 -approximation error $e_0(K) = K^{-\frac{1}{2}} \|x - x_K\|_1$, for $0 \leq K \leq N$ and x_K the best K -term ℓ_2 -approximation of x . Let Φ be a RIP_{p,2} matrix on s sparse signals with constants δ_s , for $s \in \{K, 2K, 3K\}$ and $2 \leq p < \infty$. Given a measurement vector $y = \Phi x + n$ corrupted by a noise n with bounded ℓ_p -norm, i.e., $\|n\|_p \leq \epsilon$, the solution $x_p^* = \Delta_p(y, \epsilon)$ of BPDQ_p obeys*

$$\|x_p^* - x\|_2 \leq A_p e_0(K) + B_p \epsilon / \mu_{p,2},$$

for values $A_p(\Phi, K) = \frac{2(1+C_p-\delta_{2K})}{1-\delta_{2K}-C_p}$, $B_p(\Phi, K) = \frac{4\sqrt{1+\delta_{2K}}}{1-\delta_{2K}-C_p}$, and $C_p = C_p(\Phi, 2K, K)$ given in the proof of Lemma 2 (Appendix D).

⁴The probability density function f of such a distribution is $f(x) \propto \exp(-|x/b|^p)$ for a standard deviation $\sigma \propto b$.

As shown in Appendix E, this theorem follows from a generalization of the fundamental result proved by Candès [22] to the particular geometry of Banach spaces ℓ_p .

IV. EXAMPLE OF $\text{RIP}_{p,2}$ MATRICES

Interestingly, it turns out that SGR matrices $\Phi \in \mathbb{R}^{m \times N}$ also satisfy the $\text{RIP}_{p,2}$ with high probability provided that m is sufficiently large compared to the sparsity K of the signals to measure. This is made formal in the following Proposition, for which the proof⁵ is given in Appendix A.

Proposition 1. *Let $\Phi \in \mathbb{R}^{m \times N}$ be a Standard Gaussian Random (SGR) matrix, i.e., its entries are iid $\mathcal{N}(0, 1)$. Then, if $m \geq (p-1)2^{p+1}$ for $2 \leq p < \infty$ and $m \geq 0$ for $p = \infty$, there exists a constant $c > 0$ such that, for*

$$\Theta_p(m) \geq c \delta^{-2} \left(K \log \left[e \frac{N}{K} (1 + 12\delta^{-1}) \right] + \log \frac{2}{\eta} \right), \quad (4)$$

with $\Theta_p(m) = m^{2/p}$ for $1 \leq p < \infty$ and $\Theta_p(m) = \log m$ for $p = \infty$, Φ is $\text{RIP}_{p,2}$ of order K and radius δ with probability higher than $1 - \eta$. Moreover, the value $\mu_{p,2} = \mathbb{E} \|\xi\|_p$ is the expectation value of the ℓ_p -norm of a SGR vector $\xi \in \mathbb{R}^m$.

Roughly speaking, this proposition tells us that to generate a matrix that is $\text{RIP}_{p,2}$ with high probability, we need a number of measurements m that grows polynomially in $K \log N/K$ with an “order” $p/2$ for $2 \leq p < \infty$, while the limit case $p = \infty$ grows exponentially in $K \log N/K$.

Notice that an asymptotic estimation of $\mu_{p,2}$, i.e., for $m \rightarrow \infty$, can be found in [34] for $1 \leq p < \infty$. However, as presented in the following Lemma (proved in Appendix C), non-asymptotic bounds for $\mu_{p,2} = \mathbb{E} \|\xi\|_p$ can be expressed in terms of

$$(\mathbb{E} \|\xi\|_p^p)^{1/p} = (m \mathbb{E} |g|^p)^{1/p} = \nu_p m^{1/p},$$

with $g \sim \mathcal{N}(0, 1)$ and $\nu_p^p = \mathbb{E} |g|^p = 2^{\frac{p}{2}} \pi^{-\frac{1}{2}} \Gamma(\frac{p+1}{2})$.

Lemma 1. *If $\xi \in \mathbb{R}^m$ is a SGR vector, then, for $1 \leq p < \infty$,*

$$\left(1 + \frac{2^{p+1}}{m} \right)^{\frac{1}{p}-1} (\mathbb{E} \|\xi\|_p^p)^{\frac{1}{p}} \leq \mathbb{E} \|\xi\|_p \leq (\mathbb{E} \|\xi\|_p^p)^{\frac{1}{p}}.$$

⁵Interestingly, this proof shows also that SGR matrices are $\text{RIP}_{p,2}$ with high probability for $1 < p < 2$ when m exceeds a similar bound to (4).

In particular, as soon as $m \geq \beta^{-1} 2^{p+1}$ for $\beta \geq 0$, $\mathbb{E}\|\xi\|_p \geq (\mathbb{E}\|\xi\|_p^p)^{\frac{1}{p}} (1+\beta)^{\frac{1}{p}-1} \geq (\mathbb{E}\|\xi\|_p^p)^{\frac{1}{p}} (1-\frac{p-1}{p}\beta)$.
For $p = \infty$, there exists a $\rho > 0$ such that $\rho^{-1} \sqrt{\log m} \leq \mathbb{E}\|\xi\|_\infty \leq \rho \sqrt{\log m}$.

An interesting aspect of matrices respecting the $\text{RIP}_{p,2}$ is that they approximately preserve the decorrelation of sparse vectors of disjoint supports.

Lemma 2. *Let $u, v \in \mathbb{R}^N$ with $\|u\|_0 = s$ and $\|v\|_0 = s'$ and $\text{supp}(u) \cap \text{supp}(v) = \emptyset$, and $2 \leq p < \infty$. If Φ is $\text{RIP}_{p,2}$ of order $s + s'$ with constant $\delta_{s+s'}$, and of orders s and s' with constants δ_s and $\delta_{s'}$, then*

$$|\langle J(\Phi u), \Phi v \rangle| \leq \mu_{p,2}^2 C_p \|u\|_2 \|v\|_2, \quad (5)$$

with $(J(u))_i = \|u\|_p^{2-p} |u_i|^{p-1} \text{sign } u_i$ and $C_p = C_p(\Phi, s, s')$ is given explicitly in Appendix D.

It is worth mentioning that the value C_p behaves as $\sqrt{(\delta_s + \delta_{s+s'})(1 + \delta_{s'})(p-2)}$ for large p , and as $\delta_{s+s'} + \frac{3}{4}(1 + \delta_{s+s'})(p-2)$ for $p \simeq 2$. Therefore, this result may be seen as a generalization of the one proved in [22] (see Lemma 2.1) for $p = 2$ with $C_2 = \delta_{s+s'}$. As shown in Appendix D, this Lemma uses explicitly the 2-smoothness of the Banach spaces ℓ_p when $p \geq 2$ [35], [36], in connection with the *normalized duality mapping* J that plays a central role in the geometrical description of ℓ_p .

Lemma 2 is at the heart of the proof of Theorem 2, which prevents the later from being valid for $p = \infty$. This is related to the fact that the ℓ_∞ Banach space is not 2-smooth and no duality mapping exists. Therefore, any result for $p = \infty$ would require different tools than those developed here.

V. BPDQ_p AND QUANTIZATION ERROR REDUCTION

Let us now observe the particular behavior of the BPDQ_p decoders on quantized measurements of a sparse or compressible signal assuming that α is known at the decoding step. In this Section, we consider that $p \geq 2$ everywhere.

First, if we assume in the model (1) that the quantization distortion $n = Q_\alpha[\Phi x] - \Phi x$ is uniformly distributed in each quantization bin, the simple Lemma below provides precise estimator ϵ for any ℓ_p -norm of n .

Lemma 3. *If $\xi \in \mathbb{R}^m$ is a uniform random vector with $\xi_i \sim_{\text{iid}} U([- \frac{\alpha}{2}, \frac{\alpha}{2}])$, then, for $1 \leq p < \infty$,*

$$\zeta_p = \mathbb{E}\|\xi\|_p^p = \frac{\alpha^p}{2^{p(p+1)}} m. \quad (6)$$

In addition, for any $\kappa > 0$, $\mathbb{P}[\|\xi\|_p^p \geq \zeta_p + \kappa \frac{\alpha^p}{2^p} \sqrt{m}] \leq e^{-2\kappa^2}$, while, $\lim_{p \rightarrow \infty} (\zeta_p + \kappa \frac{\alpha^p}{2^p} \sqrt{m})^{\frac{1}{p}} = \frac{\alpha}{2}$.

The proof is given in Appendix F.

According to this result, we may set the ℓ_p -norm bound ϵ of the program BPDQ_p to

$$\epsilon = \epsilon_p(\alpha) \triangleq \frac{\alpha}{2(p+1)^{1/p}} (m + \kappa(p+1) \sqrt{m})^{\frac{1}{p}}, \quad (7)$$

so that, for $\kappa = 2$, we know that x is a feasible solution of the BPDQ_p fidelity constraint with a probability exceeding $1 - e^{-8} > 1 - 3.4 \times 10^{-4}$.

Second, Theorem 2 points out that, when Φ is RIP_{p,2} with $2 \leq p < \infty$, the approximation error of the BPDQ_p decoders is the sum of two terms: one that expresses the *compressibility error* as measured by $e_0(K)$, and one, the *noise error*, proportional to the ratio $\epsilon/\mu_{p,2}$. In particular, by Lemma 1, for m respecting (4), a SGR sensing matrix of m rows induces with a controlled probability

$$\|x - x_p^*\|_2 \leq A_p e_0(K) + B_p \frac{\epsilon_p(\alpha)}{\mu_{p,2}}. \quad (8)$$

Combining (7) and the result of Lemma 1, we may bound the noise error for uniform quantization more precisely. Indeed, for $2 \leq p < \infty$, if $m \geq (p-1)2^{p+1}$, $\mu_{p,2} \geq \frac{p-1}{p} \nu_p m^{\frac{1}{p}}$ with $\nu_p = \sqrt{2} \pi^{-\frac{1}{2p}} \Gamma(\frac{p+1}{2})^{\frac{1}{p}}$.

