View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Infoscience - École polytechnique fédérale de Lausanne

SCHOOL OF ENGINEERING - STI
ELECTRICAL ENGINEERING INSTITUTE
SIGNAL PROCESSING LABORATORY

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

*Dr Laurent Jacques*

EPFL - FSTI - IEL - LTS
Station 11
Switzerland-1015 LAUSANNE

*Phone:* *+4121 693 26 57*
*Fax:* *+4121 6937600*
*e-mail:* `firstname.lastname@epfl.ch`

# Dequantizing Compressed Sensing: When Oversampling and Non-Gaussian Constraints Combine

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

Ecole Polytechnique Fédérale de Lausanne (EPFL)
Signal Processing Laboratory

Technical Report LTS-2009-002

February 18, 2009

Part of this work is also available on http://arxiv.org/abs/0902.2367
This work has been supported by Belgian National Science Foundation (F.R.S.-FNRS).

# Dequantizing Compressed Sensing:

## When Oversampling and Non-Gaussian Constraints Combine

L. Jacques[1,2], D. Hammond[1], M. J. Fadili[3]

[1]Institute of Electrical Engineering, Ecole Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland

[2]Communications and Remote Sensing Laboratory, Université catholique de Louvain (UCL), B-1348 Louvain-la-Neuve, Belgium

[3]GREYC CNRS-ENSICAEN-Université de Caen, 14050 Caen France

April 3, 2009

**Abstract**

The theory of Compressed Sensing (CS) is based on reconstructing sparse signals from random linear measurements. 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.

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 $p$ ($\mathrm{BPDQ}_p$), 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 particular data-fidelity constraint imposing that the difference between the original and the reproduced measurements has bounded $\ell_p$ norm, for $2 \le p \le \infty$.

We show that, in an oversampled situation, i.e. when the ratio between the number of measurements and the sparsity of the signal becomes large, the performance of the $\mathrm{BPDQ}_p$ decoders are significantly better than that of BPDN. Indeed, in this case the reconstruction error due to quantization is divided by $\sqrt{p+1}$. The condition guaranteeing this reduction relies on a modified Restricted Isometry Property ($\mathrm{RIP}_p$) of the sensing matrix bounding the projections of sparse signals in the $\ell_p$ norm. Surprisingly, Gaussian random matrices are also $\mathrm{RIP}_p$ with high probability. To demonstrate the theoretical power of $\mathrm{BPDQ}_p$, we report numerical simulations on signal and image reconstruction problems.

## Contents

# 1 Introduction

The theory of Compressed Sensing (CS) [CR06, Don06] is based on reconstructing sparse or compressible signals from few linear measurements compared to the dimensionality of the signal space.

In short, on one side, a signal is compressed at the acquisition level from few linear and non-adaptive measures. This generalizes the common (Dirac) signal sampling, i.e. the signal is sensed by its comparison (correlation) with the elements of a certain sensing basis.

On the other side, we have a “beyond Nyquist” statement. It tells us that signal reconstruction is still possible thanks to the signal sparsity or compressibility prior. Indeed, if the matrix underlying the sensing stage 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 reconstruction is guaranteed from non-linear techniques based on convex optimization. What makes CS more than an interesting theoretical concept, is that some class of randomly generated matrices (e.g. Gaussian, Bernoulli, Fourier ensemble, ...) respect the RIP with overwhelming probability. Roughly speaking, the common requirement is then that the number of measurements be higher than a few multiples of the signal sparsity. The same RIP also ensures that the reconstruction is robust if noise corrupt the measurements and/or if the signal deviates from the exactly sparse model, i.e. in the compressible setup.

In a realistic acquisition system, quantization of these measurements is a natural process that Compressed Sensing theory has to handle conveniently. Any coder or device that can integrate a CS encoding step may transmit data to the decoder in a digital way so as for instance to limit the transmission error. This will require modification to the algorithm used to recover the signal. The commonly used Basis Pursuit algorithm for CS recovery finds the sparsest signal (in $\ell_1$ norm) that could have produced the observed measurements. Directly using the quantized measurements in Basis Pursuit fails, however, as there may be no signal close to the desired signal whose (unquantized) measurements reproduce the observed quantized values! This problem may be resolved by relaxing the data fidelity constraint.

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. While this approach can give acceptable results, it is theoretically unsatisfying as the measurement error created by quantization is highly non-Gaussian, being essentially uniform and bounded by the quantization bin width.

A more appealing modification is to impose the Quantization Consistency (QC) constraint, i.e. the fact that the requantized measurements of the reconstructed signal must match the original quantized measurements. This idea, in some form, has appeared previously in the literature. At the beginning of the Compressed Sensing developments, Candès et al. mentioned that the $\ell_2$-norm of BPDN should be replaced by the $\ell_\infty$-norm to handle more naturally the quantization distortion of the measurements [CT04]. More recently, In [PG08], 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 [DPM09], an adaptation of both BPDN and the Subspace Pursuit integrates an explicit QC contstraint. 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 compressed sensing field. Examples include oversampled Analog to Digital Converter (ADC) conversion of signal [TV94], and in image restoration problems [WAB06, WBA08].

In this paper, we propose to tackle the problem by the introduction of a new class of convex optimization programs, or decoders, coined Basis Pursuit DeQuantizer of moment $p$ ($\mathrm{BPDQ}_p$) that model the quantization distortion more faithfully. These proceed by minimizing the sparsity of the reconstructed signal subject to a particular data-fidelity constraint. This constraint imposes that the difference between the original and the reproduced measurements has bounded $\ell_p$ norm, for $p \geq 2$. As $p$ approaches infinity, this fidelity term reproduces the QC constraint.

We show theoretically that, given a sufficient number of measurements, the performance of the $\mathrm{BPDQ}_p$ decoders for appropriate $p$ are significantly better than that of BPDN, i.e. $\mathrm{BPDQ}_2$. This difference is especially significant in an oversampled situation, i.e. when the ratio between the number of measurements and the sparsity of the signal becomes large. In this case, we show that if the signal is exactly sparse, the approximation error scales inversely with $\sqrt{p+1}$.

The proof of these results relies on introducing a modified Restricted Isometry Property ($\mathrm{RIP}_p$) that bounds the $\ell_p$ norm of random projections of sparse signals. We show that matrices generated by a Gaussian random process satisfy this new property with controllable high probability. We conclude with numerical simulations demonstrating the use of the $\mathrm{BPDQ}_p$ for example signal and image reconstruction problems.

The paper is structured as follows. In Section 2, we review the principles of Compressed Sensing, i.e. the sensing model, the reconstruction and the requirements for its robustness with respect to noise and to deviation to the exactly sparse signal model. The common treatment of the quantization distortion as a noise is also recalled.

The Section 3 introduces the BPDQ decoders. Their stability, i.e. the $\ell_2-\ell_1$ instance optimality, is deduced by the new Restricted Isometry Property of moment $p$ ($\mathrm{RIP}_p$). Standard Gaussian Random matrices are shown to satisfy this property with high probability for a sufficiently large number of measurements. The section finishes with the key result of this paper, i.e. it is shown that the approximation error of $\mathrm{BPDQ}_p$ scales inversely with $\sqrt{p+1}$ when the signal measurements are uniformly quantized.

In Section 4, we briefly explore the non-convex world, i.e. when we deal with equivalent decoders based on the $\ell_0$ non-convex sparsity constraint. Those can be solved approximately by a reweighting procedure of the $\ell_1$ BPDQ programs. The $\mathrm{RIP}_p$ is also essential to bound the approximation error but only for exactly $K$-sparse signal model. It is shown that the same approximation error reduction by $\sqrt{p+1}$ occurs at smaller oversampling ratio $m/K$.

Section 5 describes the numerical techniques that we emply to solve the BPDQ programs. These rely on the Proximal methods [Com04]. In particular, the global optimization is solved iteratively using the Douglash Rashford splitting method. The fidelity constraint expressed in the $\ell_p$-norm corresponds to an orthogonal projection on a $\ell_p$ tube of $\mathbb{R}^N$ that we compute iteratively by using Newton's method to solve the associated Lagrange multiplier problem. Section 7 confirms experimentally the theoretical power of $\mathrm{BPDQ}_p$ on 1-D signals and images reconstruction problems.

## 2 Compressed Sensing and Quantization of Measurements

In Compressed Sensing (CS) theory [CR06, Don06], the signal $x \in \mathbb{R}^N$ to be acquired and subsequently reconstructed is assumed sparse or compressible in an orthogonal[1] 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 $\Psi$ gives an exact or accurate representation of this signal even for small $K < N$. For simplicity, only the canonical basis $\Psi = \mathrm{Id}$ will be considered here.

At the acquisition stage, $x$ is measured by $m$ linear measurements (with $K \le m \le N$) provided by a sensing matrix $\Phi \in \mathbb{R}^{m\times N}$, i.e. we know from $x$ only $m$ measurements (or *questions*) $\langle \varphi_i, x\rangle = \sum_k \varphi^*_{ik} x_k$ where $(\varphi_i)_{i=0}^{m-1}$ are the rows of $\Phi$.

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:

$$y_{\mathrm{q}} \;=\; Q_\alpha[\Phi x] \;=\; \Phi x + n, \tag{1}$$

where $y_{\mathrm{q}} \in (\alpha\mathbb{N} + \frac{\alpha}{2})^m$ 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 $\alpha$, and $n \in \mathbb{R}^m$ is the *quantization distortion*.

This model 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 compressed sensing theory. The study of more realistic non-uniform quantization is deferred as a question for future research.

In much previous work in compressed sensing, the reconstruction of $x$ from $y_{\mathrm{q}}$ is obtained by handling the quantization distortion $n$ as a noise of bounded power (i.e. $\ell_2$-norm) $\|n\|_2^2 = \sum_k |n_k|^2$. For such a noise, 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) [CRT06]:

$$\Delta(y,\epsilon) \;=\; \operatorname*{arg\,min}_{u\in\mathbb{R}^N} \|u\|_1 \;\text{ s.t. }\; \|y - \Phi u\|_2 \le \epsilon, \tag{\textbf{BPDN}}$$

This convex optimization program can be solved numerically by methods like Second Order Cone Programing (for BPDN) or by the Proximal methods [Com04] described in Section 5. We will often refer to the constraint $\|y - \Phi u\|_2 \le \epsilon$ in BPDN as the *fidelity term*. Notice that the noiseless situation $\epsilon = 0$ leads to the Basis Pursuit (BP) program with additional specific numerical implementation in the field of Linear Programming and interior point method.

A necessary condition for BPDN to provide a good approximation of the initial signal $x$ is the *feasibility* of this solution, i.e. we must chose $\epsilon$ s.t. $\|n\|_2 = \|y - \Phi x\|_2 \le \epsilon$. In [CRT06], an estimator of $\epsilon$ for $y = y_{\mathrm{q}}$ is obtained by considering $n$ distributed as a uniform random vector $\xi \in \mathbb{R}^m$ in the quantization bins, i.e. $\xi_i \sim_{\mathrm{iid}} U([-\frac{\alpha}{2}, \frac{\alpha}{2}])$. An easy computation shows then that

$$\|\xi\|_2^2 \;\le\; \epsilon_2^2(\alpha) \;\triangleq\; E[\|\xi\|_2^2] + \kappa\sqrt{\mathrm{Var}[\|\xi\|_2^2]} \;=\; \tfrac{\alpha^2}{12} m + \kappa\,\tfrac{\alpha^2}{6\sqrt{5}}\, m^{\frac{1}{2}},$$

[1] A generalization for redundant basis, or dictionary, exists [RSV08, YZ09].

with a probability higher than $1 - e^{-c_0\kappa^2}$ for a certain constant $c_0 > 0$ (by the Chernoff-Hoeffding bound [Hoe63]). Therefore, CS usually handle quantization distortion by setting $\epsilon = \epsilon_2(\alpha)$, typically for $\kappa = 2$.

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

**Definition 1.** *A matrix $\Phi \in \mathbb{R}^{m\times N}$ satisfies the Restricted Isometry Property (RIP) of order $K$ and radius $\delta \in (0,1)$, if there exists a constant $\mu$ such that*

$$\mu\sqrt{1-\delta}\,\|u\|_2 \;\leq\; \|\Phi u\|_2 \;\leq\; \mu\sqrt{1+\delta}\,\|u\|_2, \tag{2}$$

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

To clarify the writting of our study, we adopt here a slightly different definition of the common RIP [Can08]. The original definition considers indeed normalized matrices $\bar\Phi = \Phi/\mu$ having unit-norm columns (in expectation) so that $\mu$ is absorbed in the normalizing constant. We will see in Section 3.2 that the value of $\mu$ is intrinsically linked to the norm used to measure the projected vector $\Phi u$ in (2).

