Weighted Fidelity in Non-Uniformly Quantized Compressed Sensing
Laurent Jacques, David Hammond, Jalal M. Fadili

To cite this version:

HAL Id: hal-00808632
https://hal.archives-ouvertes.fr/hal-00808632
Submitted on 5 Apr 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
WEIGHTED FIDELITY IN NON-UNIFORMLY QUANTIZED COMPRESSED SENSING

Laurent Jacques\(^1\), David Kenrik Hammond\(^2\), Jalal Fadili\(^3\)

\(^1\)ICTEAM/ELEN, Université catholique de Louvain (UCL), Belgium. \(^2\)Neuroinformatics Center, University of Oregon, USA.
\(^3\)GREYC CNRS-ENSICAEN-Université de Caen, 14050 Caen France

ABSTRACT

Following the Compressed Sensing (CS) paradigm, this paper studies the problem of recovering sparse or compressible signals from (scalar) non-uniformly quantized measurements. We show that a simple adaptation of the Basis Pursuit De-Quantizer introduced earlier, that is, a sign sensitive weighting of their \(\ell_p\)-norm fidelity constraint, yields good SNR improvements in the signal reconstruction. As a good indication of this improvement origin, we prove theoretically that a similar decoder, using a particular side-position-to-level oracle, displays a reduction of the reconstruction error when both the number of measurements and the moment \(p\) of the constraint increase. This follows the oversampling principle underlined in our previous study for uniformly quantized CS, with an additional gain provided by the non-uniform quantization. We conclude this paper by showing the efficiency of the approach on 1-D and 2-D signal examples.

1. INTRODUCTION

The recent theory of Compressed Sensing (CS) \([1, 2]\) shows how sparse or compressible signals can be reconstructed from few linear measurements compared to the dimension \(N\) of the signal space. The gist of this approach relies in the use of a sensing basis sufficiently incoherent with the sparsity basis of the signal. This happens with high probability for a large class of random matrix constructions as soon as the number of measurements \(M\) is higher than "few multiples" of the signal sparsity \(K\). For instance, for Random Gaussian matrices, \(M = O(K \log N/K)\).

Similarly to recent studies \([3–5]\) in the CS literature, this work is interested in controlling the signal reconstruction stability when the compressive measurements undergo a scalar quantization, possibly non-uniform, of given rate \(R\).

More precisely, given a signal \(x \in \mathbb{R}^N\), we first assume it to be sparse, or sparsely approximable (compressible), in a certain orthogonal basis \(\Psi = (\Psi_1, \cdots, \Psi_N) \in \mathbb{R}^{N \times N}\) (e.g., in the wavelet basis or in the pixel domain). In other words, this signal is decomposed as \(x = \Psi c = \sum_j \Psi_j c_j\) with an approximation error \(\|c - c_K\| (c_K\) being the best \(K\) term approximation of \(c\)) quickly decreasing when the integer \(K\) increases. For the sake of simplicity, the sparsity basis is assumed to be canonical \((\Psi = I)\) and \(c\) is identified with \(x\). All the results can be easily extended to the situation \(\Psi \neq I\).

Second, we are interested in the Compressed Sensing of \(x \in \mathbb{R}^N\) with a certain sensing matrix \(\Phi \in \mathbb{R}^{M \times N}\) \([1, 2]\). Each compressed sensing measurement, i.e., each component of the measurement vector \(\Phi x\), undergoes a general scalar (uniform or non-uniform) quantization \(Q\) described in Section 2, i.e., our sensing model is

\[ y = Q[\Phi x]. \]  

Conventions: All space dimensions are denoted by capital letters \((e.g., K, M, N, D \in \mathbb{N})\), vectors and matrices are written in bold symbols. For any vector \(u = (u_1, \cdots, u_D)^T \in \mathbb{R}^D\) (with \((\cdot)^T\) the transposition), the \(\ell_p\)-norm \((p \geq 1)\) of \(u\) is \(\|u\|_p = \sum_i |u_i|^p\), with \(\|u\|_2 = \|u\|\) and \(\|u\|_0 = \#\{i : u_i \neq 0\}\) the \(\ell_0\) ("not-a") norm of \(u\). We denote also \(1 = (1, \cdots, 1)^T \in \mathbb{R}^D\), and \(I\) for the identity matrix. Given a vector \(u\), \(U = \text{diag } u\) is the diagonal matrix such that \(U_{ii} = u_i\).

2. QUANTIZATION FRAMEWORK

Our operator \(Q\) of interest here is a scalar quantizer of vector components. We do not impose this quantization to be uniform, that is, the quantization bin width is not necessarily constant with respect to the bin level (as in Fig. 1-left).