In addition, using a variant of the Stirling formula found in [37], we know that $|\Gamma(x) - (\frac{2\pi}{x})^{\frac{1}{2}} (\frac{x}{e})^x| \leq \frac{1}{9x} (\frac{2\pi}{x})^{\frac{1}{2}} (\frac{x}{e})^x$ for $x \geq 1$. Therefore, we compute easily that, for $x = (p+1)/2 > 1$, $\nu_p \geq c^{1/p} (\frac{p+1}{e})^{1/2} \geq c (\frac{p+1}{e})^{1/2}$ with $c = \frac{8\sqrt{2}}{9\sqrt{e}} < 1$. Finally, by (7), we see that,

$$\begin{aligned} \frac{\epsilon_p(\alpha)}{\mu_{p,2}} &\leq \frac{p}{p-1} \frac{9e}{16\sqrt{2}} \left(\frac{1}{p+1} + \kappa \frac{1}{\sqrt{m}} \right)^{1/p} \frac{\alpha}{\sqrt{p+1}} \\ &< C \frac{\alpha}{\sqrt{p+1}}, \end{aligned} \quad (9)$$

with $C = 9e/(8\sqrt{2}) < 2.17$, where we used the bound $\frac{p}{p-1} \leq 2$ and the fact that $(\frac{1}{p+1} + \kappa \frac{1}{\sqrt{m}})^{1/p} < 1$ if $m > (\frac{p+1}{p} \kappa)^2 = O(1)$.

In summary, we can formulate the following principle.

Oversampling Principle. *The noise error term in the $\ell_2 - \ell_1$ instance optimality relation (8) in the case of uniform quantization of the measurements of a sparse or compressible signal is divided by $\sqrt{p+1}$ in oversampled SGR sensing, i.e., when the oversampling factor m/K is higher than a minimal value increasing with p .*

Interestingly, this follows the improvement achieved by adding a QC constraint in the decoding of oversampled ADC signal conversion [8].

The oversampling principle requires some additional explanations. Taking a SGR matrix, by Proposition 1, if m_p is the smallest number of measurements for which such a randomly generated matrix Φ is $\text{RIP}_{p,2}$ of radius $\delta_p < 1$ with a certain nonzero probability, taking $m > m_p$ allows one to generate a new random matrix with a smaller radius $\delta < \delta_p$ with the same probability of success.

Therefore, increasing the *oversampling factor* m/K provides two effects. First, it enables one to hope for a matrix Φ that is $\text{RIP}_{p,2}$ for high p , providing the desired error division by $\sqrt{p+1}$. Second, as shown in Appendix B, since $\delta = O(m^{-1/p} \sqrt{\log m})$, oversampling gives a smaller δ hence counteracting the increase of p in the factor C_p of the values $A_p \geq 2$ and $B_p \geq 4$. This decrease of δ also favors BPDN, but since the values $A = A_2 \geq 2$ and $B = B_2 \geq 4$ in (3) are also bounded from below this effect is limited. Consequently, as the number of measurements increases the improvement in reconstruction error for BPDN will saturate, while for BPDQ_p the error will be divided by $\sqrt{p+1}$.

From this result, it is very tempting to choose an extremely large value for p in order to decrease the noise error term (8). There are however two obstacles with this. First, the instance optimality result of Theorem 2 is not directly valid for $p = \infty$. Second, and more significantly, the necessity of satisfying $\text{RIP}_{p,2}$ implies that we cannot take p arbitrarily large in Proposition 1. Indeed, for a given oversampling factor m/K , a SGR matrix Φ can be $\text{RIP}_{p,2}$ only over a finite interval $p \in [2, p_{\max}]$. This implies that for each particular reconstruction problem, there should be an optimal maximum value for p . We will demonstrate this effect experimentally in Section VII.

We remark that the compressibility error is not significantly reduced by increasing p when the number of measurements is large. This makes sense as the ℓ_p -norm appears only in the fidelity term of the decoders, and we know that in the case where $\epsilon = 0$ the compressibility error remains in the BP decoder [22]. Finally, note that due to the embedding of the ℓ_p -norms, *i.e.*, $\|\cdot\|_p \leq \|\cdot\|_{p'}$ if $p \geq p' \geq 1$, increasing p until p_{\max} makes the fidelity term closer to the QC.

VI. NUMERICAL IMPLEMENTATION

This section is devoted to the description of the convex optimization tools needed to numerically solve the Basis Pursuit DeQuantizer program. While we generally utilize $p \geq 2$, the BPDQ_p program is convex for $p \geq 1$. In fact, the efficient iterative procedure we describe will converge to the global minimum of the BPDQ_p program for all $p \geq 1$.

A. Proximal Optimization

The BPDQ_p (and BPDN) decoders are special case of a general class of convex problems [18], [38]

$$\arg \min_{x \in \mathcal{H}} f_1(x) + f_2(x), \quad (\mathbf{P})$$

where $\mathcal{H} = \mathbb{R}^N$ is seen as an Hilbert space equipped with the inner product $\langle x, z \rangle = \sum_i x_i z_i$. We denote by $\text{dom } f = \{x \in \mathcal{H} : f(x) < +\infty\}$ the domain of any $f : \mathcal{H} \rightarrow \mathbb{R} \cup \{+\infty\}$. In (\mathbf{P}) , the functions $f_1, f_2 : \mathcal{H} \rightarrow \mathbb{R} \cup \{+\infty\}$ are assumed (i) convex functions which are not infinite everywhere, i.e., $\text{dom } f_1, \text{dom } f_2 \neq \emptyset$, (ii) $\text{dom } f_1 \cap \text{dom } f_2 \neq \emptyset$, and (iii) these functions are lower semi-continuous (lsc) meaning that $\liminf_{x \rightarrow x_0} f(x) = f(x_0)$ for all $x_0 \in \text{dom } f$. The class of functions satisfying these three properties is denoted $\Gamma_0(\mathbb{R}^N)$. For BPDQ_p, these two non-differentiable functions are $f_1(x) = \|x\|_1$ and $f_2(x) = \iota_{T^p(\epsilon)}(x) = 0$ if $x \in T^p(\epsilon)$ and ∞ otherwise, i.e., the indicator function of the closed convex set $T^p(\epsilon) = \{x \in \mathbb{R}^N : \|y_q - \Phi x\|_p \leq \epsilon\}$.

It can be shown that the solutions of problem (\mathbf{P}) are characterized by the following fixed point equation: x solves (\mathbf{P}) if and only if

$$x = (\mathbb{1} + \beta \partial(f_1 + f_2))^{-1}(x), \quad \text{for } \beta > 0. \quad (10)$$

The operator $\mathcal{J}_{\beta \partial f} = (\mathbb{1} + \beta \partial f)^{-1}$ is called the *resolvent operator* associated to the *subdifferential operator* ∂f , β is a positive scalar known as the proximal step size, and $\mathbb{1}$ is the identity map on \mathcal{H} . We recall that the subdifferential of a function $f \in \Gamma_0(\mathcal{H})$ at $x \in \mathcal{H}$ is the set-valued map $\partial f(x) = \{u \in \mathcal{H} : \forall z \in \mathcal{H}, f(z) \geq f(x) + \langle u, z - x \rangle\}$, where each element u of ∂f is called a subgradient.

The resolvent operator is actually identified with the *proximity operator* of βf , i.e., $\mathcal{J}_{\beta \partial f} = \text{prox}_{\beta f}$, introduced in [39] as a generalization of convex projection operator. It is defined as the unique solution $\text{prox}_f(x) = \arg \min_{z \in \mathcal{H}} \frac{1}{2} \|z - x\|_2^2 + f(z)$ for $f \in \Gamma_0(\mathcal{H})$. If $f = \iota_C$ for some closed convex set $C \subset \mathcal{H}$, $\text{prox}_f(x)$ is equivalent to orthogonal projection onto C . For $f(x) = \|x\|_1$, $\text{prox}_{\gamma f}(x)$ is given by component-wise soft-thresholding of x by threshold γ [18]. In addition, proximity operators of lsc convex functions exhibit nice properties with respect to translation, composition with frame operators, dilation, etc. [40], [38].

In problem (\mathbf{P}) with $f = f_1 + f_2$, the resolvent operator $\mathcal{J}_{\beta \partial f} = (\mathbb{1} + \beta \partial f)^{-1}$ typically cannot be calculated in closed-form. Monotone operator splitting methods do not attempt to evaluate this resolvent mapping directly, but instead perform a sequence of calculations involving separately the individual

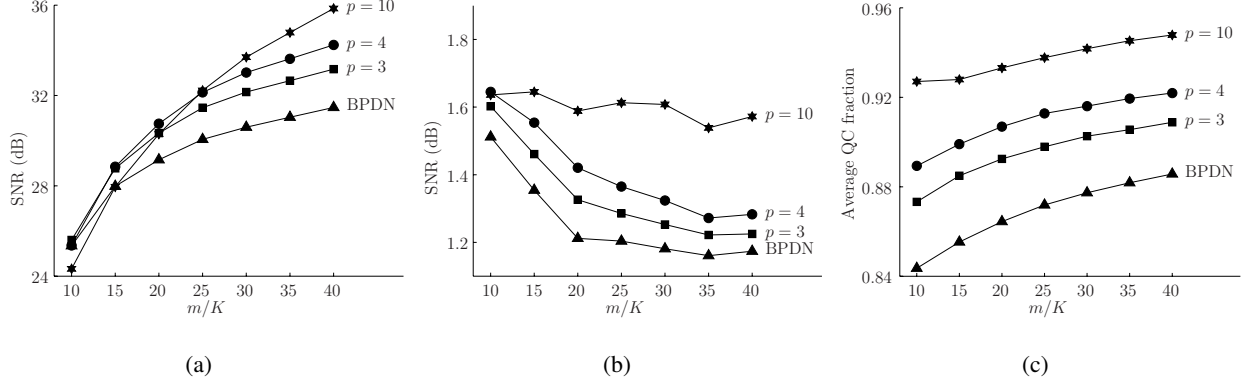


Fig. 1. Quality of BPDQ_p for different m/K and p . Mean (a) and standard deviation (b) of SNR. (c) Fraction of coefficients satisfying QC.

resolvent operators $\mathcal{J}_{\beta\partial f_1}$ and $\mathcal{J}_{\beta\partial f_2}$. The latter are hopefully easier to evaluate, and this holds true for our functionals in BPDQ_p.

