

DEQUANTIZING COMPRESSED SENSING WITH NON-GAUSSIAN CONSTRAINTS

L. Jacques^{1,2}, D. K. Hammond¹, M. J. Fadili³

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

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

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

ABSTRACT

In this paper, following the Compressed Sensing (CS) paradigm, 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 while enforcing a data fidelity term of bounded ℓ_p -norm, for $2 < p \leq \infty$.

We show that in oversampled situations, i.e. when the number of measurements is higher than the minimal value required by CS, the performance of the BPDQ $_p$ decoders outperforms that of BPDN, with reconstruction error due to quantization divided by $\sqrt{p+1}$. This reduction relies on a modified Restricted Isometry Property of the sensing matrix expressed in the ℓ_p -norm (RIP $_p$); a property satisfied by Gaussian random matrices with high probability. We conclude with numerical experiments comparing BPDQ $_p$ and BPDN for signal and image reconstruction problems.

Index Terms— Compressed Sensing, Quantization, Sampling, Uniform noise, Convex Optimization, Basis Pursuit.

1. INTRODUCTION

The theory of Compressed Sensing (CS) [1] enables reconstruction of sparse or compressible signals from a small number of linear measurements, relative to the dimension of the signal space. In this setting, knowledge of a signal $x \in \mathbb{R}^N$ is contained in the $m \leq N$ linear measurements provided by a sensing matrix $\Phi \in \mathbb{R}^{m \times N}$, i.e. we know only the m inner products $\langle \varphi_i, x \rangle$; where $(\varphi_i)_{i=0}^{m-1}$ are the rows of Φ .

In any realistic digital acquisition system, these analog measurements must be quantized before they may be stored or transmitted. The study of signal recovery from quantized measurements is thus of fundamental interest. In this paper we are interested in the noiseless and uniformly quantized

sensing (or coding) model:

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

where $y_q \in (\alpha\mathbb{N} + \frac{\alpha}{2})^m$ 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 α , 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.

A CS reconstruction program (or decoder) relies on the assumption that the sensed signal x is sparse or compressible in an orthogonal (or redundant [2]) basis $\Psi \in \mathbb{R}^{N \times N}$, i.e. the best K -term approximation x_K in Ψ is an exact or accurate representation of this signal even for small $K < N$. For simplicity, only the canonical basis $\Psi = \text{Id}$ is considered here.

The commonly used Basis Pursuit algorithm for CS recovery finds the sparsest signal (in ℓ_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 whose (unquantized) measurements reproduce the observed quantized values! This problem may be resolved by relaxing the data fidelity constraint. Using a quadratic constraint yields the standard Basis Pursuit DeNoise (BPDN) program [3]:

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

where $\|\cdot\|_2$ is the ℓ_2 -norm. The value of ϵ depends on the magnitude of quantization distortion, and should be chosen just large enough to ensure that the measurements of the original true signal satisfy the data fidelity constraint. In [3], an estimator of ϵ is obtained by considering n as a uniform random vector X with $X_i \sim_{\text{iid}} U([-\frac{\alpha}{2}, \frac{\alpha}{2}])$, i.e. $\epsilon^2 = E[\|X\|_2^2] + \kappa \sqrt{\text{Var}[\|X\|_2^2]} = \frac{\alpha^2}{12} m + \kappa \frac{\alpha^2}{6\sqrt{5}} m^{\frac{1}{2}}$. In that case, $u = x$ respects the constraint of BPDN with a probability higher than $1 - e^{-c_0 \kappa^2}$ for a certain constant $c_0 > 0$.

The stability of BPDN is guaranteed if the sensing matrix $\Phi \in \mathbb{R}^{m \times N}$ satisfies the Restricted Isometry Property (RIP) of order K and radius $\delta \in (0, 1)$, i.e. if there exists a constant μ such that $\mu \sqrt{1 - \delta} \|u\|_2 \leq \|\Phi u\|_2 \leq \mu \sqrt{1 + \delta} \|u\|_2$, for

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

all K -sparse signals $u \in \mathbb{R}^N$. Generally, CS is described with normalized matrices $\bar{\Phi} = \Phi/\mu$ having unit-norm columns (in expectation) so that μ is absorbed in the normalizing constant.