Fortunately, it is actually not too difficult to find matrices satisfying the RIP. For instance, a matrix $\Phi \in \mathbb{R}^{m\times N}$ with each of its entries drawn independtly from a (sub) Gaussian random variable satisfies this property with an overwhelming probability as soon as [BDD08, DT09]

$$m \;\geq\; O(K\;\log N/K).$$

This is the case of Standard Gaussian Random (SGR) matrices where $\Phi_{ij} \sim N(0,1)$, and of the Bernoulli matrices with $\Phi_{ij} = \pm 1$ with equal probability, with for both $\mu = \sqrt{m}$ in the RIP definition. This results from the measure concentration property of such random distribution over the union of all the $K$-dimensional subspaces of $\mathbb{R}^N$ [LT91]. We will reprove this basic fact as a special case of a more general RIP adapted to other norms of the projection $\Phi u$ of sparse vectors $u$. Let us mention that other random constructions satisfying the RIP exist (e.g. Fourier ensemble) [CR06, CRT06].

The following theorem expresses the announced stability result, i.e. the $\ell_2 - \ell_1$ instance optimality[2] of BPDN.

**Theorem 1** ([Can08])**.** *Let $x \in \mathbb{R}^N$ be a compressible signal with a $K$-term $\ell_1$-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 $\ell_2$-approximation of $x$. Let $\Phi$ be a RIP 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 the $\ell_2 - \ell_1$ instance optimality*

$$\|x^* - x\|_2 \;\leq\; A\,e_0(K) \;+\; B\,\tfrac{\epsilon}{\mu}, \tag{3}$$

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

[2]Adopting the definition of mixed-norm instance optimality [CDD09].

We call the first term in the RHS of (3), the *compressibility error*, while the second term is named the *noise error*.

*Remark*: The situation where $x$ represents the vector of coefficients of a signal $s = \Psi x$ in a non-trivial sparsity basis $\Psi$ instead of the signal itself can be solved by considering the new sensing basis $\Theta = \Phi\Psi$. This one is still RIP with overwhelming probability when $\Phi$ is a SGR matrix and when $\Psi$ is an orthonormal basis[3]. In that case however, the $\ell_2$ approximation error on the left of (3) is equal to the approximation error of the signal (by Parseval), while the $\ell_1$-measure of the signal compressibility on the right is still expressed in the coefficient domain.

However, 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. $Q_\alpha[\Phi x^*] = y_{\rm q}$. This will be met iff

$$\|y_{\rm q} - \Phi x^*\|_\infty \leq \tfrac{\alpha}{2}$$

which is not necessarily implied by the BPDN $\ell_2$ 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 signal encoded in the available measurements.

Second, from a Bayesian Maximum a Posteriori (MAP) standpoint, BPDN can be viewed as solving an ill-posed inverse problem where the $\ell_2$-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.

# 3 A New Class of Decoders: BPDQ$_p$

## 3.1 Generalizing the Fidelity Constraint

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.

In the objective of reconstructing an approximation of a sparse or compressible signal $x$ from its measurements $y = \Phi x + n$ when the $p^{\rm th}$ moment of the noise $n$ is bounded, i.e. when $\|n\|_p^p = \sum_k |n_k|^p \leq \epsilon$ for some estimators $\epsilon > 0$, we introduce the novel programs

$$\Delta_p(y,\epsilon) \;=\; \operatorname*{arg\,min}_{u\in\mathbb{R}^N} \|u\|_1 \;\text{ s.t. }\; \|y - \Phi u\|_p \;\leq \epsilon, \quad \text{for } p \geq 1. \qquad (\mathbf{BPDQ}_p)$$

The fidelity constraint expressed in the $\ell_p$-norm is now tuned to noises that follow centered Generalized Gaussian Distribution[4] (GGD) of *shape parameter* $p$ [VA89], with the uniform noise case corresponding to $p \to \infty$.

We dub this class of decoders *Basis Pursuit DeQuantizer* of *moment* $p$ (or BPDQ$_p$) since, for reasons that will become clear in the next sections, their approximation error when $\Phi x$ is uniformly quantized has an interesting decreasing behavior when both the moment $p$ and the oversampling factor $m/K$ increase.

[3] Results are also proved for the case where $\Psi$ is a redundant basis or a dictionary [RSV08, YZ09]. In that case the RIP radius $\delta$ is increased by a quantity linked to the coherence of the basis.

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

This arises from the conjunction of two effects. First, when the matrix $\Phi$ satisfies a certain extension of the RIP, the BPDQ$_p$ decoders satisfies an instance optimality, similar to this of BPDN (Section 2), with respect to $\epsilon$ and to the possible compressibility of $x$. Second, the $p^{\rm th}$ moment of a uniformly random vector arising from quantization distortion can be precisely bounded for all $p$.

### 3.2 RIP$_p$ and $\ell_2 - \ell_1$ Instance Optimality

In order to study the approximation error of the BPDQ$_p$ decoders, we must introduce the Restricted Isometry Property of moment $p$ (or RIP$_p$).

**Definition 2.** *A matrix $\Phi \in \mathbb{R}^{m\times N}$ satisfies the RIP$_p$ ($1 \leq p \leq \infty$) property of order $K$ and radius $\delta$, if there exists a constant $\mu_p > 0$ such that*

$$\mu_p \sqrt{1-\delta}\, \|x\|_2 \;\leq\; \|\Phi x\|_p \;\leq\; \mu_p \sqrt{1+\delta}\, \|x\|_2, \quad \forall\, x \in \mathbb{R}^N \;\; \text{s.t.} \;\; \|x\|_0 \leq K, \tag{4}$$

*where $\|\cdot\|_p$ is the $\ell_p$ norm on $\mathbb{R}^m$.*

The common RIP previously introduced (see Section 2) is thus the RIP$_2$. Note that the constant $\mu_p$ appearing above will play an important role in our theory, and as we will see later is intimately related to the error of the recovered signal. Interestingly, as for the RIP, SGR matrices $\Phi \in \mathbb{R}^{m\times N}$ satisfy also the RIP$_p$ with high probability provided that $m$ is sufficiently big compared to the sparsity $K$ of the signals to measure. This is shown in the following Proposition, for which the proof is given in Appendix A.

**Proposition 1.** *Let $\Phi \in \mathbb{R}^{m\times N}$ be a Standard Gaussian Random (SGR) matrix, i.e. a matrix with each entry $\Phi_{ij}$ drawn according to a normalized Gaussian distribution $N(0,1)$. Then, if $m \geq (p-1)2^{p+1}$ and $m \geq O\big((\delta^{-2}\, K \log N/K)^{\max(p/2,1)}\big)$ for $1 \leq p < \infty$, or $\log m \geq O(\delta^{-2}\, K \log N/K)$ for $p = \infty$, $\Phi$ is RIP$_p$ of order $K$ and radius $\delta$ with a high probability. Moreover, the value $\mu_p = E[\|\xi\|_p]$ is the expectation value of the $\ell_p$-norm of a SGR vector $\xi \in \mathbb{R}^m$.*

Notice that $\mu_p$ can be approximated thanks to the following result (see Appendix B for its explanation) that is mainly due to the study made in [FWV07].

**Lemma 1.** *For the random vector $\xi \in \mathbb{R}^m$ with $\xi_i \sim_{\rm iid} \mathcal{P}$, for a certain distribution $\mathcal{P}$, there exists a $\rho > 0$ such that $\rho^{-1} \sqrt{\log m} \leq E[\|\xi\|_\infty] \leq \rho \sqrt{\log m}$. In addition, for $1 \leq p < \infty$, if $\mathcal{P}$ is a normalized Gaussian distribution $N(0,1)$,*

$$\begin{aligned} E[\|\xi\|_p] &\;\simeq\; E[\|\xi\|_p^p]^{\frac{1}{p}} \;=\; \sqrt{2}\, \pi^{-\frac{1}{2p}}\, \Gamma(\tfrac{p+1}{2})^{\frac{1}{p}}\; m^{\frac{1}{p}}, \\ \mathrm{Var}[\|\xi\|_p] &\;\simeq\; 2\, \pi^{-\frac{1}{p}} p^{-2}\; \frac{\sqrt{\pi}\,\Gamma(\frac{2p+1}{2}) \;-\; \Gamma(\frac{p+1}{2})^2}{\Gamma(\frac{p+1}{2})^{2-\frac{2}{p}}}\; m^{\frac{2}{p}-1}, \end{aligned} \tag{5}$$

*where the symbol $\simeq$ means that the equalities holds asymptotically with $m$.*

However, non-asymptotic bounds for $\mu_p$ can be deduced from $E[\|\xi\|_p^p]^{\frac{1}{p}}$ thanks to the following result (proof in Appendix C).

**Lemma 2.** *If $\xi \in \mathbb{R}^m$ is a SGR vector, then, for $1 \leq p < \infty$,*

$$\big(1 \;+\; \tfrac{2^{p+1}}{m}\big)^{\frac{1}{p}-1}\, E[\|\xi\|_p^p]^{\frac{1}{p}} \;\leq\; E[\|\xi\|_p] \;\leq\; E[\|\xi\|_p^p]^{\frac{1}{p}}.$$

*In particular, as soon as* $m \geq \beta^{-1}\, 2^{p+1}$ *for* $\beta \geq 0$,

$$E[\|\xi\|_p] \;\geq\; E[\|\xi\|_p^p]^{\frac{1}{p}}\,(1+\beta)^{\frac{1}{p}-1} \;\geq\; E[\|\xi\|_p^p]^{\frac{1}{p}}\,(1-\tfrac{p-1}{p}\beta).$$

In Figure 1, we have tested numerically the evolution of the ratio $\mu_p/E[\|\xi\|_p^p]^{1/p} \leq 1$ for $m \in \{128, 256, 512\}$ and $p \in [2, 16]$. The value of $\mu_p$ has been estimated by averaging of $\|\xi\|_p$ over 10 000 trials of SGR vectors $\xi \in \mathbb{R}^m$, while $E[\|\xi\|_p^p]^{1/p}$ is given by (5). We observe clearly that increasing $m$ makes the ratio closer to 1 for a larger range of $p$. However, $\mu_p/E[\|\xi\|_p^p]^{1/p}$ deviates slower from 1 when $p$ increases than the expected bound $(1 + m^{-1}2^{p+1})^{\frac{1}{p}-1}$ (not shown here).

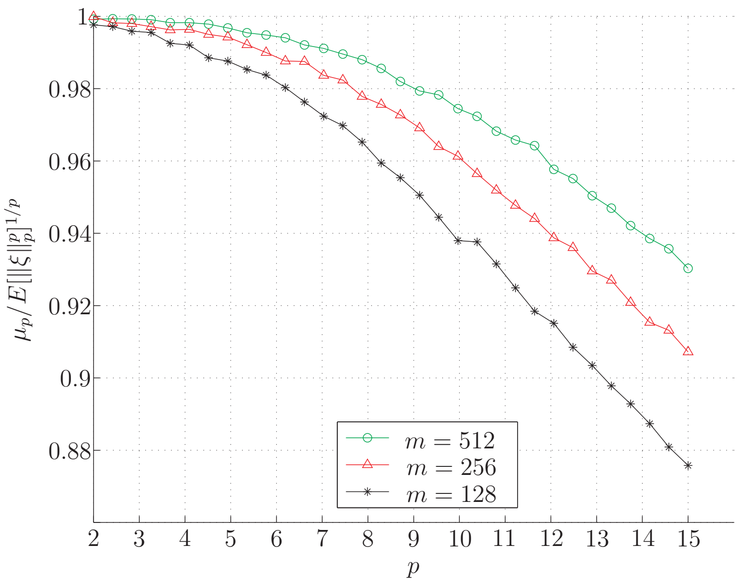



Figure 1: Numerical estimation of $\mu_p/E[\|\xi\|_p^p]^{1/p}$ (averaged over 10 000 trials) for $m \in \{128, 256, 512\}$ and $p \in [2, 16]$.

An interesting aspect of matrices respecting the RIP$_p$ is that they approximately preserve the decorrelation of sparse vectors of disjoint supports.