Since for BPDQ_p, both f_1 and f_2 are non-differentiable, we use a particular monotone operator splitting method known as the Douglas-Rachford (DR) splitting. It can be written as the following compact recursion formula [18]

$$x^{(t+1)} = (1 - \frac{\alpha_t}{2}) x^{(t)} + \frac{\alpha_t}{2} S_\gamma^\odot \circ \mathcal{P}_{T_p(\epsilon)}^\odot(x^{(t)}), \quad (11)$$

where $A^\odot \triangleq 2A - \mathbb{1}$ for any operator A , $\alpha_t \in (0, 2)$ for all $t \in \mathbb{N}$, $S_\gamma = \text{prox}_{\gamma f_1}$ is the component-wise soft-thresholding operator with threshold $\gamma > 0$ and $\mathcal{P}_{T_p(\epsilon)} = \text{prox}_{f_2}$ is the orthogonal projection onto the tube $T^p(\epsilon)$. From [19], one can show that the sequence $(x^{(t)})_{t \in \mathbb{N}}$ converges to some point x^* and $\mathcal{P}_{T_p(\epsilon)}(x^*)$ is a solution of BPDQ_p. In the next Section, we provide a way to compute $\mathcal{P}_{T_p(\epsilon)}(x^*)$ efficiently.

B. Proximity operator of the ℓ_p fidelity constraint

Each step of the DR iteration (11) requires computation of $\text{prox}_{f_2} = \mathcal{P}_{T^p(\epsilon)}$ for $T^p(\epsilon) = \{x \in \mathbb{R}^N : \|y_q - \Phi x\|_p \leq \epsilon\}$. We present an iterative method to compute this projection for $2 \leq p \leq \infty$.

Notice first that, defining the unit ℓ_p ball $B^p = \{y \in \mathbb{R}^m : \|y\|_p \leq 1\} \subset \mathbb{R}^m$, we have

$$f_2(x) = \iota_{T^p(\epsilon)}(x) = (\iota_{B^p} \circ A_\epsilon)(x),$$

with the affine operator $A_\epsilon(x) \triangleq \frac{1}{\epsilon}(\Phi x - y_q)$.

The proximity operator of a pre-composition of a function $f \in \Gamma_0(\mathcal{H})$ with an affine operator can be computed from the proximity operator of f . Indeed, let $\Phi' \in \mathbb{R}^{m \times N}$ and the affine operator $A(x) \triangleq \Phi'x - y$ with $y \in \mathbb{R}^m$. If Φ' is a tight frame of \mathcal{H} , i.e., $\Phi'\Phi'^* = c\mathbb{1}$ for some $c > 0$, we have

$$\text{prox}_{f \circ A}(x) = x + c^{-1}\Phi'^*(\text{prox}_{cf} - \mathbb{1})(A(x)) ,$$

[40], [18]. Moreover, for a general bounded matrix Φ' , we can use the following lemma.

Lemma 4 ([18]). *Let $\Phi' \in \mathbb{R}^{m \times N}$ be a matrix with bounds $0 \leq c_1 < c_2 < \infty$ such that $c_1 \mathbb{1} \leq \Phi'\Phi'^* \leq c_2 \mathbb{1}$ and let $\{\beta_t\}_{t \in \mathbb{N}}$ be a sequence with $0 < \inf_t \beta_t \leq \sup_t \beta_t < 2/c_2$. Define*

$$\begin{aligned} u^{(t+1)} &= \beta_t(\mathbb{1} - \text{prox}_{\beta_t^{-1}f})(\beta_t^{-1}u^{(t)} + A(p^{(t)})), \\ p^{(t+1)} &= x - \Phi'^*u^{(t+1)}. \end{aligned} \tag{12}$$

If the matrix Φ' is a general frame of \mathcal{H} , i.e., $0 < c_1 < c_2 < \infty$, then $f \circ A \in \Gamma_0(\mathcal{H})$. In addition, $u^{(t)} \rightarrow \bar{u} \in \mathbb{R}^m$ and $p^{(t)} \rightarrow \text{prox}_{f \circ A}(x) = x - \Phi'^\bar{u}$ in (12). More precisely, both $u^{(t)}$ and $p^{(t)}$ converge linearly and the best convergence rate is attained for $\beta_t \equiv 2/(c_1 + c_2)$ with $\|u^{(t)} - \bar{u}\| \leq \left(\frac{c_2 - c_1}{c_2 + c_1}\right)^t \|u^{(0)} - \bar{u}\|$. Otherwise, if Φ' is just bounded (i.e., $c_1 = 0 < c_2 < \infty$), and if $f \circ A \in \Gamma_0(\mathcal{H})$, apply (12), and then $u^{(t)} \rightarrow \bar{u}$ and $p^{(t)} \rightarrow \text{prox}_{f \circ A}(x) = x - \Phi'^*\bar{u}$ at the rate $O(1/t)$.*

In conclusion, computing prox_{f_2} may be reduced to applying the orthogonal projection $\text{prox}_{\iota_{B^p}} = \mathcal{P}_{B^p}$ by setting $f = \iota_{B^p}$, $\Phi' = \Phi/\epsilon$ and $y = y_q/\epsilon$ inside the iterative method (12) with a number of iterations depending on the selected application (see Section VII).

For $p = 2$ and $p = \infty$, the projector \mathcal{P}_{B^p} has an explicit form. Indeed, if y is outside the closed unit ℓ_p -ball in \mathbb{R}^m , then $\mathcal{P}_{B^2}(y) = \frac{y}{\|y\|_2}$; and $(\mathcal{P}_{B^\infty}(y))_i = \text{sign}(y_i) \times \min\{1, |y_i|\}$ for $1 \leq i \leq m$.

Unfortunately, for $2 < p < \infty$ no known closed-form for the projection exists. Instead, we describe an iterative method. Set $f_y(u) = \frac{1}{2}\|u - y\|_2^2$ and $g(u) = \|u\|_p^p$.

If $\|y\|_p \leq 1$, $\mathcal{P}_{B^p}(y) = y$. For $\|y\|_p > 1$, the projection \mathcal{P}_{B^p} is the solution of the constrained minimization problem $u^* = \arg \min_u f_y(u)$ s.t. $g(u) = 1$. Let $L(u, \lambda)$ be its Lagrange function (for $\lambda \in \mathbb{R}$)

$$L(u, \lambda) = f_y(u) + \lambda(g(u) - 1). \tag{13}$$

Without loss of generality, by symmetry, we may work in the positive⁶ orthant $u_i \geq 0$ and $y_i \geq 0$, since the point y and its projection u^* belong to the same orthant of \mathbb{R}^m , i.e., $y_i u_i^* \geq 0$ for all $1 \leq i \leq m$.

⁶The general solution can be obtained by appropriate axis mirroring.

As f_y and g are continuously differentiable, the Karush-Kuhn-Tucker system corresponding to (13) is

$$\begin{aligned}\nabla_u L(u^*, \lambda^*) &= \nabla_u f_y(u^*) + \lambda^* \nabla_u g(u^*) = 0 \\ \nabla_\lambda L(u^*, \lambda^*) &= g(u^*) - 1 = 0,\end{aligned}\tag{14}$$

where the solution u^* is non-degenerate by strict convexity in u [41], and λ^* the corresponding Lagrange multiplier.

Let us write $z = (u, z_{m+1} = \lambda) \in \mathbb{R}^{m+1}$ and $F = \nabla_z L : \mathbb{R}^{m+1} \rightarrow \mathbb{R}^{m+1}$ as

$$F_i(z) = \begin{cases} z_i + p z_{m+1} z_i^{p-1} - y_i & \text{if } i \leq m, \\ (\sum_{j=1}^m z_j^p) - 1 & \text{if } i = m+1. \end{cases}$$

The KKT system (14) is equivalent to $F(z^*) = 0$, where the desired projection u^* is then given by the first m coordinates of z^* . This defines a system of $m+1$ equations with $m+1$ unknowns (u^*, λ^*) that we can solve efficiently with the Newton method. This is the main strategy underlying sequential quadratic programming used to solve general-type constrained optimization problems [41].

Given an initialization point z^0 , the successive iterates are defined by

$$z^{n+1} = z^n - V(z^n)^{-1} F(z^n),\tag{15}$$

where $V_{ij} = \frac{\partial F_i}{\partial z_j}$ is the Jacobian associated to F . If the iterates sequence $(z^n)_{n \geq 0}$ is close enough to (u^*, λ^*) , we know that the Jacobian is nonsingular as u^* is non-degenerate. Moreover, since that Jacobian has a simple block-invertible form, we may compute ([42], p.125)

$$V^{-1}(z) = \frac{1}{\mu} \begin{pmatrix} \mu D^{-1} u + (z_{m+1} - \bar{b}^T u) \bar{b} \\ (\bar{b}^T u - z_{m+1}) \end{pmatrix},\tag{16}$$

where $D \in \mathbb{R}^{m \times m}$ is a diagonal matrix with $D_{ii}(z) = 1 + p(p-1)z_{m+1}z_i^{p-2}$, $b \in \mathbb{R}^m$ with $b_i(z) = pz_i^{p-1}$ for $1 \leq i \leq m$, $\bar{b} = D^{-1}b$, $\mu = b^T D^{-1}b = \bar{b}^T D \bar{b}$. This last expression can be computed efficiently as D is diagonal.

We initialize the first m components of z^0 by the direct radial projection of y on the unit ℓ_p -ball, $u^0 = y/\|y\|_p$, and initialize $z_{m+1}^0 = \arg \min_\lambda \|F(u^0, \lambda)\|_2$.

In summary, to compute \mathcal{P}_{B^p} , we run (15) using (16) to calculate each update step. We terminate the iteration when the norm of $\|F(z^n)\|_2$ falls below a specified tolerance. Since the Newton method converges superlinearly, we obtain error comparable to machine precision with typically fewer than 10 iterations.

VII. EXPERIMENTS

As an experimental validation of the BPDQ_p method, we ran two sets of numerical simulations for reconstructing signals from quantized measurements. For the first experiment we studied recovery of exactly sparse random 1-D signals, following very closely our theoretical developments. Setting the dimension $N = 1024$ and the sparsity level $K = 16$, we generated 500 K -sparse signals where the non-zero elements were drawn from the standard Gaussian distribution $\mathcal{N}(0, 1)$, and located at supports drawn uniformly in $\{1, \dots, N\}$. For each sparse signal x , m quantized measurements were recorded as in model (1) with a SGR matrix $\Phi \in \mathbb{R}^{m \times N}$. The bin width was set to $\alpha = \|\Phi x\|_\infty / 40$.