Interestingly, Standard Gaussian Random (SGR) matrices, i.e. with entries drawn from $\Phi_{ij} \sim_{\text{iid}} N(0, 1)$, satisfy the RIP with a controllable high probability (with $\mu = \sqrt{m}$), as soon as $m \geq O(K \log N/K)$ [3]. Moreover, other random constructions satisfying the RIP exist (e.g. Bernoulli matrix, Fourier ensemble, etc.) [1, 3].

For completeness, we include the following theorem expressing the aforementioned stability result, i.e. the $\ell_2 - \ell_1$ instance optimality [4] of BPDN.

Theorem 1 ([5]). *Let $x \in \mathbb{R}^N$ be a compressible 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 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*

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

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 $B < 8.5$.

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 *Quantization Consistency* (QC), i.e. $Q_\alpha[\Phi x^*] = y_q$. This will be met iff $\|y_q - \Phi x^*\|_\infty \leq \frac{\alpha}{2}$, which is not necessarily implied by the BPDN ℓ_2 fidelity constraint. 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 distortion¹

Recently, a few works have focused on this problem. In [6], 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 [7], an adaptation of both BPDN and the Subspace Pursuit integrates the QC constraint explicitly. In [8], sparse signal estimation from the quantization of noisy measurements is realized with a ℓ_1 -regularized maximum likelihood optimization. However, despite interesting experimental results, no theoretical guarantees are given about the approximation error reached by these solutions. In oversampled ADC conversion of signal [9] and in image restoration problems [10], dequantization obtained from

global optimization with equivalent QC constraint expressed in ℓ_∞ -norm can also be found.

This paper, linked to the companion technical report [11], is structured as follows. In Section 2, we present a new class of abstract decoders, coined Basis Pursuit DeQuantizer of moment p (BPDQ $_p$), that model the quantization distortion more faithfully. Section 3 introduces a modified Restricted Isometry Property (RIP $_p$) expressed in the ℓ_p -norm. With this tool, we then prove then the stability of the BPDQ $_p$ programs, i.e. their $\ell_2 - \ell_1$ instance optimality. In Section 4, we show that, given a sufficient number of measurements, the approximation error due to quantization scales inversely with $\sqrt{p+1}$. Finally, Section 5 reports numerical simulations on signal and image reconstruction problems.

2. BASIS PURSUIT DEQUANTIZERS

We introduce a new class of optimization programs (or decoders) that generalize the fidelity term of the BPDN program to noises that follow a centered Generalized Gaussian Distribution (GGD) of *shape parameter* $p \geq 1$ [11], with the uniform noise case corresponding to $p \rightarrow \infty$. These decoders reconstruct an approximation of the sparse or compressible signal x from its distorted measurements $y = \Phi x + n$ when the distortion has a bounded p^{th} moment, i.e. $\|n\|_p \leq \epsilon$. Formally, the decoder is

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

We dub this class of decoders *Basis Pursuit DeQuantizer of moment p* (or BPDQ $_p$) since, as shown in Section 4, their approximation error when Φx is uniformly quantized decreases as both the moment p and the oversampling factor m/K increase.

3. RIP $_p$ AND $\ell_2 - \ell_1$ INSTANCE OPTIMALITY

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

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

$$\mu_p \sqrt{1-\delta} \|x\|_2 \leq \|\Phi x\|_p \leq \mu_p \sqrt{1+\delta} \|x\|_2, \quad (3)$$

for all $x \in \mathbb{R}^N$ with $\|x\|_0 \leq K$, and where $\|\cdot\|_p$ is the ℓ_p -norm on \mathbb{R}^m .