**Lemma 3.** *Let* $u, v \in \mathbb{R}^N$ *with* $\|u\|_0 = s$ *and* $\|v\|_0 = s'$ *and* $\operatorname{supp}(u) \cap \operatorname{supp}(v) = \emptyset$, *and* $2 \leq p < \infty$. *If* $\Phi$ *is RIP*$_p$ *of order* $s + s'$ *with constant* $\delta_{s+s'}$, *and of orders* $s$ *and* $s'$ *with constants* $\delta_s$ *and* $\delta_{s'}$, *then*

$$|\langle J(\Phi u), \Phi v\rangle| \;\leq\; \mu_p^2\, C_p\, \|u\|_2 \|v\|_2, \tag{6}$$

*with* $(J(u))_i = \|u\|_p^{2-p}\, |u_i|^{p-1} \operatorname{sign} u_i$ *and*

$$\begin{aligned} C_p(\delta_s, \delta_{s'}, \delta_{s'+s}) \;&=\; \min\Big\{ \;\big[\,(\delta_s + \delta_{s+s'})\,\big((p-2) + (p-1)\delta_{s'} + \delta_{s+s'}\big)\big]^{\frac{1}{2}}, \\ &\qquad \tfrac{1}{2}\,\big[\,(p-2+p\,\delta_{s+s'})\,\big(2(p-2) + (p-2)\delta_{s'} + p\,\delta_{s+s'}\big)\big]^{\frac{1}{2}}\;\Big\}. \end{aligned}$$

As shown in its proof in Appendix D, this Lemma uses explicitly the 2-smoothness of the Banach spaces $\ell_p$ when $p \geq 2$ [Byn76, Xu91], in connection with the *normalized duality mapping* $J$ that plays a central role in the geometrical description of $\ell_p$.

The relation (6) may be seen as a generalization of the one proved in [Can08] (see Lemma 2.1) for $p = 2$. We can check indeed that for this Hilbert space case $C_2 = \delta_{s+s'}$. We may finally notice that $C_p$ behaves as $\sqrt{(\delta_s + \delta_{s+s'})\,(1+\delta_{s'})\,p}$ for large $p$, and as $\delta_{s+s'} + \frac{3}{4}(1+\delta_{s+s'})(p-2)$ for $p \simeq 2$.

We are now ready to bound the approximation error commited by $\text{BPDN}_p$. This result, a generalization to Banach spaces $\ell_p$ of the fundamental result proved by Candès in [Can08], expresses the $\ell_2 - \ell_1$ instance optimality of the $\text{BPDQ}_p$ decoders, i.e. their stability under reconstruction of compressible (or sparse) signals under measurement corrupted by bounded $p^{\text{th}}$ moment.

**Theorem 2.** *Let $x \in \mathbb{R}^N$ be a compressible signal with a $K$-term $\ell_1$-approximation error $e_0(K) = K^{-\frac{1}{2}}\,\|x - x_K\|_1$, for $0 \le K \le N$ and $x_K$ the best $K$-term $\ell_2$-approximation of $x$. Let $\Phi$ be a $\text{RIP}_p$ matrix on $s$ sparse signals with constants $\delta_s$, for $s \in \{K, 2K, 3K\}$ and $2 \le p < \infty$. Given a measurement vector $y = \Phi x + n$ corrupted by a noise $n$ with bounded $p^{\text{th}}$ moment, i.e. $\|n\|_p \le \epsilon$, the solution $x^*_p = \Delta_p(y,\epsilon)$ obeys the $\ell_2 - \ell_1$ instance optimality*

$$\|x^*_p - x\|_2 \;\le\; A_p\, e_0(K) \;+\; B_p\,\frac{\epsilon}{\mu_p},$$

*for values*

$$A_p(\delta_K, \delta_{2K}, \delta_{3K}) \;=\; \tfrac{2(1+C_p-\delta_{2K})}{1-\delta_{2K}-C_p}, \qquad B_p(\delta_K, \delta_{2K}, \delta_{3K}) \;=\; \tfrac{4\sqrt{1+\delta_{2K}}}{1-\delta_{2K}-C_p},$$

*and $C_p = C_p(\delta_K, \delta_{2K}, \delta_{3K})$ given in Lemma 3.*

Notice that this theorem is unfortunately not valid for the case $p = \infty$. This is mainly due to the fact that $\ell_\infty$ does not fit the Lemma 3 and the proof of this theorem, i.e. this Banach space is not 2-smooth and no duality mapping exists. Therefore, a result involving $\ell_\infty$ would require different tools than the ones we developped here.

The proof of this theorem, given in Appendix E, adapts the one of [Can08] to the particular geometry of $\ell_p$.

### 3.3 Quantization Error Reduction

Let us now observe the particular behavior of the $\text{BPDN}_p$ decoders on quantized measurements of a sparse or compressible signal.

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 $\epsilon$ for all the $p^{\text{th}}$ moments of $n$.

**Lemma 4.** *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 $1 \le p < \infty$,*

$$\zeta_p = E[\|\xi\|_p^p] \;=\; \tfrac{\alpha^p}{2^p(p+1)}\, m. \tag{7}$$

*In addition, for any $\kappa > 0$, $P\big[\|\xi\|_p^p \ge \zeta_p + \kappa\,\frac{\alpha^p}{2^p}\sqrt{m}\,\big] \;\le\; e^{-2\kappa^2}$, while, $\lim_{p\to\infty}(\zeta_p \,+\, \kappa\,\frac{\alpha^p}{2^p}\sqrt{m}\,)^{\frac{1}{p}} \;=\; \|\xi\|_\infty = \frac{\alpha}{2}$.*

The proof is given in Appendix F.

According to this result, we may set the $p^{\text{th}}$ *moment* bound $\epsilon$ of the program $\text{BPDQ}_p$ to

$$\epsilon \;=\; \epsilon_p(\alpha) \;\triangleq\; \tfrac{\alpha}{2\,(p+1)^{1/p}}\,\big(\,m + \kappa\,(p+1)\sqrt{m}\,\big)^{\frac{1}{p}}, \tag{8}$$

so that, for $\kappa = 2$, we know that $x$ is a solution of the BPDQ$_p$ fidelity constraint with a probability $e^{-8} \leq 3.4\,10^{-4}$.

Second, Theorem 2 points out that, when $\Phi$ is RIP$_p$ 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 $\epsilon/\mu_p$. In particular, from Lemma 1, taking a SGR matrix, for $m \geq O\big((\delta^{-2}\,K \log N/K)^{p/2}\big)$, we have

$$\|x - x_p^*\|_2 \;\leq\; A_p\, e_0(K) \;+\; B_p \frac{\epsilon_p(\alpha)}{\mu_p}. \tag{9}$$

Therefore, combining these two results, for a uniform quantization distortion $n$, the noise error can be more precisely bounded. Indeed, for $2 \leq p < \infty$, by Lemma 2, if $m \geq (p-1)2^{p+1}$, $\mu_p \geq \frac{p-1}{p}\,\nu_p m^{\frac{1}{p}}$ with $\nu_p = \sqrt{2}\,\pi^{-\frac{1}{2p}}\,\Gamma(\frac{p+1}{2})^{\frac{1}{p}}$. In addition, using Stirling formula of the Gamma function, i.e.[5] $\Gamma(z) = (\frac{2\pi}{z})^{\frac{1}{2}}(\frac{z}{e})^z(1 + O(\frac{1}{z}))$, we get easily

$$\nu_p = 2^{-\frac{1}{p}}\, e^{-\frac{p+1}{2p}}\,\sqrt{p+1}\,\big(1 + O(p^{-2})\big) \;\geq\; 2^{-\frac{1}{2}}\, e^{-\frac{3}{4}}\,\sqrt{p+1}\,\big(1 + O(p^{-2})\big).$$

Finally, by (8), we see that,

$$\frac{\epsilon_p(\alpha)}{\mu_p} \;\leq\; \frac{e^{3/4}p}{\sqrt{2}(p-1)}\,\big(\tfrac{1}{p+1} + \kappa\,\tfrac{1}{\sqrt{m}}\big)^{1/p}\;\frac{\alpha}{\sqrt{p+1}}\,\big(1 + O(p^{-2})\big) <\; C\frac{\alpha}{\sqrt{p+1}}\,\big(1 + O(p^{-2})\big), \tag{10}$$

with $C < 2.994$. We used the bound $\frac{p}{p-1} \leq 2$ and the fact that $(\frac{1}{p+1} + \kappa\,\frac{1}{\sqrt{m}})^{1/p}$ tends to 1 from below[6] when $p \to \infty$.

*In short, the noise error term in the $\ell_2 - \ell_1$ instance optimality relation (9) in the case of uniform quantization of the measurements of a sparse or compressible signal is thus divided by $\sqrt{p+1}$ if the sensing matrix $\Phi$ satisfies the RIP$_p$!*

More precisely, with a philosophy close to the oversampled ADC conversion [TV94], *this error noise reduction happens in oversampled sensing*, i.e. when the *oversampling factor $m/K$* is high.

Indeed, taking a SGR matrix, Proposition 1 teaches us the following lesson. If $m_p$ is the smallest number of measurements for which such a randomly generated matrix $\Phi$ is RIP$_p$ 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 $\Phi$ that is RIP$_p$ for high $p$, providing the desired error division by $\sqrt{p+1}$. Second, oversampling gives a smaller $\delta$, i.e. $\delta \propto m^{-1/p}$, 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 $\delta$ also favors BPDN, but since the value $A = A_2$ and $B = B_2$ in (3) are bounded from below this effect is limited.

From this result, it is very tempting to chose an extremely large value for $p$ in order to decrease the noise error term (9). There are two problems with this. First, the instance optimality result is not directly valid for $p = \infty$. Second, and more significantly, the necessity of satisfying RIP$_p$

[5] A more precise formula can be established from [Spi71], i.e. $|\Gamma(x) - (\frac{2\pi}{x})^{\frac{1}{2}}(\frac{x}{e})^x| \leq \frac{1}{9x}\,(\frac{2\pi}{x})^{\frac{1}{2}}(\frac{x}{e})^x$ for $x \geq 1$.

[6] As soon as $\kappa \leq \frac{2}{3}\sqrt{m}$, i.e. $m \geq 9$ if $\kappa = 2$, $\frac{1}{p+1} + \kappa\,\frac{1}{\sqrt{m}} \leq (\frac{1}{p+1} + \kappa\,\frac{1}{\sqrt{m}})^{1/p} \leq 1$ for $p \geq 2$.

implies that we cannot cannot take $p$ arbitrarily large. Indeed, for a given oversampling factor $m/K$, a SGR matrix $\Phi$ can be $\mathrm{RIP}_p$ 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 demomstrate this effect experimentally in Section 7.

We may notice finally that the error arising from the compressible status of the signal, i.e. the compressibility error, is not significally reduced by increasing $p$ when the number of measurement is large. This makes sense since the $\ell_p$ norm touches only the fidelity term of the decoders and we know that in the case where $\epsilon = 0$, the compressibility error remains in the decoder BP [Can08]. However, because of the embedding of the $\ell_p$-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.

# 4 Answers from the Non Convex World

In this Section we provide some stability results for theoretical decoders inspired by the $\mathrm{BPDQ}_p$ programs but minimizing the $\ell_0$ "norm". The framework is however less general than the one of Section 3 since the absence of the $\ell_1$ sparsity promoting measure forces us to consider the reconstruction of exactly sparse signals only. We gives also some practical elements to numerically solve these non-convex decoders. Pragmatically, this is actually the reason why we consider these programs as important to study here.

Let $x \in \mathbb{R}^N$ be a signal exactly sparse in the canonical basis $\Psi = \mathrm{Id}$, i.e. $\|x\|_0 \leq K$. This signal is assumed as before acquired by $m$ noisy measurements

$$y \;=\; \Phi x + n,$$

for a certain sensing matrix $\Phi \in \mathbb{R}^{m\times N}$ and a noise $n$ with power (or second moment) $\|n\|_2 \leq \epsilon$.

Theoretically, the decoder

$$\Delta^{\mathrm{th}}(y,\epsilon) \;=\; \operatorname*{arg\,min}_{u\in\mathbb{R}^N} \|u\|_0 \;\text{ s.t. }\; \|y - \Phi u\|_2 \;\leq \epsilon,$$

leads to a much smaller approximation error than the one provided by BPDN in Section 2 when $\Phi$ satisfies the RIP.