The decoding was accomplished with BPDQ_p for various moments $p \geq 2$ using the optimization algorithm described in Section VI. In particular, the overall Douglas-Rachford procedure (11) was run for 500 iterations. At each DR step, the method in (12) was iterated until the relative error $\frac{\|p^{(t)} - p^{(t-1)}\|_2}{\|p^{(t)}\|_2}$ fell below 10^{-6} ; the required number of iterations was dependent on m but was fewer than 700 in all cases examined.

In Figure 1, we plot the average quality of the reconstructions of BPDQ_p for various values of $p \geq 2$ and $m/K \in [10, 40]$. We use the quality measure $\text{SNR}(\hat{x}; x) = 20 \log_{10} \frac{\|x\|_2}{\|x - \hat{x}\|_2}$, where x is the true original signal and \hat{x} the reconstruction. As can be noticed, at higher oversampling factors m/K the decoders with higher p give better reconstruction performance. Equivalently, it can also be observed that at lower oversampling factors, increasing p beyond a certain point degrades the reconstruction performance. These two effects are consistent with the remarks noted at the end of Section V, as the sensing matrices may fail to satisfy the RIP_{p,2} if p is too large for a given oversampling factor.

One of the original motivations for the BPDQ_p decoders is that they are closer to enforcing quantization consistency than the BPDN decoder. To check this, we have examined the “quantization consistency fraction”, *i.e.*, the average fraction of remeasured coefficients $(\Phi \hat{x})_i$ that satisfy $|(\Phi \hat{x})_i - y_i| < \frac{\alpha}{2}$. These are shown in Figure 1 (c) for various p and m/K . As expected, it can be clearly seen that increasing p increases the QC fraction.

An even more explicit illustration of this effect is afforded by examining histograms of the normalized residual $\alpha^{-1}(\Phi \hat{x} - y)_i$ for different p . For reconstruction exactly satisfying QC, these normalized residuals should be supported on $[-1/2, 1/2]$. In Figure 2 we show histograms of normalized residuals for $p = 2$ and $p = 10$, for the case $m/K = 40$. The histogram for $p = 10$ is indeed closer to uniform on $[-1/2, 1/2]$.

For the second experiment, we apply a modified version of the BPDQ_p to an undersampled MRI

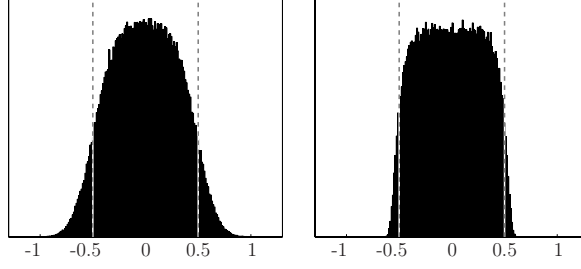


Fig. 2. Histograms of $\alpha^{-1}(\Phi \hat{x} - y)_i$. Left, $p = 2$. Right, $p = 10$.

reconstruction problem. Using an example similar to [43], the original image is a 256×256 pixel “synthetic angiogram”, *i.e.*, $N = 256^2$, comprised of 10 randomly placed non-overlapping ellipses. The linear measurements are the real and imaginary parts of a fraction ρ of the Fourier coefficients at randomly selected locations in Fourier space, giving $m = \rho N$ independent measurements. These random locations form the index set $\Omega \subset \{1, \dots, N\}$ with $|\Omega| = m$. Experiments were carried out with $\rho \in \{1/6, 1/8, 1/12\}$, but we show results only for $\rho = 1/8$. These were quantized with a bin width $\alpha = 50$, giving at most 12 quantization levels for each measurement.

For this example, we modify the BPDQ $_p$ program III by replacing the ℓ_1 term by the total variation (TV) semi-norm [44]. This yields the problem

$$\operatorname{argmin}_u \|u\|_{TV} \quad \text{s.t.} \quad \|y - \Phi u\|_p \leq \epsilon,$$

where $\Phi = F_\Omega$ is the restriction of Discrete Fourier Transform matrix F to the rows indexed in Ω .

This may be solved with the Douglas-Rachford iteration (11), with the modification that S_γ be replaced by the proximity operator associated to γ times the TV norm, *i.e.*, by $\operatorname{prox}_{\gamma \|\cdot\|_{TV}}(y) = \operatorname{argmin}_u \frac{1}{2} \|y - u\|^2 + \gamma \|u\|_{TV}$. The latter is known as the Rudin-Osher-Fatemi model, and numerous methods exist for solving it exactly, including [45], [46], [47], [48]. In this work, we use an efficient projected gradient descent algorithm on the dual problem, see *e.g.*, [18]. Note that the sensing matrix F_Ω is actually a tight frame, *i.e.*, $F_\Omega F_\Omega^* = \mathbb{I}$, so we do not need the nested inner iteration (12).

We show the SNR of the BPDQ $_p$ reconstructions as a function of p in Figure 3, averaged over 50 trials where both the synthetic angiogram image and the Fourier measurement locations are randomized. This figure also depicts the SNR improvement of BPDQ $_p$ -based reconstruction over BPDN. For these simulations we used 500 iterations of the Douglas-Rachford recursion (11). This quantitative results are

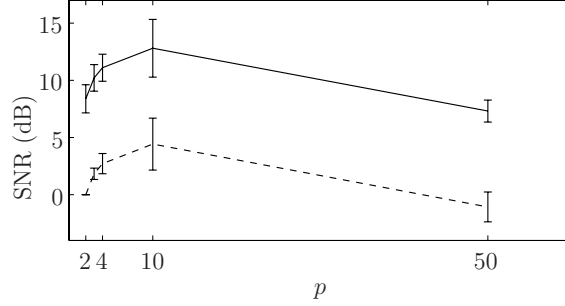


Fig. 3. Average SNR (solid) and SNR improvement over BPDN (dashed) as a function of p , for the synthetic angiogram reconstruction simulations. Error bars indicate 1 standard deviation.

confirmed by visual inspection of Figure 4, where we compare 100×100 pixel details of the reconstruction results with BPDN and with BPDQ_p for $p = 10$, for one particular instance of the synthetic angiogram signal.

Note that this experiment lies far outside of the justification provided by our theoretical developments, as we do not have any proof that the sensing matrix F_Ω satisfies the $\text{RIP}_{p,2}$, and our theory was developed only for ℓ_1 synthesis-type regularization, while the TV regularization is of analysis type. Nonetheless, we obtain results analogous to the previous 1-D example; the BPDQ_p reconstruction shows improvements both in SNR and visual quality compared to BPDN. These empirical results suggest that the BPDQ_p method may be useful for a wider range of quantized reconstruction problems, and also provoke interest for further theoretical study.

VIII. CONCLUSION AND FURTHER WORK

The objective of this paper was to show that the BPDN reconstruction program commonly used in Compressed Sensing with noisy measurements is not always adapted to quantization distortion. We introduced a new class of decoders, the Basis Pursuit DeQuantizers, and we have shown both theoretically and experimentally that BPDQ_p exhibit a substantial reduction of the reconstruction error in oversampled situations.

A first interesting question for further study would be to characterize the evolution of the optimal moment p with the oversampling ratio. This would allow for instance the selection of the best BPDQ_p decoder in function of the precise CS coding/decoding scenario. Second, it is also worth investigating the existence of other $\text{RIP}_{p,2}$ random matrix constructions, *e.g.*, using the Random Fourier Ensemble.

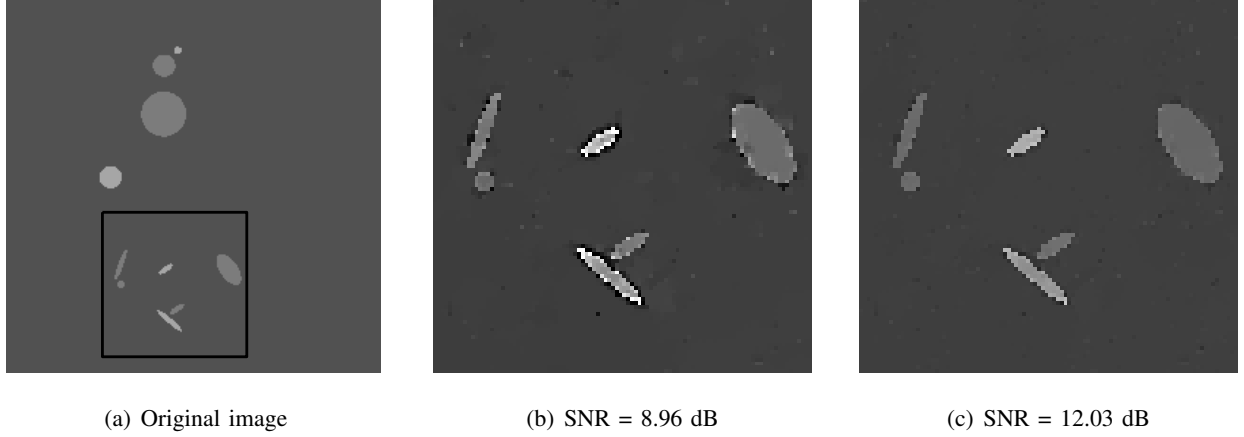


Fig. 4. Reconstruction of synthetic angiograms from undersampled Fourier measurements, using TV regularization. (a) Original, showing zoom area (b) BPDN (zoom) (c) BPDQ₁₀ (zoom).

Third, a more realistic coding/decoding scenario should set α theoretically in function of the bit budget (rate) available to quantize the measurements, of the sensing matrix and of some a priori on the signal energy. This should be linked also to the way our approach can integrate the saturation of the quantized measurements [16]. Finally, we would like to extend our approach to non-uniform scalar quantization of random measurements, generalizing the quantization consistency and the optimization fidelity term to this more general setting.

IX. ACKNOWLEDGMENTS

LJ and DKH are very grateful to Prof. Pierre Vandergheynst (Signal Processing Laboratory, LTS2/EPFL, Switzerland) for his useful advices and his hospitality during their postdoctoral stay in EPFL.

APPENDIX A

PROOF OF PROPOSITION 1

Before proving Proposition 1, let us recall some facts of measure concentrations [49], [50].

In particular, we are going to use the concentration property of any Lipschitz function over \mathbb{R}^m , i.e., F such that $\|F\|_{\text{Lip}} \triangleq \sup_{u,v \in \mathbb{R}^m, u \neq v} \frac{|F(u) - F(v)|}{\|u - v\|_2} < \infty$. If $\|F\|_{\text{Lip}} \leq 1$, F is said 1-Lipschitz.