The common RIP previously introduced is thus the RIP $_2$. Interestingly, SGR matrices $\Phi \in \mathbb{R}^{m \times N}$ satisfy also the RIP $_p$ of order K and radius $0 < \delta < 1$ with high probability provided that $m \geq O((\delta^{-2} K \log N/K)^{p/2})$ for $2 \leq p < \infty$, or $\log m \geq O(\delta^{-2} K \log N/K)$ for $p = \infty$; see [11] for details. Moreover, for these matrices an asymptotic (in m) approximation for μ_p is

$$\sqrt{2} \pi^{-\frac{1}{2p}} \Gamma\left[\frac{p+1}{2}\right]^{\frac{1}{p}} m^{\frac{1}{p}} \quad (4)$$

¹We do not adapt here other decoders, such as the Dantzig selector.

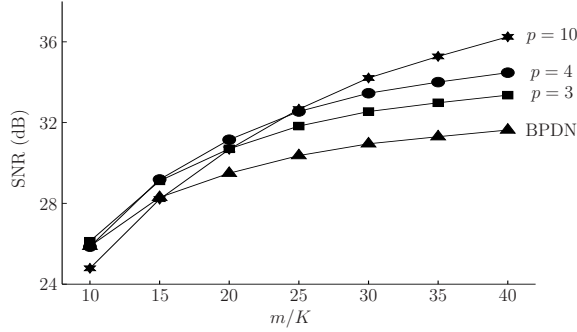


Fig. 1: Quality of BPDQ_p for different m/K and p .

if $2 \leq p < \infty$ and $\mu_\infty \geq \rho^{-1} \sqrt{\log m}$ for a certain $\rho > 0$ [11]. This results from the specialization to SGR vectors of a study made in [12]. We are now ready to state our main result.

Theorem 2. *Let $x \in \mathbb{R}^N$ be a compressible 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 matrix on s sparse signals with radius δ_s , for $s \in \{K, 2K, 3K\}$ and $2 \leq p < \infty$. Given a measurement vector $y = \Phi x + n$ with $\|n\|_p \leq \epsilon$, the solution $x_p^* = \Delta_p(y, \epsilon)$ obeys the $\ell_2 - \ell_1$ instance optimality*

$$\|x_p^* - x\|_2 \leq A_p e_0(K) + B_p \frac{\epsilon}{\mu_p}, \quad (5)$$

for $A_p = \frac{2(1+C_p-\delta_{2K})}{1-\delta_{2K}-C_p}$, $B_p = \frac{4\sqrt{1+\delta_{2K}}}{1-\delta_{2K}-C_p}$, and with $C_p = C_p(\delta_K, \delta_{2K}, \delta_{3K})$ well behaved.

The approximation error reached by BPDQ is thus bounded in (5) by the sum of the *compressibility* and the *noise errors*. As shown in [11], this theorem uses explicitly the 2-smoothness of the Banach spaces ℓ_p when $2 \leq p < \infty$ [13] and their possible embedding in ℓ_2 . The value C_p behaves as $\sqrt{(\delta_K + \delta_{3K})(1 + \delta_{2K})} p$ for large p , and as $\delta_{3K} + \frac{3}{4}(1 + \delta_{3K})(p - 2)$ for $p \simeq 2$. For $p = 2$, Theorem 2 reduces to Theorem 1 if $\delta_{3K} = \sqrt{2} \delta_{2K}$.

4. QUANTIZATION ERROR REDUCTION

We now turn to the behavior of the BPDN_p decoders on quantized measurements of sparse or compressible signals. First, assuming the quantization distortion $n = Q_\alpha[\Phi x] - \Phi x$ is uniformly distributed in each quantization bin, we proved in [11] that for

$$\epsilon = \epsilon_p(\alpha) \triangleq \frac{\alpha}{2^{(p+1)^{1/p}}} (m + \kappa(p+1) \sqrt{m})^{\frac{1}{p}}, \quad (6)$$

$u = x$ is a solution of the BPDQ_p fidelity constraint with a probability at least $1 - e^{-2\kappa^2}$.