Indeed, writing $x^* = \Delta^{\mathrm{th}}(y,\epsilon)$,

$$\|x - x^*\|_2 \;\leq\; \tfrac{2}{\sqrt{1-\delta}}\,\epsilon.$$

This is easily obtained by observing that $x - x^*$ is $2K$-sparse since $\|x^*\|_0 \leq \|x\|_0$ because $x$ satisifies the fidelity constraint of the decoder. Therefore, if $\Phi$ is RIP of radius $0 < \delta < 1$, $\|x - x^*\|_2 \leq \frac{1}{\sqrt{1-\delta}}\,\|\Phi(x - x^*)\|_2 \leq \frac{1}{\sqrt{1-\delta}}\big(\|\Phi x - y\|_2 + \|y - \Phi x^*\|_2\big) \leq \frac{2}{\sqrt{1-\delta}}\,\epsilon$. For instance, for $\delta = 0.2$, $\frac{2}{\sqrt{1-\delta}} \simeq 2.24$, that is much lesser than the value 8.5 multiplying $\epsilon$ in Theorem 1!

When the noise $n$ has a bounded $p^{\mathrm{th}}$ moment, i.e. $\|n\|_p \leq \epsilon$, as ensured by Lemma 4 for $\epsilon = \epsilon_p(\alpha)$ in the quantization model

$$y \;=\; y_{\mathrm{q}} \;=\; Q_\alpha[\Phi x] \;=\; \Phi x + n,$$

the same stability can be proved for the decoders

$$\Delta_p^{\mathrm{th}}(y,\epsilon) \;=\; \operatorname*{arg\,min}_{u\in\mathbb{R}^N} \|u\|_0 \;\text{ s.t. }\; \|y - \Phi u\|_p \;\leq \epsilon.$$

The vector $x - x_p^*$, with $x_p^* = \Delta_p^{\rm th}(y, \epsilon)$, is $2K$-sparse[7]. Therefore, if $\Phi$ is $\mathrm{RIP}_p$ over $2K$ sparse signals with constant $\delta$ for $2 \le p < \infty$,

$$\begin{aligned}
\|x - x_p^*\|_2 &\le \tfrac{1}{\mu_p\sqrt{1-\delta}}\,\|\Phi(x - x_p^*)\|_p\\
&\le \tfrac{1}{\mu_p\sqrt{1-\delta}}\,\big(\|\Phi x - y\|_p \;+\; \|\Phi x_p^* - y\|_p\big)\\
&\le \tfrac{2}{\sqrt{1-\delta}}\,\frac{\epsilon}{\mu_p},
\end{aligned}$$

using the triangle inequality.

Therefore, the same ratio $\epsilon/\mu_p$ is reached in the error bound for $2 \le p < \infty$. When measurements are uniformly quantized by $Q_\alpha$, we can then apply the same arguments than in the end of Section 3.3 and observe that, as soon as $\Phi$ is $\mathrm{RIP}_p$,

$$\|x - x_p^*\|_2 \;\le\; \tfrac{2C}{\sqrt{1-\delta}}\,\frac{\alpha}{\sqrt{p+1}}.$$

Intrinsically, $\delta$ is of course also function of $p$ and $m$. Indeed, Proposition 1 teaches us that the minimal number of measurement to guarantee the $\mathrm{RIP}_p$ of radius $\delta$ is $O\big((\delta^{-2}\,K \log N/K)^{\frac{p}{2}}\big)$ when $p \ge 2$. A small $\delta$ means thus a large measurement number $m$ according to the approximative rule $\delta \propto m^{-1/p}$.

In other words, when the number of measurements $m$ is sufficiently large compared to the sparsity $K$, i.e. more than a few multiples of $K$, this *oversampled* situation allows the search for an optimal $p^* > 2$ compatible with the $\mathrm{RIP}_{p^*}$ of $\Phi$ and reducing the bound on the approximation error.

The situation is even better than for the $\mathrm{BPDQ}_p$ decoder. Indeed, $p$ is not altering explicitly the factor $\frac{2C}{\sqrt{1-\delta}}$, as it was for value $A_p$ and $B_p$ in (9).

*Remark*: The $\ell_\infty$ case is somehow rehabilitated in the stability result compared to the one provided in Theorem 2. Indeed, for $p = \infty$, since $\epsilon_\infty = \frac{\alpha}{2}$, if $\Phi$ is $\mathrm{RIP}_\infty$, i.e. if $\log m \ge O(\delta^{-2}\,K \log N/K)$ of SGR matrix, there exist a $\rho > 0$ such that

$$\|x - x_\infty^*\|_2 \;\le\; \tfrac{2}{\sqrt{1-\delta}}\,\frac{\epsilon_\infty(\alpha)}{\mu_\infty} \;\le\; \frac{\rho\,\alpha}{\sqrt{(1-\delta)\,\log m}}.$$

Interestingly, the error is explictely divided by $\sqrt{\log m}$. However, $\alpha$ may grow like $\sqrt{\log m}$ if we fix for instance the rate (or bit budget) $B$ at which we want to quantize the measurements. Indeed, in that case, $\alpha \propto E[\|\Phi x\|_\infty]/B$, and Lemma 1 points out that there exist $\rho'$ such that $E[\|\Phi x\|_\infty] \le \rho' \sqrt{\log m}$.

Notice however that this effect could vanish with other random matrices such as the Bernoulli generated matrix where each entry is $\pm 1$ with equal probability. In that case, $\|\Phi x\|_\infty$ is upper bounded by the constant $\|x\|_1$, i.e the maximum of $\Phi x$ occurs when one row of $\Phi$ is the sign of the $x$ components. A this time, we do not know if such matrices verify the $\mathrm{RIP}_p$ for $2 \le p \le \infty$. If it is the case, we could reach asymptotically exact recovery from quantized measurement when $m \to \infty$ whatever is the bin width $\alpha$, and so, even for 1-bit quantization! Research is ongoing on this aspect currently.

[7] Actually, for the quantization model with $\epsilon = \epsilon_p(\alpha)$, it is $2K$-sparse with a probability higher than $1 - e^{-8}$ due to the stochastic validity of the bound $\epsilon_p(\alpha)$ with $\kappa = 2$ (Lemma (7)) guaranteeing that $x$ satisifies the fidelity constraint.

## 5 Numerical Implementation

### 5.1 Proximal Optimization

The BPDQ$_p$ (and BPDN) decoders are special case of a general class of convex problems [FS09, CP08]

$$\arg\min_{u\in\mathcal{H}}\ f_1(u) + f_2(u), \tag{P}$$

where $\mathcal{H} = \mathbb{R}^N$ is seen as an Hilbert space equipped with the inner product $\langle u, v\rangle = \sum_i u_i v_i$. We denote by $\operatorname{dom} f = \{x \in \mathcal{H} : f(x) < \infty\}$ the domain of any $f : \mathcal{H} \to \mathbb{R}$. In (P), the functions $f_1, f_2 : \mathcal{H} \to \mathbb{R}$ are assumed *(i)* closed convex functions which are not infinite everywhere, i.e. $\operatorname{dom} f_1, \operatorname{dom} f_2 \neq \emptyset$, *(ii)* $\operatorname{dom} f_1 \cap \operatorname{dom} f_2 \neq \emptyset$, and *(iii)* these functions are lower semi-continuous (lsc) meaning that they respect $\liminf_{x\to x_0} f(x) = f(x_0)$ for all $x_0 \in \operatorname{dom} f$. The class of functions satisfying these three properties is denoted $\Gamma_0(\mathbb{R}^N)$. For BPDQ$_p$, these two non-differentiable functions are $f_1(u) = \|u\|_1$ and $f_2(u) = \imath_{T^p(\epsilon)}(u) = 0$ if $u \in T^p(\epsilon)$ and $\infty$ else, i.e. the indicator function of the set (or tube) $T^p(\epsilon) = \{u \in \mathbb{R}^N : \|y_{\rm q} - \Phi u\|_p \leq \epsilon\}$.

The problem (P) is related to the notion of Proximity operator introduced in [Mor62] as a generalization of convex projection operator.

**Definition 3** (Proximity operator [Mor62])**.** *Let $f \in \Gamma_0(\mathcal{H})$. Then, for every $x \in \mathcal{H}$, the function $z \mapsto \frac{1}{2}\|x-z\|^2 + f(z)$ achieves its infimum at a unique point denoted by $\operatorname{prox}_f x$. The uniquely-valued operator $\operatorname{prox}_f : \mathcal{H} \to \mathcal{H}$ thus defined is the proximity operator of $\varphi$.*

Since $\operatorname{prox}_f z = \arg\min_u \frac{1}{2}\|z - u\|_2^2 + f(u)$, it is clear that if $f = \imath_C$ for some convex set $C \subset \mathcal{H}$, the proximity operator reduces to the orthogonal projection operator onto $C$. We will also use the case where $f(u) = \|u\|_1$, for which the $\operatorname{prox}_{\gamma f_1} z$ is the componentwise soft thresholding of $z$ of threshold $\gamma$ [FS09]. In addition, proximity operators of convex functions respect some nice properties with respect to translation, conjugation with frame operators, dilation, etc. [CP08]

The solutions of problem (P) is also characterized by the following fixed point equation:

$$x \text{ solves (P)} \quad \Leftrightarrow \quad x \ = \ (\mathrm{Id} \ + \ \beta\partial(f_1 + f_2))^{-1}(x), \qquad \text{for } \beta > 0. \tag{11}$$

The operator $J_{\beta\partial f} = (\mathrm{Id} + \beta\partial f)^{-1}$ is called the *resolvent operator* associated to the *subdifferential operator* $\partial f$, $\beta$ is a positive scalar known as the proximal stepsize, and Id 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 $\partial f : \mathcal{H} \to 2^{\mathcal{H}}$

$$\partial f(x) \ = \ \{u \in \mathcal{H} \ : \ \forall z \in \mathcal{H}, f(z) \geq f(x) + \langle u, z - x\rangle\}\,, \tag{12}$$

where each element $u$ of $\partial f$ is called a subgradient. The resolvent operator is thus the proximity operator of $\beta f$, i.e. $J_{\beta\partial f} = \operatorname{prox}_{\beta f}$.

The fixed point relation (11) suggests the iterative *proximal point algorithm*, defined by $x_{n+1} = (\mathrm{Id} \ + \ \beta\partial(f))^{-1}(x_n)$. The main difficulty with the method is that $\mathrm{Id} + \beta\partial f$ may be hard to invert in general, depending on the function $f$. This is for instance the case in most inverse problems arising in image and signal processing. In problem (P) however, $f = f_1 + f_2$ with $f_1, f_2 \in \Gamma_0(\mathcal{H})$ so that we can solve it by monotone operator splitting proximal methods [Com04]. Splitting methods for problem (P) are algorithms that do not attempt to evaluate the resolvent mapping $(\mathrm{Id} + \beta\partial f)^{-1}$ of the combined function $f$, but instead perform a sequence of calculations involving separately the individual resolvent operators $J_{\beta\partial f_1}$ and $J_{\beta\partial f_2}$. The latter are hopefully easier to evaluate.

Since for BPDQ$_p$, both $f_1$ and $f_2$ are non-differentiable functions of $\Gamma_0(\mathcal{H})$, we use the particular Douglas-Rachford, or Douglas/Peaceman-Rachford (DR), splitting. It provides the following compact recursion formula [FS09]

$$u^{(t+1)} \;=\; (1 - \tfrac{\alpha_t}{2})u^{(t)} \;+\; \tfrac{\alpha_t}{2}(2S_\gamma - \mathrm{Id}) \circ (2\mathcal{P}_{T_p(\epsilon)} - \mathrm{Id})(u^{(t)}),$$

where $\alpha_t \in (0,2), \forall t \in \mathbb{N}$, $\gamma > 0$, $S_\gamma = \mathrm{prox}_{\gamma f_1}$ is the component-wise soft-thresholding operator with threshold $\gamma$ and $\mathcal{P}_{T_p(\epsilon)} = \mathrm{prox}_{f_2}$ is the orthogonal projection onto the tube $T^p(\epsilon)$. From [Com04], one can show that the sequence $(u^{(t)})_{t\in\mathbb{N}}$ converges to some point $u^*$ and $\mathcal{P}_{T_p(\epsilon)}(u^*)$ is a solution of BPDQ$_p$. In the next Section, we provide a way to compute $\mathcal{P}_{T_p(\epsilon)}(u^*)$ efficiently.

# 6 Projection onto $\ell^p$ ball via Newton's method

The previous Section showed that each step of the Douglas-Rachford algorithm requires computation of $\mathrm{prox}_{f_2} = \mathcal{P}_{T^p(\epsilon)}$ for $T^p(\epsilon) = \{u \in \mathbb{R}^N : \|y_\mathrm{q} - \Phi u\|_p \leq \epsilon\}$. We present here a way to compute iteratively this projection for any $2 \leq p \leq \infty$.