Lemma 5 (Ledoux, Talagrand [49] (Eq. 1.6)). *If F is Lipschitz with $\lambda = \|F\|_{\text{Lip}}$, then, for the random*

vector $\xi \in \mathbb{R}^m$ with $\xi_i \sim_{\text{iid}} \mathcal{N}(0, 1)$,

$$P_\xi[|F(\xi) - \mu_F| > r] \leq 2e^{-\frac{1}{2}r^2\lambda^{-2}}, \quad \text{for } r > 0,$$

with $\mu_F = \mathbb{E}F(\xi) = \int_{\mathbb{R}^m} F(x) \gamma^m(x) \mathrm{d}^m x$ and $\gamma^m(x) = (2\pi)^{-m/2} e^{-\|x\|_2^2/2}$.

A useful tool that we will use is the concept of a *net*. An ϵ -net ($\epsilon > 0$) of $A \subset \mathbb{R}^K$ is a subset \mathcal{S} of A such that for every $t \in A$, one can find $s \in \mathcal{S}$ with $\|t - s\|_2 \leq \epsilon$. In certain cases, the size of a ϵ -net can be bounded.

Lemma 6 ([50]). *There exists a ϵ -net \mathcal{S} of the unit sphere of \mathbb{R}^K of size $|\mathcal{S}| \leq (1 + \frac{2}{\epsilon})^K$.*

We will use also this fundamental result.

Lemma 7 ([50]). *Let \mathcal{S} be a ϵ -net of the unit sphere in \mathbb{R}^K . Then, if for some vectors v_1, \dots, v_K in the Banach space B normed by $\|\cdot\|_B$, we have $1 - \epsilon \leq \|\sum_{i=1}^K s_i v_i\|_B \leq 1 + \epsilon$ for all $s = (s_1, \dots, s_K) \in \mathcal{S} \subset \mathbb{R}^K$, then*

$$(1 - \beta) \|t\|_2 \leq \left\| \sum_{i=1}^K t_i v_i \right\|_B \leq (1 + \beta) \|t\|_2,$$

for all $t \in \mathbb{R}^K$, with $\beta = \frac{2\epsilon}{1-\epsilon}$.

In our case, the Banach space B is $\ell_p^m = (\mathbb{R}^m, \|\cdot\|_p)$ for $1 \leq p \leq \infty$, i.e \mathbb{R}^m equipped with the norm $\|u\|_p^p = \sum_i |u_i|^p$. With all these concepts, we can now demonstrate the main proposition.

Proof of Proposition 1: Let $p \geq 1$. We must prove that for a SGR matrix $\Phi \in \mathbb{R}^{m \times N}$, i.e., with $\Phi_{ij} \sim_{\text{iid}} \mathcal{N}(0, 1)$, with the right number of measurements m , there exist a radius $0 < \delta < 1$ and a constant $\mu_{p,2} > 0$ such that

$$\mu_{p,2} \sqrt{1 - \delta} \|x\|_2 \leq \|\Phi x\|_p \leq \mu_{p,2} \sqrt{1 + \delta} \|x\|_2, \quad (17)$$

for all $x \in \mathbb{R}^N$ with $\|x\|_0 \leq K$.

We begin with a unit sphere $S_T = \{u \in \mathbb{R}^N : \text{supp } u = T, \|u\|_2 = 1\}$ for a fixed support $T \subset \{1, \dots, N\}$ of size $|T| = K$. Let \mathcal{S}_T be an ϵ -net of S_T . We consider the SGR random process that generates Φ and, by an abuse of notation, we identify it for a while with Φ itself. In other words, we define the random matrix $\Phi = (\Phi_1, \dots, \Phi_N) \in \mathbb{R}^{m \times N}$ where, for all $1 \leq i \leq N$, $\Phi_j \in \mathbb{R}^m$ is a random vector of probability density function (or *pdf*) $\gamma^m(u) = \prod_{i=1}^m \gamma(u_i)$ for $u \in \mathbb{R}^m$ and

$\gamma(u_i) = \frac{1}{\sqrt{2\pi}} e^{-u_i^2/2}$ (the standard Gaussian pdf). Therefore, Φ is related to the pdf $\gamma_\Phi(\phi) = \prod_{j=1}^N \gamma^m(\phi_j)$, $\phi = (\phi_1, \dots, \phi_N) \in \mathbb{R}^{m \times N}$.

Since the Frobenius norm $\|\phi\|_{\mathcal{F}} = (\sum_{jk} |\phi_{jk}|^2)^{1/2}$ of ϕ and the pdf $\gamma_\Phi(\phi) \propto e^{-\|\phi\|_{\mathcal{F}}^2/2}$ are invariant under a global rotation in \mathbb{R}^N of all the rows of ϕ , it is easy to show that for unit vector $s \in \mathbb{R}^N$, $P_\Phi[|F(\Phi s) - \mu_F| > r] = P_\Phi[|F(\Phi_1) - \mu_F| > r] \leq 2e^{-\frac{1}{2}r^2\lambda^{-2}}$, using Lemma 5 on the SGR vector Φ_1 .

The above holds for a single s . To obtain a result valid for all $s \in \mathcal{S}_T$ we may use the union bound. As $|\mathcal{S}_T| \leq (1 + 2/\epsilon)^K$ by Lemma 6, setting $r = \epsilon\mu_F$ for $\epsilon > 0$, we obtain

$$P_\Phi[|\mu_F^{-1} F(\Phi s) - 1| > \epsilon] \leq 2e^{K \log(1+2\epsilon^{-1}) - \frac{1}{2}\epsilon^2\mu_F^2\lambda^{-2}},$$

for all $s \in \mathcal{S}_T$.

Taking now $F(\cdot) = \|\cdot\|_p$ for $1 \leq p \leq \infty$, we have $\mu_F = \mu_{p,2} = \mathbb{E}\|\xi\|_p$ for a SGR vector $\xi \in \mathbb{R}^m$. The Lipschitz value is $\lambda = \lambda_p = 1$ for $p \geq 2$, and $\lambda = \lambda_p = m^{\frac{2-p}{2p}}$ for $1 \leq p \leq 2$. Consequently,

$$(1 - \epsilon) \leq \left\| \frac{1}{\mu_{p,2}} \Phi s \right\|_p \leq (1 + \epsilon), \quad (18)$$

for all $s \in \mathcal{S}_T$, with a probability higher than $1 - 2 \exp(K \log(1 + 2\epsilon^{-1}) - \frac{1}{2}\epsilon^2\mu_{p,2}^2\lambda_p^{-2})$.

We apply Lemma 7 by noting that, as s has support of size K , (18) may be written as

$$1 - \epsilon \leq \left\| \sum_{i=1}^K s_i v_i \right\|_p \leq 1 + \epsilon$$

where $v_i \in \mathbb{R}^m$ are the columns of $\frac{1}{\mu_{p,2}} \Phi$ corresponding to the support of s (we abuse notation to let s_i range only over the support of s). Then according to Lemma 7 we have, with the same probability bound and for $(\sqrt{2} - 1)\delta = \frac{2\epsilon}{1-\epsilon}$,

$$\begin{aligned} \sqrt{1-\delta} \|x\|_2 &\leq (1 - (\sqrt{2} - 1)\delta) \|x\|_2 \leq \|\Phi x\|_p \\ &\leq (1 + (\sqrt{2} - 1)\delta) \|x\|_2 \leq \sqrt{1+\delta} \|x\|_2, \end{aligned} \quad (19)$$

for all $x \in \mathbb{R}^N$ with $\text{supp } x = T$.

The result can be made independent of the choice of $T \subset \{1, \dots, N\}$ by considering that there are $\binom{N}{K} \leq (eN/K)^K$ such possible supports. Therefore, applying again an union bound, (19) holds for all K -sparse x in \mathbb{R}^N with a probability higher than $1 - 2e^{-\frac{1}{2}\epsilon^2\mu_{p,2}^2\lambda_p^{-2} + K \log[e \frac{N}{K} (1+2\epsilon^{-1})]}$.

Let us bound this probability first for $1 \leq p < \infty$. For $m \geq \beta^{-1} 2^{p+1}$ and $\beta^{-1} = p - 1$, Lemma 1 (page 9) tells us that $\mu_{p,2} \geq \frac{p-1}{p} \nu_p m^{\frac{1}{p}}$ with $\nu_p = \sqrt{2} \pi^{-\frac{1}{2p}} \Gamma(\frac{p+1}{2})^{\frac{1}{p}}$. A probability of success $1 - \eta$ with

$\eta < 1$ is then guaranteed if we select, for $1 \leq p < 2$,

$$m > \frac{2}{\epsilon^2 \nu_p^2} \left(\frac{p}{p-1} \right)^2 \left(K \log \left[e^{\frac{N}{K}} (1 + 2\epsilon^{-1}) \right] + \log \frac{2}{\eta} \right),$$

since $\lambda_p = m^{\frac{2-p}{2p}}$, and for $2 \leq p < \infty$,

$$m^{\frac{2}{p}} > \frac{2}{\epsilon^2 \nu_p^2} \left(\frac{p}{p-1} \right)^2 \left(K \log \left[e^{\frac{N}{K}} (1 + 2\epsilon^{-1}) \right] + \log \frac{2}{\eta} \right), \quad (20)$$

since $\lambda_p = 1$.

From now, $A \geq cB$ or $A \leq cB$ means that there exists a constant $c > 0$ such that these inequalities hold. According to the lower bound found in Section V, $\nu_p > c\sqrt{p+1}$ implying that $\nu_p^{-2} \leq c$. Since $(p/(p-1))^2 \leq 4$ for any $p \geq 2$ and $\epsilon^{-1} \leq \frac{\sqrt{2}+1}{\sqrt{2}-1} \delta^{-1} \leq 6\delta^{-1}$, we find the new sufficient conditions,

$$m > c\delta^{-2} \left(\frac{p}{p-1} \right)^2 \left(K \log \left[e^{\frac{N}{K}} (1 + 12\delta^{-1}) \right] + \log \frac{2}{\eta} \right),$$

for $1 \leq p < 2$, and

$$m^{2/p} > c\delta^{-2} \left(K \log \left[e^{\frac{N}{K}} (1 + 12\delta^{-1}) \right] + \log \frac{2}{\eta} \right),$$

for $2 \leq p < \infty$.