Second, by Theorem 2, when Φ is RIP_p with $2 \leq p < \infty$, i.e. when for $m \geq O((\delta^{-2} K \log N/K)^{p/2})$ for SGR matrices, we have

$$\|x - x_p^*\|_2 \leq A_p e_0(K) + B_p \frac{\epsilon_p(\alpha)}{\mu_p}. \quad (7)$$

Third, from (4), we have the approximation $\mu_p \simeq c_p m^{\frac{1}{p}}$ with $c_p = \sqrt{2} \pi^{-\frac{1}{2p}} \Gamma[\frac{p+1}{2}]^{\frac{1}{p}} \geq 2^{-\frac{1}{2}} e^{-\frac{3}{4}} \sqrt{p+1} (1 + O(p^{-2}))$, using Stirling formula $\Gamma(z) = (\frac{2\pi}{z})^{\frac{1}{2}} (\frac{z}{e})^z (1 + O(\frac{1}{z}))$.

Finally, bounding the different functions involved yields

$$\frac{\epsilon_p(\alpha)}{\mu_p} \lesssim C \frac{\alpha}{\sqrt{p+1}} (1 + O(p^{-2})), \quad C < 1.497. \quad (8)$$

In short, the noise error term in the $\ell_2 - \ell_1$ instance optimality relation (7) for the quantized model (1) is thus divided by $\sqrt{p+1}$ if the sensing matrix Φ satisfies the RIP_p!

More precisely, with a philosophy close to the oversampled ADC conversion [9], *this error noise reduction happens in oversampled sensing*, i.e. when the *oversampling factor* m/K is high. Indeed, in that case a SGR matrix Φ satisfies the RIP_p with high probability for high p . Moreover, oversampling gives a smaller δ , 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 δ also favors BPDN, but since the value $A = A_2$ and $B = B_2$ in (2) are bounded from below this effect is limited. This is confirmed experimentally in Section 5.

Finally, note that the necessity of satisfying RIP_p implies that we cannot directly set $p = \infty$ in BPDQ_p to impose Quantization Consistency (QC) of this decoder². Indeed, for a given oversampling factor m/K , a SGR matrix Φ can be RIP_p only over a finite interval $p \in [2, p_{\max}]$.

5. EXPERIMENTAL RESULTS

The BPDQ_p decoders³ are solved in practice by monotone operator splitting and proximal methods [14, 11]. More precisely, as both the ℓ_1 -norm and the indicator function of the constraint in BPDQ_p are non-differentiable, the Douglas-Rachford splitting is used. The Douglas-Rachford recursion to solve BPDQ_p can be written in the compact form

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

where $\alpha_t \in (0, 2), \forall t \in \mathbb{N}, \gamma > 0, S_\gamma$ is the component-wise soft-thresholding operator with threshold γ and $\mathcal{P}_{T_p(\epsilon)}$ is the orthogonal projection onto the closed convex constraint set $T_p(\epsilon) = \{u \in \mathbb{R}^N : \|y_q - \Phi u\|_p \leq \epsilon\}$. From [14], 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. The projection $\mathcal{P}_{T_p(\epsilon)}$ was computed iteratively using Newton's method to solve the Lagrange multiplier equations arising from minimizing the distance to the constraint set.

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 $\{1, \dots, N\}$. The non-zero elements have been drawn from a standard Gaussian distribution $N(0, 1)$. For each sparse

²Observing also that $\lim_{p \rightarrow \infty} \epsilon_p(\alpha) = \alpha/2$.

³Our code is freely available on <http://wiki.epfl.ch/bpdq>.

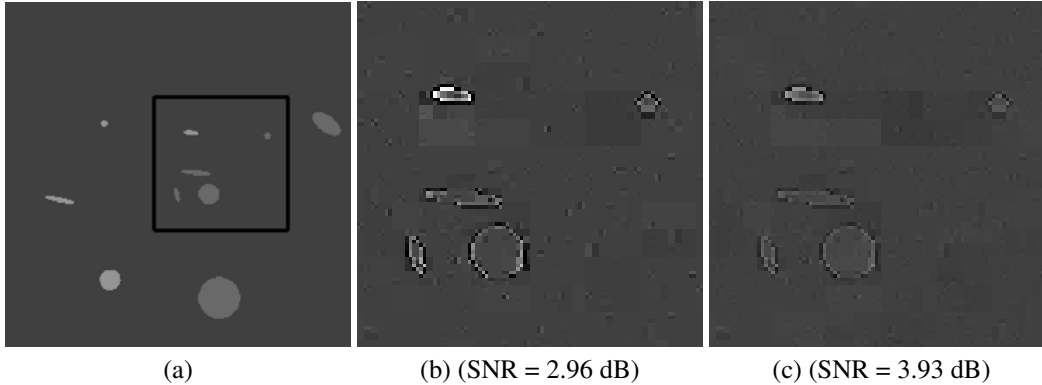


Fig. 2: Reconstruction from quantized undersampled Fourier measurements. (a) Original; details from (b) BPDN; and (c) BPDQ₁₀