Setting $B^p(y,\epsilon) \subset \mathbb{R}^m$ to be the $\ell^p$ ball of radius $\epsilon$ centered at $y \in \mathbb{R}^m$, i.e.,

$$B^p(y,\epsilon) = \{y' \in \mathbb{R}^m : \|y' - y\|_p \leq \epsilon\},$$

we then have the composition relation $f_2(x) = (\imath_{B^p(y_\mathrm{q},\epsilon)} \circ \Phi)(x)$. This composition is still compatible with the computation of the proximity operator of $f_2$ under the assumption that $\Phi$ is a frame.

**Lemma 5.** *[FS09] Let a bounded affine operator $A \triangleq \Phi \cdot - v$, with $v \in \mathbb{R}^m$, be associated to a frame $\Phi$ of $\mathcal{H} = \ell^2(\mathbb{R}^m)$ with upper and lower bounds $c_1$ and $c_2$, i.e. $c_1 \,\mathrm{Id} \leq \Phi \circ \Phi^* \leq c_2 \,\mathrm{Id}$, and $f \in \Gamma_0(\mathcal{H})$. Then, $f \circ A \in \Gamma_0(\mathcal{H})$. Moreover, if the frame is tight, i.e. if $c_1 = c_2 = c > 0$, then,*

$$\mathrm{prox}_{f\circ\Phi}(x) = x + c^{-1}\Phi^* \circ \big(\mathrm{prox}_{cf} - \mathrm{Id}\big) \circ (\Phi x - v). \tag{13}$$

*Otherwise, let $0 < \inf_t \mu_t \leq \sup_t \mu_t < 2/c_2$ and define*

$$\begin{aligned} u^{(t+1)} &= \mu_t\big(\mathrm{Id} - \mathrm{prox}_{\mu_t^{-1} f}\big) \circ \big(\mu_t^{-1} u^{(t)} + \Phi \circ (x - \Phi^* u^{(t)}) - v\big), && (14)\\ p^{(t+1)} &= x - \Phi^* u^{(t+1)}. && (15)\end{aligned}$$

*Then $u^{(t)} \to \bar{u}$ and $p^{(t)} \to \mathrm{prox}_{f\circ\Phi} = x - \Phi^*\bar{u}$. More precisely, both $u^{(t)}$ and $p^{(t)}$ converge linearly and the best convergence rate is attained for $\mu_t \equiv 2/(c_1 + c_2)$*

$$\|u^{(t)} - \bar{u}\| \leq \big(\tfrac{c_2 - c_1}{c_2 + c_1}\big)^t \|u^{(0)} - \bar{u}\| \;. \tag{16}$$

Therefore, in presence of a frame operator $\Phi \in \mathbb{R}^{m\times N}$, which happens with very high probability for a SGR matrix $\Phi$, the problem of computing $\mathrm{prox}_{f_2}$ may be reduced to computing the orthogonal projection onto the $\ell^p$ ball $\mathcal{P}_{B^p(y_\mathrm{q},\epsilon)}$. Moreover, by simple recentering and normalization we can consider projection onto the unit $\ell^p$ ball. Explicitly,

$$\mathcal{P}_{B^p(y_\mathrm{q},\epsilon)}(y) \;=\; y_\mathrm{q} \;+\; \epsilon\, \mathcal{P}_{B^p}\big(\tfrac{y - y_\mathrm{q}}{\epsilon}\big),$$

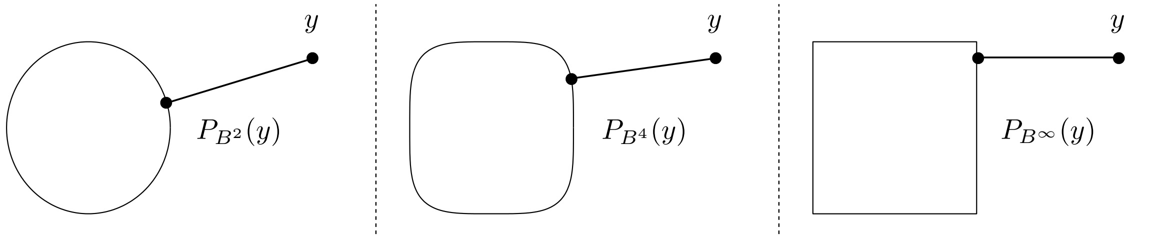



Figure 2: Projection onto $\ell^p$ ball for $p = 2, 4, \infty$

where $B^p = B^p(0, 1)$ is the centered, normalized ball. For $p = 2$ and $p = \infty$ it is straightforward to calculate the projection explicitly. In particular (assuming $y$ is outside the unit $\ell^p$ ball of interest) we have for $p = 2$

$$\mathcal{P}_{B^2}(y) = \frac{y}{\|y\|_2},$$

and for $p = \infty$, component by component,

$$(\mathcal{P}_{B^\infty}(y))_i = \begin{cases} 1 \text{ if } y_i > 1, \\ y_i \text{ if } |y_i| \le 1, \\ -1 \text{ if } y_i < -1. \end{cases}$$

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 $g(u) = \|u\|_p^p$. First note that for $y$ outside of the $\ell^p$ ball, i.e. for $\|y\|_p > 1$, the projection $P_{B^p}$ will lie on the surface of the ball, not in the interior. The projection is thus the solution of the contstrained optimization problem

$$\mathcal{P}_{B^p}(y) \;=\; \arg\min_u f_y(u) \quad \text{s.t.} \quad g(u) = 1$$

As $f_y$ and $g$ are both smoothly differentiable, according to the theory of Lagrange Multipliers, for $\lambda \in \mathbb{R}$, the solution $u^* \in \mathbb{R}^m$ will satisfy the $(m + 1)$ scalar equations

$$\nabla f_y(u^*) = \lambda \nabla g(u^*), \tag{17}$$
$$g(u^*) = 1. \tag{18}$$

Clearly, the point $y$ and its projection $u^*$ belong to the same orthant of $\mathbb{R}^m$, i.e. $y_i u_i^* \ge 0$ for all $1 \le i \le m$. Therefore, without loss of generality, we assume in the sequel that we are working in the positive[8] orthant $u_i \ge 0$ and $y_i \ge 0$.

We will express the previous equations as a single nonlinear multivariate equation, and then use Newton's method to achieve an iterative solution to it. First, let $z = (u, \lambda) \in (\mathbb{R}^+)^m \times \mathbb{R}$ define the augmented variable space. Define $F : \mathbb{R}^{m+1} \to \mathbb{R}^{m+1}$ by

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

[8]The solution of the actual orthant can be obtained with appropriate axis mirroring.

It is easy to see that $F(z) = 0$ is equivalent to equations (17) and (18). Accordingly, the desired projection $u^*$ will be given as the first $m$ coordinates of the value $z^*$ solving $F(z^*) = 0$.

We apply straightforward Newton's method. Given an initialization point $z^0$, the successive iterates are defined by

$$z^{n+1} = z^n - J(z^n)^{-1} F(z^n), \tag{19}$$

where $J_{ij} = \frac{\partial F_i}{\partial z_j}$ is the Jacobian of $F$.

Calculating each component shows that Jacobian has a simple block-invertible form that allows efficient calculation of the inverse applied to any vector. In particular,

$$J = \begin{pmatrix} D & b \\ b^T & 0 \end{pmatrix},$$

where $D \in \mathbb{R}^{m\times m}$ is a diagonal matrix with $D_{ii}(z) = 1 + p(p-1) z_{m+1} z_i^{p-2}$, and $b \in \mathbb{R}^m$ with $b_i(z) = p z_i^{p-1}$ for $1 \le i \le m$. Moreover, using the block inverse formula for block $2\times 2$ matrices[9] and writing $\bar{b} = D^{-1} b$, we see

$$J^{-1} \;=\; \tfrac{1}{\mu} \begin{pmatrix} \mu D^{-1} \;-\; \bar{b}\bar{b}^T & \bar{b} \\ \bar{b}^T & -1 \end{pmatrix},$$

where $\mu = b^T D^{-1} b = \bar{b}^T D\, \bar{b}$. Writing $u \in \mathbb{R}^{m+1}$ as $u = (\vec{u}, u_{m+1})^T$, we have

$$J^{-1} u \;=\; \tfrac{1}{\mu} \begin{pmatrix} \mu D^{-1}\vec{u} \;+\; \big(u_{m+1} - \bar{b}^T \vec{u}\big)\bar{b} \\ (\bar{b}^T \vec{u} - u_{m+1}) \end{pmatrix}, \tag{20}$$

which can be computed efficiently as $D$ is diagonal.

We initialize the first $m$ components of $z^0$ by computing the direct radial projection of $y$ to the $\ell^p$ ball, $\vec{z}^{\,0} = y/\|y\|_p$. The $m+1$ component, the lagrange multiplier, is then initialized as the best least squares solution of (17), i.e.

$$z^0_{m+1} \;=\; \operatorname*{argmin}_{\lambda} \sum_{i=1}^{m} \big(y_i - z_i^0 - \lambda p (z_i^0)^{p-1}\big)^2 \;=\; \|b(\vec{z}^{\,0})\|_2^{-2} \,\big(b^T(\vec{z}^{\,0})(y - \vec{z}^{\,0})\big).$$

The algorithm is then run, using (20) to calculate each update step according to (19). We terminate iteration when the norm of $\|F(z^n)\|_2$ falls below a specified tolerance. In practice, we find extremely rapid convergence, achieving results with error comparable to machine precision with typically fewer than 10 iterations.

## 7 Experiments

For the first experiment, setting the dimension $N = 1024$ and the sparsity level $K = 16$, we have generated 500 $K$-sparse signals with support selected uniformly at random in $\{1, \cdots, N\}$. The non-zero elements have been drawn from a standard Gaussian distribution $N(0, 1)$. For each sparse signal, $m$ quantized measurements have been recorded as in model (1) with a SGR matrix

[9] See for instance page 125 of [Zwi03].