Second, in the specific case where $p = \infty$, since there exists a $\rho > 0$ such that $\mu_{\infty,2} \geq \rho^{-1} \sqrt{\log m}$, with $\lambda_{\infty} = 1$, $\log m > c\delta^{-2} \left(K \log \left[e^{\frac{N}{K}} (1 + 12\delta^{-1}) \right] + \log \frac{2}{\eta} \right)$. ■

Let us make some remarks about the results and the requirements of the last proposition. Notice first that for $p = 2$, we find the classical result proved in [23]. Second, as for the comparison between the common $\text{RIP}_{2,2}$ proof [23] and the tight bound found in [24], the requirements on the measurements above are possibly pessimistic, *i.e.*, the exponent $2/p$ occurring in (20) is perhaps too small. Proposition 1 has however the merit to prove that random Gaussian matrices satisfy the $\text{RIP}_{p,2}$ in a certain range of dimensionality.

APPENDIX B

LINK BETWEEN δ AND m FOR SGR $\text{RIP}_{p,2}$ MATRICES

For $2 \leq p < \infty$, Proposition 1 shows that, if $\delta^2 \geq cm^{-2/p} \left(K \log \left[e^{\frac{N}{K}} (1 + 12\delta^{-1}) \right] + \log \frac{2}{\eta} \right)$ for a certain constant $c > 0$, a SGR matrix $\Phi \in \mathbb{R}^{m \times N}$ is $\text{RIP}_{p,2}$ of order K and radius $0 < \delta < 1$ with a probability higher than $1 - \eta$. Let us assume that $\delta > dm^{-1/p}$ for some $d > 0$. We have, $\log \delta^{-1} < \frac{1}{p} \log m - \log d$, and therefore, the same event occurs with the same probability bound when $\delta^2 \geq cm^{-2/p} \left(K \log \left[13e^{\frac{N}{K}} \right] + \frac{K}{p} \log m - K \log d + \log \frac{2}{\eta} \right)$. For high m and for fixed K, N and η , this provides $\delta = O(m^{-1/p} \sqrt{\log m})$, which meets the previous assumption.

APPENDIX C

Proof of Lemma 1: The result for $p = \infty$ is due to [49] (see Eq (3.14)). Let $\xi \in \mathbb{R}^m$ be a SGR vector, i.e., $\xi_i \sim_{iid} \mathcal{N}(0, 1)$ for $1 \leq i \leq m$, and $1 \leq p < \infty$. First, the inequality $\mathbb{E}\|\xi\|_p \leq (\mathbb{E}\|\xi\|_p^p)^{1/p}$ follows from the application of the Jensen inequality $\varphi(\mathbb{E}\|\xi\|_p) \leq \mathbb{E}\varphi(\|\xi\|_p)$ with the convex function $\varphi(\cdot) = (\cdot)^p$. Second, the lower bound on $\mathbb{E}\|\xi\|_p$ arises from the observation that for $f : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ with $f(t) = t^{\frac{1}{p}}$, and for a given $t_0 > 0$,

$$f(t) \geq f(t_0) + f'(t_0)(t - t_0) + pf''(t_0)(t - t_0)^2, \quad (21)$$

for all $t \geq 0$.

Indeed, observe first that since $f^{(n)}(\alpha t') = \alpha^{\frac{1}{p}-n} f^{(n)}(t')$ for $\alpha > 0$ and $n \in \mathbb{N}$, it is sufficient to prove the result for $t_0 = 1$. Proving (21) amounts then to prove $f(t) = t^{\frac{1}{p}} \geq \frac{2p-1}{p}t - \frac{p-1}{p}t^2$, or equivalently, $t^{\frac{1}{p}-1} + \frac{p-1}{p}t \geq \frac{2p-1}{p}$. The LHS of this last inequality takes its minimum in $t = 1$ with value $\frac{2p-1}{p}$, which provides the result.

Since $\mu_{p,2} = \mathbb{E}\|\xi\|_p = \mathbb{E}f(\|\xi\|_p^p)$ and $\mathbb{E}(\|\xi\|_p^p - \bar{\mu}_{p,2}) = 0$, using (21) we find

$$\mu_{p,2} \geq (t_0)^{\frac{1}{p}-2} \left((2 - \frac{1}{p})\bar{\mu}_{p,2}t_0 + (\frac{1}{p} - 1)(\bar{\mu}_{p,2}^2 + \bar{\sigma}_p^2) \right)$$

writing $\bar{\mu}_{p,2} = \mathbb{E}\|\xi\|_p^p$ and $\bar{\sigma}_p^2 = \mathbb{E}(\|\xi\|_p^p - \bar{\mu}_{p,2})^2 = \text{Var}\|\xi\|_p^p$. The RHS of the last inequality is maximum for $t_0 = \bar{\mu}_{p,2} (1 + \bar{\mu}_{p,2}^{-2} \bar{\sigma}_p^2)$. For that value, we get finally

$$\mu_{p,2} \geq (\mathbb{E}\|\xi\|_p^p)^{\frac{1}{p}} \left(1 + (\mathbb{E}\|\xi\|_p^p)^{-2} \text{Var}\|\xi\|_p^p \right)^{\frac{1}{p}-1}.$$

Because of the decorrelation of the components of ξ , the last inequality simplifies into

$$\mu_{p,2} \geq m^{\frac{1}{p}} (\mathbb{E}|g|^p)^{\frac{1}{p}} \left(1 + m^{-1} (\mathbb{E}|g|^p)^{-2} \text{Var}|g|^p \right)^{\frac{1}{p}-1},$$

with $g \sim \mathcal{N}(0, 1)$.

Moreover, since $\mathbb{E}|g|^p = 2^{\frac{p}{2}} \pi^{-\frac{1}{2}} \Gamma(\frac{p+1}{2})$ and using the following approximation of the Gamma function [37] $|\Gamma(x) - (\frac{2\pi}{x})^{\frac{1}{2}} (\frac{x}{e})^x| \leq \frac{1}{9x} (\frac{2\pi}{x})^{\frac{1}{2}} (\frac{x}{e})^x$, valid for $x \geq 1$, we observe that

$$0.9 \left(\frac{2\pi}{x} \right)^{\frac{1}{2}} \left(\frac{x}{e} \right)^x \leq \Gamma(x) \leq 1.1 \left(\frac{2\pi}{x} \right)^{\frac{1}{2}} \left(\frac{x}{e} \right)^x,$$

that holds also if $x = \frac{p+1}{2}$ with $p \geq 1$. Therefore, $(\mathbb{E}|g|^p)^{-2} \text{Var}|g|^p \leq \left(\frac{1.1}{0.9^2} \left(\frac{e}{2} \right)^{\frac{1}{p}} \left(\frac{2p+1}{p+1} \right)^p - 1 \right) \leq \frac{1.1}{0.9^2} \left(\frac{e}{2} \right)^{\frac{1}{2}} 2^p$ and finally

$$\mu_{p,2} \geq m^{\frac{1}{p}} (\mathbb{E}|g|^p)^{\frac{1}{p}} \left(1 + c \frac{2^p}{m} \right)^{\frac{1}{p}-1}$$

for a constant $c = \frac{1.1}{0.9^2} \left(\frac{e}{2} \right)^{\frac{1}{2}} < 1.584 < 2$ independent of p and m . ■

APPENDIX D

Proof of Lemma 2: Notice first that since $J(\lambda w) = \lambda J(w)$ for any $w \in \mathbb{R}^m$ and $\lambda \in \mathbb{R}$, it is sufficient to prove the result for $\|u\|_2 = \|v\|_2 = 1$.

The Lemma relies mainly on the geometrical properties of the Banach space $\ell_p^m = (\mathbb{R}^m, \|\cdot\|_p)$ for $p \geq 2$. In [35], [36], it is explained that this space is p -convex and 2-smooth. The smoothness involves in particular

$$\|x + y\|_p^2 \leq \|x\|_p^2 + 2\langle J(x), y \rangle + (p-1)\|y\|_p^2, \quad (22)$$

where $J = J_2$ and J_r is the *duality* mapping of *gauge function* $t \rightarrow t^{r-1}$ for $r \geq 1$. For the Hilbert space ℓ_2 , the relation (22) reduces of course to the *polarization identity*. For ℓ_p , J_r is the differential of $\frac{1}{r}\|\cdot\|_p^r$, i.e., $(J_r(u))_i = \|u\|^{r-p} |u_i|^{p-1} \text{sign } u_i$.

The smoothness inequality (22) involves

$$2\langle J(x), y \rangle \leq \|x\|_p^2 + (p-1)\|y\|_p^2 - \|x - y\|_p^2, \quad (23)$$

where we used the change of variable $y \rightarrow -y$.

Let us take $x = \Phi u$ and $y = t\Phi v$ with $\|u\|_0 = s$, $\|v\|_0 = s'$, $\|u\|_2 = \|v\|_2 = 1$, $\text{supp } u \cap \text{supp } v = \emptyset$ and for a certain $t > 0$ that we will set later. Because Φ is assumed $\text{RIP}_{p,2}$ for s , s' and $s + s'$ sparse signals, we deduce

$$2\mu_{p,2}^{-2} t |\langle J(\Phi u), \Phi v \rangle| \leq (1 + \delta_s) + (p-1)(1 + \delta_{s'})t^2 - (1 - \delta_{s+s'})(1 + t^2),$$

where the absolute value on the inner product arises from the invariance of the RIP bound on (23) under the change $y \rightarrow -y$. The value $\mu_{p,2}^{-2} |\langle J(\Phi u), \Phi v \rangle|$ is thus bounded by an expression of type $f(t) = \frac{\alpha + \beta t^2}{t}$ with $\alpha, \beta > 0$ for $p \geq 2$ given by $\alpha = \delta_s + \delta_{s+s'}$ and $\beta = (p-2) + (p-1)\delta_{s'} + \delta_{s+s'}$. Since the minimum of f is $2\sqrt{\alpha\beta}$, we get

$$\mu_{p,2}^{-2} |\langle J(\Phi u), \Phi v \rangle| \leq [(\delta_s + \delta_{s+s'}) (\bar{p} + \bar{p} \delta_{s'} + \delta_{s'} + \delta_{s+s'})]^{\frac{1}{2}}, \quad (24)$$

with $\bar{p} = p - 2 \geq 0$.