signal, m quantized measurements have been recorded as in model (1) with a SGR matrix $\Phi \in \mathbb{R}^{m \times N}$. The bin width has been set to $\alpha = \|\Phi x\|_\infty / 40$. In Figure 1, we plot the average quality of the reconstructions of 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 undersampled MRI reconstruction problem. Using an example similar to [15], 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 α 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 Ψ 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 2, 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 Φ satisfies the RIP _{p} (3) in this case. 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.

6. CONCLUSION

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 theoret-

ically and experimentally that BPDQ _{p} exhibit a substantial reduction of the approximation error in oversampled situations. An interesting perspective is to characterize the evolution of the optimal moment p with the oversampling ratio. This could allow the selection of the best BPDQ decoder as a function of the precise CS coding/decoding scenario.

7. REFERENCES

- [1] E.J. Candès and J. Romberg, “Quantitative Robust Uncertainty Principles and Optimally Sparse Decompositions,” *Found. Comp. Math.*, **6**(2):227–254, 2006.
- [2] H. Rauhut, K. Schnass, and P. Vandergheynst, “Compressed sensing and redundant dictionaries,” *IEEE T. Inform. Theory.*, **54**(5):2210–2219, 2007.
- [3] E. Candès, J. Romberg, and T. Tao, “Stable signal recovery from incomplete and inaccurate measurements,” *Comm. Pure Appl. Math.*, **59**(8):1207–1223, 2006.
- [4] A. Cohen, R. DeVore, and W. Dahmen, “Compressed sensing and best k-term approximation,” *J. Amer. Math. Soc.*, **22**:211–231, 2009.
- [5] 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.
- [6] P. Boufounos and R.G. Baraniuk, “1-bit compressive sensing,” in *42nd Conf. Inf. Sc. Systems (CISS)*, Princeton, NJ, March 2008, 19–21.
- [7] W. Dai, H. Vinh Pham, and O. Milenkovic, “Distortion-rate functions for quantized compressive sensing,” *preprint*, 2009, arXiv:0901.0749.
- [8] A. Zymnis, S. Boyd and E. Candès, “Compressed sensing with quantized measurements,” Submitted work, 2009.
- [9] N.T. Thao and M. Vetterli, “Deterministic analysis of oversampled A/D conversion and decoding improvement based on consistent estimates,” *IEEE T. Sig. Proc.*, **42**(3):519–531, 1994.
- [10] P. Weiss, L. Blanc-Feraud, T. Andre, and M. Antonini “Compression artifacts reduction using variational methods: Algorithms and experimental study Acoustics,” *IEEE ICASSP 2008*, 1173–1176.
- [11] L. Jacques, D.K. Hammond, and M.J. Fadili, “Dequantizing compressed sensing: when oversampling and non-Gaussian constraints combine” <http://arxiv.org/abs/0902.2367>, 2009,
- [12] D. François, V. Wertz, and M. Verleysen, “The Concentration of Fractional Distances,” *IEEE T. Know. Data. Eng.*, 873–886, 2007.
- [13] W.L. Bynum, “Weak parallelogram laws for Banach spaces,” *Canad. Math. Bull.*, **19**(3):269–275, 1976.
- [14] P.L. Combettes, “Solving monotone inclusions via compositions of nonexpansive averaged operators,” *Optimization*, **53**:475–504, 2004.
- [15] M. Lustig, D. Donoho and J. M. Pauly, “Sparse MRI: The Application of Compressed Sensing for Rapid MR Imaging,” *Magnetic Resonance in Medicine*, **58**:1182–1195, 2007.