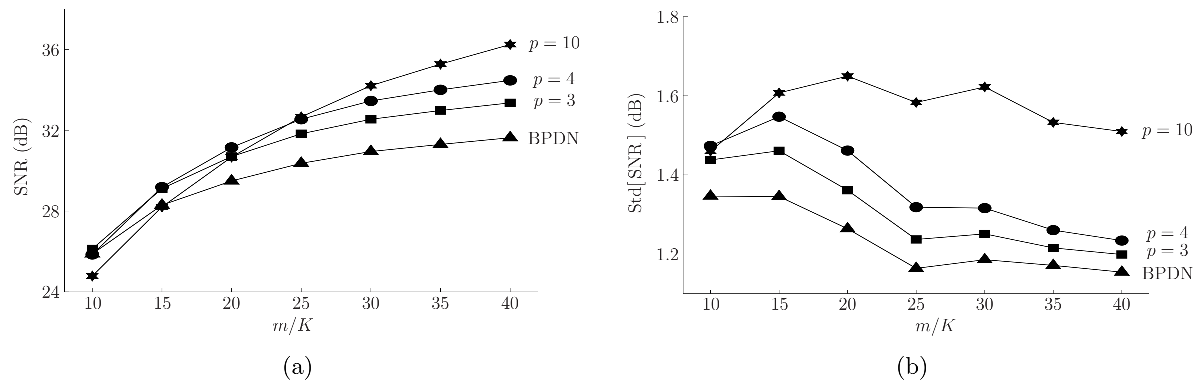


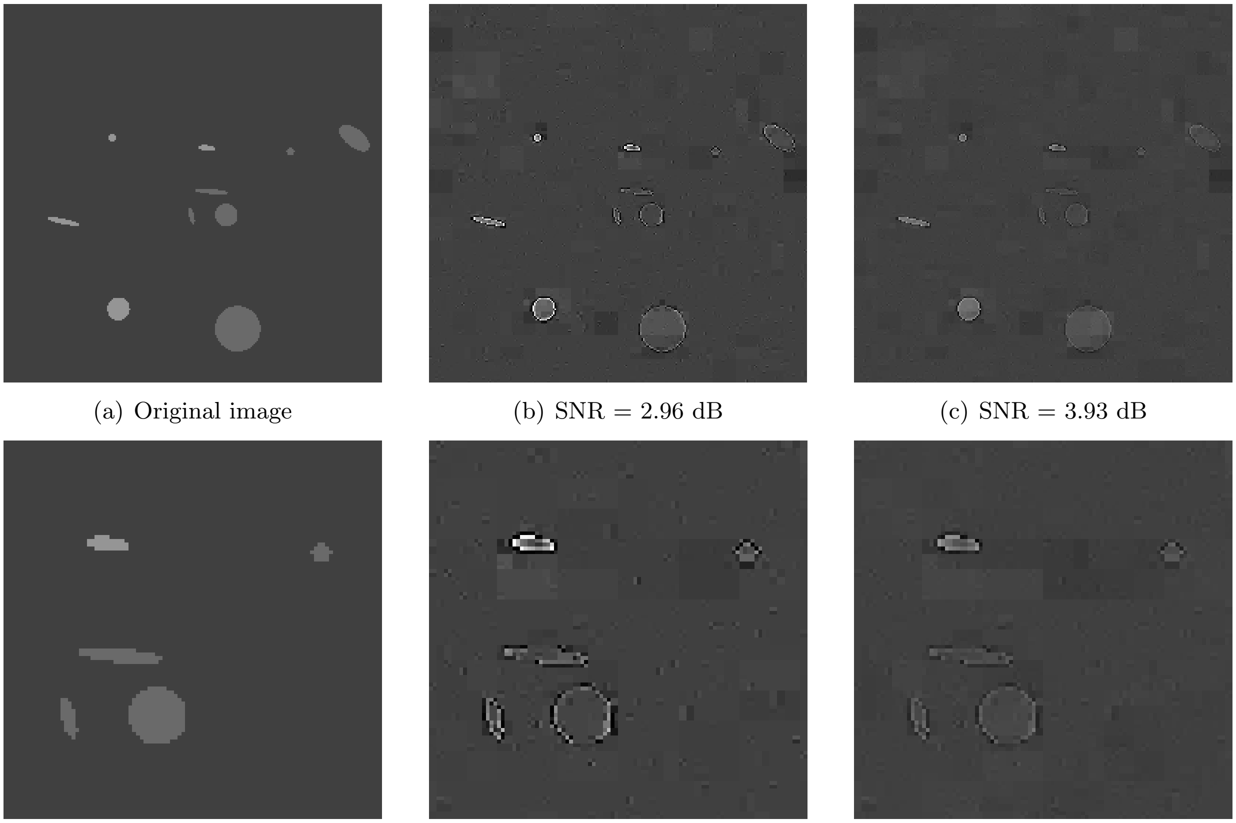

Figure 3: Quality of $\text{BPDQ}_p$ for different $m/K$ and $p$. (a) Average SNR. (b) Standard deviation of the SNR.

$\Phi \in \mathbb{R}^{m \times N}$. The bin width has been set to $\alpha = \|\Phi x\|_\infty/40$. In Figure 3, we plot the average quality of the reconstructions of $\text{BPDQ}_p$ for various $p \geq 2$ and $m/K \in [10, 40]$. We use the quality metric $\text{SNR}(\hat{x}; x) = 20 \log_{10} \frac{\|x\|_2}{\|x - \hat{x}\|_2}$, where $x$ is the true original signal and $\hat{x}$ the reconstruction. The different decoders become dominant from oversampling factors $m/K$ increasing with $p$. This confirms the fact that the noise error can be reduced when both $p$ and $m/K$ are high.

In the second experiment, we applied our methods to a model of undersampled MRI reconstruction problem. Using an example similar to [LDP07], the original signal is a 256 by 256 pixel "simulated angiogram" comprised of 10 randomly placed ellipses. The linear measurements are the real and imaginary components of one sixth of the Fourier coefficients at randomly selected locations in Fourier space, giving $m = 256^2/6$ independent measurements. These are quantized with a bin width $\alpha$ giving at most 12 quantization levels for each measurement. We use the Haar wavelet transform as a sparsity basis. The measurement matrix is then $\Phi = F\Psi$, where $\Psi$ is the Haar matrix, and $F$ is formed by the randomly selected rows of the Discrete Fourier Transform matrix. The original image has $K = 821$ nonzero wavelet coefficients, giving an oversampling ratio $m/K = 13.3$. In Figure 4, we show 100 by 100 pixel details of the results of reconstruction with BPDN, and with BPDQ for $p = 10$. Note that we do not have any proof that the sensing matrix $\Phi$ satisfies the $\text{RIP}_p$ (4). We nonetheless obtain similar results as in the previous 1-d example. The BPDQ reconstruction shows improvements both in SNR and visual quality compared to BPDN.

# 8 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 $\text{BPDQ}_p$ exhibit an substantial reduction of the approximation error in oversampled situations. An 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 decoder in function of the precise CS coding/decoding scenario.



(a) Original image (b) SNR = 2.96 dB (c) SNR = 3.93 dB

(d) Zoom on (a) (e) Zoom on (b) (f) Zoom on (c)

Figure 4: Reconstruction from quantized undersampled Fourier measurements. (a) Original. (b) BPDN. (c) $\text{BPDQ}_{10}$. (d), (e) and (f) details on (a), (b) and (c) respectively.

# 9 Acknowledgements

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. LJ is a Postdoctoral Researcher of the Belgian National Science Foundation (F.R.S.-FNRS).

# A Proof of Proposition 1

Before proving Proposition 1, let us recall some facts of measure concentrations [LT91, Led01].

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 6** (Ledoux, Talagrand [LT91] (Eq. 1.6))**.** *If $F$ is Lipschitz with $\lambda = \|F\|_{\mathrm{Lip}}$, then, for the random vector $\xi \in \mathbb{R}^m$ with $\xi_i \sim_{\mathrm{iid}} N(0,1)$,*

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

*with $\mu_F = E[F(\xi)] = \int_{\mathbb{R}^m} F(x)\,\gamma^m(x)\,\mathrm{d}^m x$ 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 $\epsilon$-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 $\|t - s\|_2 \le \epsilon$. In certain cases, the size of a $\epsilon$-net can be bounded.

**Lemma 7** ([Led01])**.** *There exists a $\epsilon$-net $\mathcal{S}$ of the unit sphere of $\mathbb{R}^K$ of size $|\mathcal{S}| \le (1+\frac{2}{\epsilon})^K \le e^{2K/\epsilon}$.*

We will use also this fundamental result.

**Lemma 8** ([Led01])**.** *Let $\mathcal{S}$ be a $\epsilon$-net of the unit sphere in $\mathbb{R}^K$. Then, if for some $v_1, \cdots, v_k$ in the Banach space $B$ normed by $\|\cdot\|_B$, we have $1 - \epsilon \le \big\|\sum_{i=1}^k s_i v_i\big\|_B \le 1 + \epsilon$ for all $s \in \mathcal{S}$, then*

$$(1-\beta)\,\|t\|_2 \;\le \Big\|\sum_{i=1}^k t_i v_i\Big\|_B \;\le\; (1+\beta)\,\|t\|_2,$$

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

In our case, the Banach space $B$ is $\ell_p(\mathbb{R}^m)$ for $1 \le p \le \infty$, 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 $p \ge 1$. We must prove that for a SGR matrix $\Phi \in \mathbb{R}^{m\times N}$, i.e. with $\Phi_{ij} \sim_{iid} N(0,1)$, with the right number of measurements $m$, there exist a radius $0 < \delta < 1$ and a constant $\mu_p > 0$ such that

$$\mu_p\sqrt{1-\delta}\,\|x\|_2 \;\le\; \|\Phi x\|_p \;\le\; \mu_p\sqrt{1+\delta}\,\|x\|_2, \quad \forall\, x \in \mathbb{R}^N \;\text{ s.t. }\; \|x\|_0 \le K. \tag{21}$$

The outline of our proof is as follows. We will start with using the spherical symmetry of the multivariate Gaussian to obtain a concentration result valid for any single unit norm vector. We will extend this using the union bound to a result valid for an $\epsilon$-net of a single unit sphere restricted to support of size $K$. We will then apply Lemma 8 to get the bound involving the $\ell_2$ norm. Finally we will use the union bound again to obtain a result valid for all $K$-sparse vectors.

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

$$P_\Phi\big[\,|F(\Phi s) - \mu_F| > r\,\big] = \int_{\mathbb{R}^m} \mathrm{d}^m\phi_1 \cdots \int_{\mathbb{R}^m} \mathrm{d}^m\phi_N\; \big[\Pi_{i=1}^N \gamma^m(\phi_j)\big]\; i_G(\phi s),$$

with $\phi = (\phi_1, \cdots, \phi_N) \in \mathbb{R}^{m\times N}$, $G = \{y \in \mathbb{R}^m : |F(y) - \mu_F| > r\}$, and $i_G(y)$ the indicator of $G$ equals to 1 if $y \in G$ and 0 elsewhere.

We show now that for any Lipschitz function $F : \mathbb{R}^m \to \mathbb{R}$ with $\lambda = \|F\|_{\rm Lip}$,

$$P_\Phi\big[\,|F(\Phi s) - \mu_F| > r\,\big] \;\leq\; 2e^{-\frac{r^2}{2\lambda^2}}, \quad \text{for } r > 0.$$

Indeed, taking a rotation matrix $R \in \mathbb{R}^{N\times N}$ such that $Rs = e_1 = (1, 0, \cdots, 0)^T \in \mathbb{R}^N$ (since $\|s\|_2 = 1$), since $\Pi_{i=1}^N \gamma^m(\phi_j) \propto e^{-\sum_j \|\phi_j\|_2^2/2} = e^{-\sum_{ij} |\phi_{ij}|^2/2}$ is invariant under the rotation of the rows of $\phi$, the $N$ changes of variables $\phi_j \to \phi_j R$, i.e. $\phi \to \phi R$, and the Lemma 6 lead to

$$\begin{aligned}
P_\Phi\big[\,|F(\Phi s) - \mu_F| > r\,\big] \;&=\; \int_{\mathbb{R}^m} \mathrm{d}^m\phi_1 \big[\int_{\mathbb{R}^m} \mathrm{d}^m\phi_2 \cdots \int_{\mathbb{R}^m} \mathrm{d}^m\phi_N \;\; \Pi_{i=2}^N \gamma^m(\phi_j)\big]\; \gamma^m(\phi_1)\, i_G(\phi e_1)\\
&=\; \int_{\mathbb{R}^m} \mathrm{d}^m\phi_1\; \gamma^m(\phi_1)\, i_G(\phi_1)\\
&=\; P_{\Phi_1}\big[\,|F(\Phi_1) - \mu_F| > r\,\big] \;\leq\; 2e^{-\frac{r^2}{2\lambda^2}}.
\end{aligned}$$

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 $|\mathcal{S}_T| \leq e^{2K/\epsilon}$ by Lemma 7, setting $r = \epsilon\mu_F$ for $\epsilon > 0$, we obtain

$$P_\Phi\big[\,|\,\mu_F^{-1}\, F(\Phi s) - 1| > \epsilon\big] \;\leq\; 2\, e^{2K/\epsilon}\, e^{-\frac{\epsilon^2\mu_F^2}{2\lambda^2}}, \quad \forall s \in \mathcal{S}_T.$$

Taking now $F(\cdot) = \|\cdot\|_p$ for $1 \leq p \leq \infty$, we have $\mu_F = \mu_p = E[\|\xi\|_p]$ 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,

$$(1-\epsilon) \;\leq\; \|\tfrac{1}{\mu_p}\Phi s\|_p \;\leq (1+\epsilon), \tag{22}$$

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

We apply Lemma 8 by noting that, as $s$ has support of size $K$, (22) may be written as

$$1-\epsilon \;\leq\; \|\sum_{i=1}^k s_i v_i\|_p \;\leq\; 1+\epsilon$$

where $v_i$ are the columns of $\frac{1}{\mu_p}\Phi$ corresponding to the support of $s$ (we abuse notation to let $s_i$ range only over the support of $s$). Then according to Lemma 8 we have, with the same probability bound and with for $(\sqrt{2}-1)\delta = \frac{2\epsilon}{1-\epsilon}$,

$$\sqrt{1-\delta}\,\|x\|_2 \;\leq\; (1-(\sqrt{2}-1)\delta)\,\|x\|_2 \;\leq\; \|\Phi x\|_p \;\leq\; (1+(\sqrt{2}-1)\delta)\,\|x\|_2 \;\leq\; \sqrt{1+\delta}\,\|x\|_2, \tag{23}$$

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

The result can be made independent of the choice of $T \subset \{1, \cdots, N\}$ by considering that there are $\binom{N}{K} \leq (eN/K)^K$ such possible supports. Therefore, applying again an union bound, (23) holds for all $K$-sparse $x$ in $\mathbb{R}^N$ with a probability higher than

$$1 - 2\, e^{K\,(1+\log\frac{N}{K}+2\epsilon^{-1}) - \frac{\epsilon^2\mu_p^2}{2\lambda_p^2}}.$$