In parallel, a change $y \rightarrow x + y$ in (23) provides

$$2\langle J(x), y \rangle \leq -\|x\|_p^2 + (p-1)\|x + y\|_p^2 - \|y\|_p^2,$$

where we used the fact that $\langle J(x), x \rangle = \|x\|_p^2$. By summing this inequality with (23), we have

$$4 \langle J(x), y \rangle \leq (p-2) \|y\|_p^2 + (p-1) \|x+y\|_p^2 - \|x-y\|_p^2.$$

Using the RIP_{p,2} on $x = \Phi u$ and $y = t\Phi v$ as above leads to

$$\begin{aligned} 4\mu_{p,2}^{-2} t |\langle J(\Phi u), \Phi v \rangle| &\leq (1 + \delta_{s'}) \bar{p} t^2 \\ &+ (p-1)(1 + \delta_{s+s'})(1 + t^2) - (1 - \delta_{s+s'})(1 + t^2) \\ &= \bar{p} + p\delta_{s+s'} + (2\bar{p} + \bar{p}\delta_{s'} + p\delta_{s+s'})t^2, \end{aligned}$$

with the same argument as before to explain the absolute value. Minimizing over t as above gives

$$2\mu_{p,2}^{-2} |\langle J(\Phi u), \Phi v \rangle| \leq [(\bar{p} + p\delta_{s+s'}) (2\bar{p} + \bar{p}\delta_{s'} + p\delta_{s+s'})]^{\frac{1}{2}}. \quad (25)$$

Together, (24) and (25) imply

$$\begin{aligned} C_p = \min \{ &[(\delta_s + \delta_{s+s'}) (\delta_{s'} + \delta_{s+s'} + \bar{p}(1 + \delta_{s'}))]^{\frac{1}{2}}, \\ &[(\delta_{s+s'} + \bar{p} \frac{1+\delta_{s+s'}}{2}) (\delta_{s+s'} + \bar{p} \frac{2+\delta_{s'}+\delta_{s+s'}}{2})]^{\frac{1}{2}} \}. \end{aligned}$$

It is easy to check that $C_p = C_p(\Phi, s, s')$ behaves as $\sqrt{(\delta_s + \delta_{s+s'}) (1 + \delta_{s'}) \bar{p}}$ for $\bar{p} \gg \frac{\delta_{s'} + \delta_{s+s'}}{(1 + \delta_{s'})}$, and as $\delta_{s+s'} + \frac{3}{4}(1 + \delta_{s+s'})\bar{p} + O(\bar{p}^2)$ for $p \simeq 2$. ■

APPENDIX E

Proof of Theorem 2: Let us write $x_K^* = x + h$. We have to characterize the behavior of $\|h\|_2$. In the following, for any vector $u \in \mathbb{R}^d$ with $d \in \{m, N\}$, we define u_A as the vector in \mathbb{R}^d equal to u on the index set $A \subset \{1, \dots, d\}$ and 0 elsewhere.

We define $T_0 = \text{supp } x_K$ and a partition $\{T_k : 1 \leq k \leq \lceil (N-K)/K \rceil\}$ of the support of $h_{T_0^c}$. This partition is determined by ordering elements of h off of the support of x_K in decreasing absolute value. We have $|T_k| = K$ for all $k \geq 1$, $T_k \cap T_{k'} = \emptyset$ for $k \neq k'$, and crucially that $|h_j| \leq |h_i|$ for all $j \in T_{k+1}$ and $i \in T_k$.

We start from

$$\|h\|_2 \leq \|h_{T_{01}}\|_2 + \|h_{T_{01}^c}\|_2, \quad (26)$$

with $T_{01} = T_0 \cup T_1$, and we are going to bound separately the two terms of the RHS. In [22], it is proved that

$$\|h_{T_{01}^c}\|_2 \leq \sum_{k \geq 2} \|h_{T_k}\|_2 \leq \|h_{T_{01}}\|_2 + 2e_0(K), \quad (27)$$

with $e_0(K) = \frac{1}{\sqrt{K}} \|x_{T_0^c}\|_1$. Therefore,

$$\|h\|_2 \leq 2\|h_{T_{01}}\|_2 + 2e_0(K).$$

Let us bound now $\|h_{T_{01}}\|_2$ by using the $\text{RIP}_{p,2}$. From the definition of the mapping J , we have

$$\begin{aligned} \|\Phi h_{T_{01}}\|_p^2 &= \langle J(\Phi h_{T_{01}}), \Phi h_{T_{01}} \rangle \\ &= \langle J(\Phi h_{T_{01}}), \Phi h \rangle - \sum_{k \geq 2} \langle J(\Phi h_{T_{01}}), \Phi h_{T_k} \rangle. \end{aligned}$$

By the Hölder inequality with $r = \frac{p}{p-1}$ and $s = p$,

$$\begin{aligned} \langle J(\Phi h_{T_{01}}), \Phi h \rangle &\leq \|J(\Phi h_{T_{01}})\|_r \|\Phi h\|_s \\ &= \|\Phi h_{T_{01}}\|_p \|\Phi h\|_p \leq 2\epsilon \|\Phi h_{T_{01}}\|_p \\ &\leq 2\epsilon \mu_{p,2} (1 + \delta_{2K})^{\frac{1}{2}} \|h_{T_{01}}\|_2, \end{aligned}$$

since $\|\Phi h\|_p \leq \|\Phi x - y\|_p + \|\Phi x_p^* - y\|_p \leq 2\epsilon$. Using Lemma 2, as $h_{T_{01}}$ is $2K$ sparse and h_{T_k} is K sparse, we know that, for $k \geq 2$,

$$|\langle J(\Phi h_{T_{01}}), \Phi h_{T_k} \rangle| \leq \mu_{p,2}^2 C_p \|h_{T_{01}}\|_2 \|h_{T_k}\|_2,$$

with $C_p = C_p(\Phi, 2K, K)$, so that, using again the $\text{RIP}_{p,2}$ of Φ and (27),

$$\begin{aligned} (1 - \delta_{2K}) \mu_{p,2}^2 \|h_{T_{01}}\|_2^2 &\leq \|\Phi h_{T_{01}}\|_p^2 \\ &\leq 2\epsilon \mu_{p,2} (1 + \delta_{2K})^{\frac{1}{2}} \|h_{T_{01}}\|_2 + \mu_{p,2}^2 C_p \|h_{T_{01}}\|_2 \sum_{k \geq 2} \|h_{T_k}\|_2 \\ &\leq 2\epsilon \mu_{p,2} (1 + \delta_{2K})^{\frac{1}{2}} \|h_{T_{01}}\|_2 \\ &\quad + \mu_{p,2}^2 C_p \|h_{T_{01}}\|_2 (\|h_{T_{01}}\|_2 + 2e_0(K)). \end{aligned}$$

After some simplifications, we get finally

$$\|h\|_2 \leq \frac{2(C_p + 1 - \delta_{2K})}{1 - \delta_{2K} - C_p} e_0(K) + \frac{4\sqrt{1 + \delta_{2K}}}{1 - \delta_{2K} - C_p} \frac{\epsilon}{\mu_{p,2}}.$$

■

APPENDIX F

Proof of Lemma 3: For a random variable $u \sim U([- \frac{\alpha}{2}, \frac{\alpha}{2}])$, we compute easily that $\mathbb{E}|u|^p = \frac{\alpha^p}{2^p(p+1)}$ and $\text{Var}|u|^p = \frac{\alpha^{2p}p^2}{2^{2p}(p+1)^2(2p+1)}$. Therefore, for a random vector $\xi \in \mathbb{R}^m$ with components ξ_i independent and identically distributed as u , $\mathbb{E}\|\xi\|_p^p = \frac{\alpha^p}{2^p(p+1)}m$ and $\text{Var}\|\xi\|_p^p = \frac{\alpha^{2p}p^2}{2^{2p}(p+1)^2(2p+1)}m$.

To prove the probabilistic inequality below (6), we define, for $1 \leq i \leq m$, the positive random variables $Z_i = \frac{2^p}{\alpha^p}|\xi_i|^p$ bounded on the interval $[0, 1]$ with $\mathbb{E}Z_i = (p+1)^{-1}$. Denoting $S = \frac{1}{m} \sum_i Z_i$, the Chernoff-Hoeffding bound [21] tells us that, for $t \geq 0$, $\mathbb{P}[S \geq (p+1)^{-1} + t] \leq e^{-2t^2m}$. Therefore,

$$\mathbb{P}[\|\xi\|_p^p \geq \frac{\alpha^p}{2^p(p+1)}m + \frac{\alpha^p}{2^p}tm] \leq e^{-2t^2m},$$

which gives, for $t = \kappa m^{-\frac{1}{2}}$,

$$\mathbb{P}[\|\xi\|_p^p \geq \zeta_p + \frac{\alpha^p}{2^p}\kappa m^{\frac{1}{2}}] \leq e^{-2\kappa^2}.$$

The limit value of $(\zeta_p + \frac{\alpha^p}{2^p}\kappa m^{\frac{1}{2}})^{1/p}$ when $p \rightarrow \infty$ is left to the reader. ■

REFERENCES

- [1] L. Jacques, D. K. Hammond, and M. J. Fadili, “Dequantizing Compressed Sensing with Non-Gaussian Constraints,” in *Proc. of IEEE International Conference on Image Processing (ICIP)*, Cairo, Egypt, Nov. 2009.
- [2] E. Candès and J. Romberg, “Quantitative Robust Uncertainty Principles and Optimally Sparse Decompositions,” *Foundations of Computational Mathematics*, vol. 6, no. 2, pp. 227–254, 2006.
- [3] D. Donoho, “Compressed Sensing,” *Information Theory, IEEE Transactions on*, vol. 52, no. 4, pp. 1289–1306, 2006.
- [4] E. J. Candès and T. Tao, “Near-Optimal Signal Recovery from Random Projections: Universal Encoding Strategies,” *IEEE Trans. Inform. Theory*, vol. 52, pp. 5406–5425, 2004.
- [5] P. Boufounos and R. G. Baraniuk, “1-bit Compressive Sensing,” in *42nd annual Conference on Information Sciences and Systems (CISS)*, Princeton, NJ, March 2008, pp. 19–21.
- [6] W. Dai, H. V. Pham, and O. Milenkovic, “Distortion-Rate Functions for Quantized Compressive Sensing,” *submitted to IEEE Information Theory Workshop (ITW) and to IEEE International Symposium on Information Theory (ISIT)*, 2009, arXiv:0901.0749.
- [7] A. Zymnis, S. Boyd, and E. Candès, “Compressed Sensing with Quantized Measurements,” 2009, http://stanford.edu/~boyd/papers/quant_compr_sens.html.
- [8] N. Thao and M. Vetterli, “Deterministic Analysis of Oversampled A/D Conversion and Decoding Improvement Based on Consistent Estimates,” *Signal Processing, IEEE Transactions on [see also Acoustics, Speech, and Signal Processing, IEEE Transactions on]*, vol. 42, no. 3, pp. 519–531, 1994.
- [9] P. Weiss, G. Aubert, and L. Blanc-Féraud, “Some Applications of ℓ_∞ -Constraints in Image Processing,” *INRIA Research Report*, vol. 6115, 2006.