To go further in the bounding of this probability, let us take first $2 \leq p < \infty$ and $m \geq \beta^{-1}\, 2^{p+1}$ with $\beta^{-1} = p-1$. Lemma 2 (page 10) tells us that $\mu_p \geq \frac{p-1}{p}\nu_p\, m^{\frac{1}{p}}$ with $\nu_p = \sqrt{2}\,\pi^{-\frac{1}{2p}}\,\Gamma(\frac{p+1}{2})^{\frac{1}{p}}$. Remembering that $\epsilon = \frac{(\sqrt{2}-1)\delta}{2+(\sqrt{2}-1)\delta}$, a probability of success $1-\eta(\delta)$ with $\eta < 1$ is then guaranteed if we select:

*(i)* for $1 \leq p < 2$, since $\lambda_p = m^{\frac{2-p}{2p}}$,

$$m \;>\; \tfrac{2}{\epsilon^2\nu_p^2}\,(\tfrac{p}{p-1})^2\,\big(K\,(1+\log\tfrac{N}{K}+2\epsilon^{-1}) \;+\; \log\tfrac{2}{\eta}\big),$$

*(ii)* for $2 \leq p < \infty$, since $\lambda_p = 1$,

$$m \;>\; \tfrac{2^{\frac{p}{2}}}{\epsilon^p\nu_p^p}(\tfrac{p}{p-1})^p\,\big(K\,(1+\log\tfrac{N}{K}+2\epsilon^{-1}) \;+\; \log\tfrac{2}{\eta}\big)^{\frac{p}{2}}, \tag{24}$$

Second, for $p = \infty$ for $p = \infty$, since there exists a $\rho > 0$ such that $\mu_\infty \;\geq\; \rho^{-1}\sqrt{\log m}$, with $\lambda_\infty = 1$,

$$\log m \;>\; \tfrac{2\rho^2}{\epsilon^2}\,\big(K\,(1+\log\tfrac{N}{K}+2\epsilon^{-1}) \;+\; \log\tfrac{2}{\eta}\big).$$

The complexities are finally deduced by isolating the values dependent on $K$, $N$ and $\epsilon$ from the others. □

We may realize 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 [BDD08], i.e. $\Phi$ satisfies the common RIP = $\text{RIP}_2$ with probability higher than $1-\eta$ if

$$m \;>\; \tfrac{8}{\epsilon^2}\,K\,(1+\log\tfrac{N}{K}+2\epsilon^{-1}) \;+\; \log\tfrac{2}{\eta}.$$

Second, we observe also that the Euclidean case $p = 2$ provides the lowest bound on $m$. Indeed, $\nu_p$ is a increasing function of $p$ with for instance $c_1 = 0.7979$ and $c_2 = 1$. Therefore, in the range $1 \leq p \leq 2$ where the complexity is constant since the exponents are not varying, the smallest proportional factor (i.e. highest $\nu_p$ and lowest $\frac{p}{p-1} = 1 + \frac{1}{p-1}$) is reached for $p = 2$, at constant $\epsilon$.

Finally, as for the comparison between the common $\text{RIP}_2$ proof [BDD08] and the tight bound found in [DT09], the requirements on the measurements above are possibly pessimistic, i.e. the exponent $p/2$ occuring in (24) is perhaps too high. The tightening of the requirements should be performed with more precise measure concentration tools that the ones used so far in this paper. Proposition (4) has however the merit to prove that random Gaussian matrices satisfy the $\text{RIP}_p$ in a certain range of dimensionality.

# B Proof of Lemma 1

*Proof.* The result for $p = \infty$ comes from [LT91] (see Eq (3.14)). For $1 \leq p < \infty$, it is a specialization of the results presented in [FWV07] (see Lemmata 1 and 2). Indeed, it is proved there that, for any probabilistic distribution $\mathcal{P}$, if $\xi \in \mathbb{R}^m$ is a random vector such that $\xi_i \sim_{\text{iid}} \mathcal{P}$, then, for $g \sim \mathcal{P}$,

$$\lim_{m\to\infty}\; m^{-\frac{1}{p}}\,E[\|\xi\|_p] \;=\; E[|g|^p]^{\frac{1}{p}},$$

and

$$\lim_{m\to\infty} m^{1-\frac{2}{p}}\mathrm{Var}[\|\xi\|_p] \;=\; \frac{\mathrm{Var}[|g|^p]}{\big(p\,E[|g|^p]^{\frac{p-1}{p}}\big)^2},$$

so that

$$\lim_{m\to\infty} \frac{\sqrt{\mathrm{Var}[\|\xi\|_p]}}{E[\|\xi\|_p]} \;=\; 0.$$

In other words, there exists a concentration phenomenon of $E[\|\xi\|_p]$ around the limit $E[|g|^p]^{\frac{1}{p}}$ as $m$ increases. The result follows then from $\mathrm{Var}[|g|^p] = E[|g|^{2p}] - E[|g|^p]^2$ and by taking $\mathcal{P} = N(0,1)$ and computing explicitly $E[|g|^p]$. Indeed,

$$\begin{aligned} E[|g|^p] \;&=\; \tfrac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} |t|^p e^{-\frac{1}{2}t^2}\,\mathrm{d}t \;=\; (\tfrac{2}{\pi})^{\frac{1}{2}}\int_0^{\infty} t^p e^{-\frac{1}{2}t^2}\,\mathrm{d}t \;=\; 2^{\frac{p}{2}}\pi^{-\frac{1}{2}}\int_0^{\infty} u^{\frac{p-1}{2}} e^{-u}\,\mathrm{d}u \\ &=\; 2^{\frac{p}{2}}\,\pi^{-\frac{1}{2}}\,\Gamma(\tfrac{p+1}{2}). \end{aligned}$$

□

# C Proof of Lemma 2

*Proof.* Let $\xi \in \mathbb{R}^m$ be a SGR vector, i.e. $\xi_i \sim_{iid} N(0,1)$ for $1 \le i \le m$, and $p \ge 1$. First, the inequality $E[\|\xi\|_p] \le E[\|\xi\|_p^p]^{1/p}$ follows from the application of the Jensen inequality $\varphi\big(E[\|\xi\|_p]\big) \le E[\varphi(\|\xi\|_p)]$ with the convex function $\varphi(\cdot) = (\cdot)^p$.

Second, the lower bound on $E[\|\xi\|_p]$ arises from the observation that for $f : \mathbb{R}^+ \to \mathbb{R}^+$ with $f(t) = t^{\frac{1}{p}}$, and for a given $t_0 > 0$,

$$f(t) \;\ge\; f(t_0) \;+\; f'(t_0)\,(t-t_0) \;+\; p\,f''(t_0)\,(t-t_0)^2, \quad \forall t \ge 0. \tag{25}$$

Indeed, observe first that since $f(\frac{t}{\alpha}) = \alpha^{-\frac{1}{p}} f(t)$ for $\alpha > 0$, if the relation above is valid for one $t_0 > 0$, then it is also true for $t_0' = \alpha t_0$, i.e.

$$f(t) \;\ge\; f(\alpha t_0) \;+\; f'(\alpha t_0)\,(t-\alpha t_0) \;+\; p\,f''(\alpha t_0)\,(t-\alpha t_0)^2,$$

where we took $t \to t/\alpha$ in the previous relation. We can thus assume $t_0 = 1$, so that, after some simplifications, we have to prove

$$f(t) = t^{\frac{1}{p}} \;\ge\; \tfrac{2p-1}{p}\,t \;-\; \tfrac{p-1}{p}\,t^2,$$

or equivalently,

$$t^{\frac{1}{p}-1} + \tfrac{p-1}{p}\,t \;\ge\; \tfrac{2p-1}{p}.$$

The LHS of this last inequality takes its minimum in $t = 1$ with value $\frac{2p-1}{p}$, which provides the result.

Let us define $\bar\mu_p = E[\|\xi\|_p^p]$. We can rewrite (25) as

$$\begin{aligned} f(t) \;\;\ge\;\; & f(t_0) \;+ f'(t_0)\,(\bar\mu_p - t_0) \;+\; p\,f''(t_0)\,(\bar\mu_p - t_0)^2 \\ & + f'(t_0)\,(t-\bar\mu_p) \;+\; 2p\,f''(t_0)\,(t-\bar\mu_p)(\bar\mu_p - t_0) \\ & + p\,f''(t_0)\,(t-\bar\mu_p)^2. \end{aligned}$$

Since $\mu_p = E[\|\xi\|_p] = E[f(\|\xi\|_p^p)]$ and $E[\|\xi\|_p^p - \bar\mu_p] = 0$, we find

$$\begin{aligned} \mu_p &\geq f(t_0) \; + f'(t_0)\,(\bar\mu_p - t_0) \; + \; p\,f''(t_0)\,(\bar\mu_p - t_0)^2 \; + \; p\,f''(t_0)\,\bar\sigma_p^2, \\ &= (t_0)^{\frac{1}{p}-2}\,\big(\,(2-\tfrac{1}{p})\,\bar\mu_p t_0 \; + \; (\tfrac{1}{p}-1)\,(\bar\mu_p^2 + \bar\sigma_p^2)\,\big) \end{aligned}$$

writing $\bar\sigma_p^2 = E[(\|\xi\|_p^p - \bar\mu_p)^2] = \mathrm{Var}[\|\xi\|_p^p]$. The RHS of the last inequality is maximum for $t_0 = \bar\mu_p\,(1 + \bar\mu_p^{-2}\,\bar\sigma_p^2)$. For that value, we get finally

$$\mu_p \;\geq\; E[\|\xi\|_p^p]^{\frac{1}{p}}\;\big(1 \;+\; E[\|\xi\|_p^p]^{-2}\,\mathrm{Var}[\|\xi\|_p^p]\big)^{\frac{1}{p}-1}.$$

Because of the decorrelation of the components of $\xi$, the last inequality simplifies into

$$\mu_p \;\geq\; m^{\frac{1}{p}}\,E[|g|^p]^{\frac{1}{p}}\;\big(1 \;+\; m^{-1}E[|g|^p]^{-2}\,\mathrm{Var}[|g|^p]\,\big)^{\frac{1}{p}-1},$$

with $g \sim N(0,1)$. From Appendix B, we know that $E[|g|^p] = 2^{\frac{p}{2}}\,\pi^{-\frac{1}{2}}\,\Gamma(\frac{p+1}{2})$.

Moreover, using the following approximation of the Gamma function [Spi71]

$$|\Gamma(x) - (\tfrac{2\pi}{x})^{\frac{1}{2}}(\tfrac{x}{e})^x| \;\leq\; \tfrac{1}{9x}\,(\tfrac{2\pi}{x})^{\frac{1}{2}}(\tfrac{x}{e})^x,$$

valid for $x \geq 1$, we observe

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

that holds also if $x = \frac{p+1}{2}$ with $p \geq 1$. Therefore,

$$E[|g|^p]^{-2}\,\mathrm{Var}[|g|^p] \;=\; \big(\,\tfrac{1.1}{0.9^2}\,(\tfrac{e}{2})^{\frac{1}{p}}\,(\tfrac{2p+1}{p+1})^p \;-\; 1\,\big) \;\leq \tfrac{1.1}{0.9^2}\,(\tfrac{e}{2})^{\frac{1}{2}}\,2^p,$$

and finally

$$\mu_p \;\geq\; m^{\frac{1}{p}}\,E[|g|^p]^{\frac{1}{p}}\,\big(1 \;+\; c\,\tfrac{2^p}{m}\,\big)^{\frac{1}{p}-1}$$

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

# D Proof of Lemma 3

*Proof.* 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 $\ell_p = \ell_p(\mathbb{R}^N)$ for $p \geq 2$. In [Byn76, Xu91], it is explained that this space is $p$-convex and 2-smooth, i.e. there exist constants $\mathrm{m}_p, \mathrm{M}_p > 0$ such that for all $x, y \in \mathbb{R}^N$,

$$\|x+y\|_p^p \;\geq\; \|x\|_p^p \;+\; p\,\langle J_p(x), y\rangle \;+\; \mathrm{m}_p\,\|y\|_p^p, \qquad (p-\text{convex})$$

$$\|x+y\|_p^2 \;\leq\; \|x\|_p^2 \;+\; 2\,\langle J(x), y\rangle \;+\; \mathrm{M}_p\,\|y\|_p^2, \qquad (2-\text{smooth})$$

where, for $r \geq 1$, $J_r$ is the *duality* mapping of *gauge function* $t \to t^{r-1}$. For the Hilbert space $\ell_2$, these two relations reduce of course to the *polarization identity*. For $\ell_p$, $J_r$ is the differential of $\frac{1}{r}\|\cdot\|_p^r$, i.e. $(J_r(u))_i = \|u\|^{r-p}\,|u_i|^{p-1}\,\mathrm{sign}\,u_i$, with the short hand $J = J_2$. In addition, $\mathrm{M}_p = (p-1)$ and $\mathrm{m}_p = (1+t_p^{p-1})(1+t_p)^{1-p}$ with $t_p$ the unique solution of the equation $(p-2)\,t^{p-1} + (p-1)t^{p-2} = 1$ for $t \in (0,1)$. Since $(1+t_p)^{p-1} \leq 2^{p-2}(1+t_p^{p-1})$, we have $\mathrm{m}_p \geq 2^{2-p}$.