- [10] P. Weiss, L. Blanc-Feraud, T. Andre, and M. Antonini, "Compression Artifacts Reduction Using Variational Methods: Algorithms and Experimental Study," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 2008, pp. 1173–1176.
- [11] V. Goyal, A. Fletcher, and S. Rangan, "Compressive Sampling and Lossy Compression," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 48–56, 2008.
- [12] E. Candès and J. Romberg, "Encoding the ℓ_p Ball From Limited Measurements," in *Data Compression Conference, 2006, March 2006, Snowbird, UT*.
- [13] P. Boufounos and R. G. Baraniuk, "Quantization of Sparse Representations," in *Data Compression Conference, 2007, March 2007, Snowbird, UT*.
- [14] H. Rauhut, K. Schnass, and P. Vandergheynst, "Compressed Sensing and Redundant Dictionaries," *Information Theory, IEEE Transactions on*, vol. 54, no. 5, pp. 2210–2219, 2008.
- [15] L. Ying and Y. Zou, "Linear Transformations and Restricted Isometry Property," *preprint*, 2009, arXiv:0901.0541.
- [16] J. Laska, P. Boufounos, M. Davenport, and R. Baraniuk, "Democracy in Action: Quantization, Saturation, and Compressive Sensing," *preprint*, 2009.
- [17] E. Candès, J. Romberg, and T. Tao, "Stable Signal Recovery from Incomplete and Inaccurate Measurements," *Comm. Pure Appl. Math*, vol. 59, no. 8, pp. 1207–1223, 2006.
- [18] M. Fadili and J.-L. Starck, "Monotone Operator Splitting for Fast Sparse Solutions of Inverse Problems," *SIAM Journal on Imaging Sciences*, 2009, submitted.
- [19] P. Combettes, "Solving Monotone Inclusions via Compositions of Nonexpansive Averaged Operators," *Optimization*, vol. 53, no. 5, pp. 475–504, 2004.
- [20] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic Decomposition by Basis Pursuit," *SIAM Review*, vol. 43, no. 1, pp. 129–159, 2001. [Online]. Available: <http://link.aip.org/link/?SIR/43/129/1>
- [21] W. Hoeffding, "Probability Inequalities for Sums of Bounded Random Variables," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 13–30, 1963.
- [22] E. Candès, "The Restricted Isometry Property and its Implications for Compressed Sensing," *Compte Rendus de l'Academie des Sciences, Paris, Serie I*, vol. 346, pp. 589–592, 2008.
- [23] R. G. Baraniuk, M. A. Davenport, R. A. DeVore, and M. B. Wakin, "A Simple Proof of the Restricted Isometry Property for Random Matrices," *Constructive Approximation*, vol. 28, no. 3, pp. 253–263, December 2008.
- [24] D. Donoho and J. Tanner, "Counting Faces of Randomly-Projected Polytopes when the Projection Radically Lowers Dimension," *Journal of the AMS*, vol. 22, no. 1, pp. 1–15, January 2009.
- [25] S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann, "Reconstruction and Subgaussian Operators in Asymptotic Geometric Analysis," *Geometric and Functional Analysis*, vol. 17, no. 4, pp. 1248–1282, 2007.
- [26] 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 Conference*, 2008.
- [27] R. Berinde and P. Indyk, "Sparse Recovery Using Sparse Matrices," MIT, MA, USA, Tech. Rep. MIT-CSAIL-TR-2008-001, October 2008.
- [28] R. Chartrand and V. Staneva, "Restricted Isometry Properties and Nonconvex Compressive Sensing," *Inverse Problems*, vol. 24, p. 035020, 2008.

- [29] A. Cohen, R. DeVore, and W. Dahmen, “Compressed Sensing and Best k -Term Approximation,” *J. Amer. Math. Soc.*, vol. 22, pp. 211–231, 2009.
- [30] M. Varanasi and B. Aazhang, “Parametric Generalized Gaussian Density Estimation,” *The Journal of the Acoustical Society of America*, vol. 86, pp. 1404–1415, 1989.
- [31] S. Foucart and M.J. Lai, “Sparsest Solutions of Underdetermined Linear Systems via ℓ_q -Minimization for $0 < q \leq 1$,” *Applied and Computational Harmonic Analysis*, vol. 26, no. 3, pp. 395–407, 2009.
- [32] M. E. Davies and R. Gribonval, “Restricted Isometry Constants Where ℓ_p -Sparse Recovery Can Fail for $0 < p \leq 1$,” *Information Theory, IEEE Transactions on*, vol. 55, no. 5, pp. 2203–2214, May 2009.
- [33] J. Fuchs, “Fast Implementation of a ℓ_1 - ℓ_1 Regularized Sparse Representations Algorithm,” in *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE Computer Society, 2009, pp. 3329–3332.
- [34] D. François, V. Wertz, and M. Verleysen, “The Concentration of Fractional Distances,” *IEEE Trans. Know. Data. Eng.*, pp. 873–886, 2007.
- [35] W. Bynum, “Weak Parallelogram Laws for Banach Spaces,” *Canad. Math. Bull.*, vol. 19, no. 3, pp. 269–275, 1976.
- [36] H. K. Xu, “Inequalities in Banach Spaces with Applications,” *Nonlinear analysis*, vol. 16, no. 12, pp. 1127–1138, 1991.
- [37] R. Spira, “Calculation of the Gamma Function by Stirling’s Formula,” *Math. Comp.*, vol. 25, no. 114, pp. 317–322, 1971.
- [38] P. Combettes and J. Pesquet, “A Proximal Decomposition Method for Solving Convex Variational Inverse Problems,” *Inverse Problems*, vol. 24, p. 27, December 2008.
- [39] 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.
- [40] P. Combettes and V. Wajs, “Signal Recovery by Proximal Forward-Backward Splitting,” *Multiscale Modeling and Simulation*, vol. 4, no. 4, p. 1168, 2006.
- [41] A. N. A. Ben-Tal, *Optimization III: Convex Analysis, Nonlinear Programming Theory, Standard Nonlinear Programming Algorithms*. Lecture notes, 2004.
- [42] D. Zwillinger, *CRC Standard Mathematical Tables and Formulae*. Chapman & Hall/CRC, 2003.
- [43] M. Lustig, D. Donoho, and J. Pauly, “Sparse MRI: The Application of Compressed Sensing for Rapid MR Imaging,” *Magnetic Resonance in Medicine*, vol. 58, no. 6, p. 1182, 2007.
- [44] L. Rudin, S. Osher, and E. Fatemi, “Nonlinear Total Variation Based Noise Removal Algorithms,” *Physica D*, Jan 1992.
- [45] C. R. Vogel and M. E. Oman, “Iterative Methods for Total Variation Denoising,” *SIAM J. Sci. Comput.*, vol. 17, no. 1, pp. 227–238, 1996.
- [46] T. F. Chan, G. H. Golub, and P. Mulet, “A Nonlinear Primal-Dual Method for Total Variation-Based Image Restoration,” *SIAM J. Sci. Comput.*, vol. 20, no. 6, pp. 1964–1977, 1999.
- [47] A. Chambolle, “An Algorithm for Total Variation Minimization and Applications,” *Journal of Mathematical Imaging and Vision*, Jan 2004.
- [48] Y. Wang, J. Yang, W. Yin, and Y. Zhang, “A New Alternating Minimization Algorithm for Total Variation Image Reconstruction,” *SIAM Journal on Imaging Sciences*, vol. 1, no. 3, pp. 248–272, 2008.
- [49] M. Ledoux and M. Talagrand, *Probability in Banach Spaces: Isoperimetry and Processes*. Springer, 1991.
- [50] M. Ledoux, “The Concentration of Measure Phenomenon,” *American Mathematical Society, Providence, RI*, 2001.



Laurent Jacques received the B.Sc. in Physics, the M.Sc. in Mathematical Physics and the PhD in Mathematical Physics from the Université catholique de Louvain (UCL), Belgium. He was a Postdoctoral Researcher with the Communications and Remote Sensing Laboratory of UCL in 2005-2006. He obtained in Oct. 2006 a four-year (3+1) Postdoctoral funding from the Belgian FRS-FNRS in the same lab. He was a visiting Postdoctoral Researcher, in spring 2007, at Rice University (DSP/ECE, Houston, TX, USA), and from 2007 to 2009, at the Swiss Federal Institute of Technology (LTS2/EPFL, Switzerland). His research focuses on Sparse Representations of signals (1-D, 2-D, sphere), Compressed Sensing, Inverse Problems, and Computer Vision.



David K. Hammond was born in Loma Linda, California. He received a B.S. with honors in Mathematics and Chemistry from the Caltech in 1999, then served as a Peace Corps volunteer teaching secondary mathematics in Malawi from 1999-2001. In 2001 he began studying at the Courant Institute of Mathematical Sciences at New York University, receiving a PhD in Mathematics in 2007. From 2007 to 2009, he was a postdoctoral researcher at the Ecole Polytechnique Federale de Lausanne. Since 2009, he is postdoc at the NeuroInformatics Center at the University of Oregon, USA. His research interests focus on image processing and statistical signal models, data processing on graph, as well as inverse problems related to EEG source localization for neuroimaging.



Jalal M. Fadili graduated from the Ecole Nationale Supérieure d'Ingénieurs (ENSI) de Caen, Caen, France, and received the M.Sc. and Ph.D. degrees in signal and image processing from the University of Caen. He was a Research Associate with the University of Cambridge (MacDonnel-Pew Fellow), Cambridge, U.K., from 1999 to 2000. He has been an Associate Professor of signal and image processing since September 2001 at ENSI. He was a visitor at several universities (QUT-Australia, Stanford University, CalTech, EPFL). His research interests include statistical approaches in signal and image processing, inverse problems, computational harmonic analysis, optimization and sparse representations. His areas of application include medical and astronomical imaging.