The smoothness inequality involves

$$2\,\langle J(x), y\rangle \;\leq\; \|x\|_p^2 \;+\; (p-1)\,\|y\|_p^2 - \|x-y\|_p^2, \tag{26}$$

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

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

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

where the absolute value on the inner product arises from the invariance of the RIP bound on (26) under the change $y \to -y$. The value $\mu_p^{-2}|\langle J(\Phi u), \Phi v\rangle|$ 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

$$|\langle J(\Phi u), \Phi v\rangle| \;\leq\; \mu_p^2\left[(\delta_s + \delta_{s+s'})\left((p-2) + (p-1)\delta_{s'} + \delta_{s+s'}\right)\right]^{\frac{1}{2}}. \tag{27}$$

In parallel, a change $y \to x+y$ in (26) 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 suming this inequality with (26), 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 $\mathrm{RIP}_p$ on $x = \Phi u$ and $y = t\Phi v$ as above leads to

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

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

$$|\langle J(\Phi u), \Phi v\rangle| \;\leq\; \tfrac{1}{2}\,\mu_p^2\left[(p-2+p\,\delta_{s+s'})\left(2(p-2) + (p-2)\delta_{s'} + p\,\delta_{s+s'}\right)\right]^{\frac{1}{2}}. \tag{28}$$

Together, (27) and (28) imply the desired result.

□

# E Proof of Theorem 2

*Proof.* Let us write $x_p^* = x + h$. 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 $A \subset \{1, \cdots, d\}$ and 0 elsewhere.

We define $T_0 = \operatorname{supp} x_K$ 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 will frequently use the relation

$$\|z\|_1 \leq \sqrt{K}\|z\|_2 \tag{29}$$

which follows from the Cauchy-Schwartz inequality, valid for vectors $z$ with support of size K.

We start from

$$\|h\|_2 \;\leq\; \|h_{T_{01}}\|_2 \;+\; \|h_{T_{01}^c}\|_2, \tag{30}$$

with $T_{01} = T_0 \cup T_1$, and we are going to bound separately the two terms of the RHS.

For all $j \in T_{k+1}$, we have $|h_j| \leq \frac{1}{K}\,\|h_{T_k}\|_1$ and therefore $\|h_{T_{k+1}}\|_2 \leq K^{-\frac{1}{2}}\|h_{T_k}\|_1$. Consequently,

$$\|h_{T_{01}^c}\|_2 \;\leq\; \sum_{k\geq 2}\|h_{T_k}\|_2 \;\leq\; K^{-\frac{1}{2}}\sum_{k\geq 1}\|h_{T_k}\|_1 \;\leq\; K^{-\frac{1}{2}}\|h_{T_0^c}\|_1. \tag{31}$$

But $\|h_{T_0^c}\|_1$ cannot be very large if $x_p^*$ arises from a $\ell_1$ minimization. As $\epsilon$ is set such that $x$ is a feasible point of the fidelity constraint in $\Delta_p(y,\epsilon)$, the solution $x_p^*$ must have a lower $\ell_1$ norm than $x$. This implies

$$\|x\|_1 \geq \|x+h\|_1 = \|(x+h)_{T_0}\|_1 + \|(x+h)_{T_0^c}\|_1 \geq \|x_{T_0}\|_1 - \|h_{T_0}\|_1 + \|h_{T_0^c}\|_1 - \|x_{T_0^c}\|_1, \tag{32}$$

implying

$$\|h_{T_0^c}\|_1 \leq \|h_{T_0}\|_1 + \|x\|_1 - \|x_{T_0}\|_1 + \|x_{T_0^c}\|_1 = \|h_{T_0}\|_1 + 2\|x_{T_0^c}\|_1 \leq \sqrt{K}\,\|h_{T_{01}}\|_2 + 2\|x_{T_0^c}\|_1. \tag{33}$$

where we use (29) for the last inequality. Inequalites (31) and (33) together imply

$$\|h\|_2 \;\leq\; \|h_{T_{01}}\|_2 \;+\; \|h_{T_{01}^c}\|_2 \;\leq\; 2\|h_{T_{01}}\|_2 \;+\frac{2}{\sqrt{K}}\|x_{T_0^c}\|_1$$

Using 29 and noting that $\|x_{T_0^c}\|_2 = e_0(K)$, this becomes

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

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

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

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

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

since $\|\Phi h\|_p \leq \|\Phi x - y\|_p + \|\Phi x_p^* - y\|_p \leq 2\epsilon$. Using Lemma 3, we know that, for $k \geq 2$,

$$|\langle J(\Phi h_{T_{01}}), \Phi h_{T_k}\rangle| \;\leq\; \mu_p^2\, C_p(\delta_K, \delta_{2K}, \delta_{3K})\,\|h_{T_{01}}\|_2\,\|h_{T_k}\|_2,$$

so that, using again the RIP$_p$ of $\Phi$,

$$
\begin{aligned}
(1-\delta_{2K})\mu_p^2\,\|h_{T_{01}}\|_2^2 \;&\le\; \|\Phi h_{T_{01}}\|_p^2 \\
&\le\; 2\,\epsilon\,\mu_p\,(1+\delta_{2K})^{\frac{1}{2}}\,\|h_{T_{01}}\|_2 \;+\; \mu_p^2\,C_p\,\|h_{T_{01}}\|_2 \sum_{k\ge 2}\|h_{T_k}\|_2 \\
&\le\; 2\,\epsilon\,\mu_p\,(1+\delta_{2K})^{\frac{1}{2}}\,\|h_{T_{01}}\|_2 \;+\; \mu_p^2\,C_p\,\|h_{T_{01}}\|_2\big(\,\|h_{T_{01}}\|_2 + 2e_0(K)\,\big).
\end{aligned}
$$

After some simplifications, we get finally

$$
\|h\|_2 \;\le\; \tfrac{2(C_p+1-\delta_{2K})}{1-\delta_{2K}-\,C_p}\;e_0(K) \;+\; \tfrac{4\sqrt{1+\delta_{2K}}}{1-\delta_{2K}-\,C_p}\;\tfrac{\epsilon}{\mu_p}.
$$

□

# F Proof of Lemma 4

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

To prove the probabilistic inequality below (7), we define, for $1 \le i \le N$, the positive random variables $Z_i = \frac{2^p}{\alpha^p}|\xi_i|^p$ bounded on the interval $[0,1]$ with $E[Z_i] = (p+1)^{-1}$. Denoting $S = \frac{1}{m}\sum_i Z_i$, the Chernoff-Hoeffding bound [Hoe63] tells us that, for $t \ge 0$,

$$
P\big[\,S \;\ge\; (p+1)^{-1} \;+\; t\,\big] \;\le\; e^{-2t^2 m}.
$$

Therefore,

$$
P\big[\|\xi\|_p^p \;\ge\; \tfrac{\alpha^p}{2^p(p+1)}\,m + \tfrac{\alpha^p}{2^p}\,t\,m\big] \;\le\; e^{-2t^2 m},
$$

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

$$
P\big[\|\xi\|_p^p \;\ge\; \zeta_p \;+\; \tfrac{\alpha^p}{2^p}\,\kappa\, m^{\frac{1}{2}}\big] \;\le\; e^{-2\kappa^2}.
$$

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

# References


[BDD08] 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*, **28**(3):253–263, December 2008.

[Byn76] WL Bynum. "Weak parallelogram laws for Banach spaces." *Canad. Math. Bull*, **19**(3):269–275, 1976.

[Can08] E. Candès. "The restricted isometry property and its implications for compressed sensing." *Compte Rendus de l'Academie des Sciences, Paris, Serie I*, **346**:589–592, 2008.

[CDD09] A. Cohen, R. DeVore, and W. Dahmen. "Compressed sensing and best k-term approximation." *J. Amer. Math. Soc.*, **22**:211–231, 2009.

[Com04] P.L. Combettes. "Solving monotone inclusions via compositions of nonexpansive averaged operators." *Optimization*, **53**(5):475–504, 2004.

[CP08] P.L. Combettes and J.C. Pesquet. "A Proximal Decomposition Method for Solving Convex Variational Inverse Problems." *Inverse Problems*, **24**:27, December 2008.

[CR06] E.J. Candès and J. Romberg. "Quantitative Robust Uncertainty Principles and Optimally Sparse Decompositions." *Foundations of Computational Mathematics*, **6**(2):227–254, 2006.

[CRT06] E. Candès, J. Romberg, and T. Tao. "Stable signal recovery from incomplete and inaccurate measurements." *Comm. Pure Appl. Math*, **59**(8):1207–1223, 2006.

[CT04] E. J. Candès and T. Tao. "Near-optimal signal recovery from random projections: universal encoding strategies." *IEEE Trans. Inform. Theory*, **52**:5406–5425, 2004.

[Don06] DL Donoho. "Compressed Sensing." *Information Theory, IEEE Transactions on*, **52**(4):1289–1306, 2006.

[DPM09] Wei Dai, Hoa Vinh Pham, and Olgica Milenkovic. "Distortion-Rate Functions for Quantized Compressive Sensing." *preprint*, 2009. arXiv:0901.0749.

[DT09] David Donoho and Jared Tanner. "Counting faces of randomly-projected polytopes when the projection radically lowers dimension." *Journal of the AMS*, **22**(1):1–15, January 2009.

[FS09] M.J. Fadili and J.-L. Starck. "Monotone operator splitting for fast sparse solutions of inverse problems." *SIAM Journal on Imaging Sciences*, 2009. submitted.

[FWV07] D. François, V. Wertz, and M. Verleysen. "The Concentration of Fractional Distances." *IEEE Trans. Know. Data. Eng.*, pp. 873–886, 2007.

[Hoe63] W. Hoeffding. "Probability inequalities for sums of bounded random variables." *Journal of the American Statistical Association*, **58**(301):13–30, 1963.

[LDP07] M. Lustig, D. Donoho, and J.M. Pauly. "Sparse MRI: The application of compressed sensing for rapid MR imaging." *MAGNETIC RESONANCE IN MEDICINE*, **58**(6):1182, 2007.

[Led01] M. Ledoux. "The concentration of measure phenomenon, volume 89 of Mathematical Surveys and Monographs." *American Mathematical Society, Providence, RI*, **208**, 2001.

[LT91] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes.* Springer, 1991.

[Mor62] J.J. Moreau. "Fonctions convexes duales et points proximaux dans un espace hilbertien." *CR Acad. Sci. Paris Ser. A Math*, **255**:2897–2899, 1962.

[PG08] Boufounos P. and Baraniuk R. G. "1-Bit Compressive Sensing." In *42nd annual Conference on Information Sciences and Systems (CISS)*, pp. 19–21, Princeton, NJ, March 2008.

[RSV08] H. Rauhut, K. Schnass, and P. Vandergheynst. "Compressed Sensing and Redundant Dictionaries." *Information Theory, IEEE Transactions on*, **54**(5):2210–2219, 2008.

[Spi71] R. Spira. "Calculation of the gamma function by Stirling's formula." *Math. Comp*, **25**(114):317–322, 1971.

[TV94] N.T. 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]*, **42**(3):519–531, 1994.

[VA89] M.K. Varanasi and B. Aazhang. "Parametric generalized Gaussian density estimation." *The Journal of the Acoustical Society of America*, **86**:1404–1415, 1989.

[WAB06] P. Weiss, G. Aubert, and L. Blanc-Féraud. "Some Applications of $\ell_\infty$-Constraints in Image Processing." *INRIA Research Report*, **6115**, 2006.

[WBA08] 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*, pp. 1173–1176, 2008.

[Xu91] H. K. Xu. "Inequalities in Banach spaces with applications." *Nonlinear analysis*, **16**(12):1127–1138, 1991.

[YZ09] L. Ying and Y.M. Zou. "Linear Transformations and Restricted Isometry Property." *preprint*, 2009. arXiv:0901.0541.

[Zwi03] D. Zwillinger. *CRC Standard Mathematical Tables and Formulae.* Chapman & Hall/CRC, 2003.