More precisely, \(Q\) relies on the definition of a set of \(B = 2^R\) levels \(\omega_k\) (coded on \(R = \log_2 B\) bits) and of a set of \(B+1\) thresholds \(t_k \in \mathbb{R} \cup \{\pm \infty\} = \mathbb{R}\), with \(\omega_k < \omega_{k+1}\) and \(t_k \leq \omega_k < t_{k+1}\) for all \(1 \leq k \leq B\) and given a rate \(R \in \mathbb{N}\). The \(k\)th quantizer bin (or region) is \(\mathcal{R}_k = [t_k, t_{k+1})\). As illustrated on Fig. 1-left, the quantizer is a mapping between \(\mathbb{R}\) and the set of levels \(\Omega = \{\omega_k : 1 \leq k \leq B\}\), i.e., \(Q[\lambda] = \omega_k \Leftrightarrow \lambda \in \mathcal{R}_k = Q^{-1}[\omega_k]\). For vectors \(u \in \mathbb{R}^M\), \(Q[u] \in \Omega^M\) with \(Q[u]_k = Q[\omega_k]\) in \(\Omega\), for \(1 \leq k \leq M\).

Generally, whatever the definition of the quantization levels and thresholds, many scalar quantization schemes relies on a High Resolution/Rate Assumption (HRA) \([6]\). They assume that the distortion noise, that is, the difference between the initial and the quantized values, is uniform within each quantization bins, possibly with different widths. Therefore, the creation of the non-uniform quantizer, which can rely on a non-uniformity assumption (like with Lloyd-Max itera-
3. ASYMMETRIC QUANTIZATION CONSISTENCY

Beyond any stochastic modeling of the quantization distortion, we know that $Q[t] = \omega_i$ is equivalent to the condition $t \in \mathcal{N}_i = Q^{-1}[\omega_i] = [t_i, t_{i+1})$. In other words,

$$Q[t] = \omega_i \Leftrightarrow \begin{cases} t - \omega_i < t_{i+1} - \omega_i & \text{if } t \geq \omega_i \\ t - \omega_i \geq t_{i+1} - \omega_i & \text{if } t < \omega_i \end{cases}.$$ \hspace{1cm} (2)

This condition can be simplified thanks to the following asymmetric norm $\| \cdot \|_{\infty,s}$. Let $S = \text{diag}(s) \in \mathbb{R}_+^{D \times 2D}$ be a diagonal matrix of positive diagonal entries $s = (s^+, s^-)^T \in \mathbb{R}_+^{2D}$, with $s^\pm \in \mathbb{R}^D$. The sign-dependent weighted $\ell_p$ norm is given by

$$\|u\|_{p,s} = \left\| S \begin{pmatrix} (u^+) \\ -(u^-) \end{pmatrix} \right\|_p, \quad p \in [1, \infty],$$ \hspace{1cm} (3)

where $(u^+) = (u_i^+) = \max(u_i, 0)$ for $p = 1$, $\| \cdot \|_{\infty,s}$ is an asymmetric norm but not a norm\footnote{There is no positive homogeneity; $[\lambda u]_{p,s} \neq |\lambda| [u]_{p,s}$ for $\lambda \in \mathbb{R}$.}. Moreover, for $p \geq 1$, $\| \cdot \|_{\infty,s}$ is convex.

Let us define $\Delta_i^+ = t_{i+1} - \omega_i$, $\Delta_i^- = \omega_i - t_i$, and $k(y_i)$, the quantization bin index of the $i$th component of $y$, that is, such that $y_i \in \mathcal{N}_k(y_i)$. Our asymmetric weighting matrix is $S(y) = \text{diag}(s(y)) \in \mathbb{R}^{2M \times 2M}$, with $s = s(y) = (s^+(y), s^-(y))^T$ and $s_i^\pm(y) = 1/\Delta_i^\pm(k(y_i))$. Therefore, from (2),

$$Q[\Phi x] = y \Leftrightarrow \|\Phi x - y\|_{\infty,s} \leq 1.$$ \hspace{1cm} (QC)

This is the Quantization Consistency (QC) constraint that any reconstructed vector $x^*$ should respect in order to be consistent with the quantized measurement $y$. Notice that keeping the same $S$, $\lim_{p \to \infty} \|\Phi x - y\|_{p,s} = \|\Phi x - y\|_{\infty,s} \leq 1$.

4. RECONSTRUCTION AND WEIGHTED FIDELITY

There is a straightforward way for estimating a signal $x$ sensed through the model (1). We can indeed alter the Basis Pursuit DeQuantizer introduced in [5] by replacing their $\ell_p$-norm fidelity constraint by the weighted asymmetric norm introduced in (3), that is,

$$\arg \min_{u \in \mathbb{R}^N} \|u\|_1 \text{ s.t. } \|\Phi u - y\|_{p,s} \leq \epsilon, \quad \text{(WBPDQ)}$$

with $s$ defined as in Sec. 3. Notice that for $p = 2$ and $s^\pm = 1$, WBPDQ reduces to the Basis Pursuit DeNoiser (BPDN) [1, 2]. As for the BPDQ decoders, we are going to discover which moment $p$ minimizes the reconstruction error; and the answer is not necessarily $p = \infty$ despite the QC relation!

Unfortunately, in spite of good numerical results (see Sec. 7), we did not find a convincing way for bounding theoretically the WBPDQ approximation error. We were instead able to characterize the behavior of the following brother program helped with the side-position-to-level (SPTL) oracle $\sigma = \text{sign}(\Phi x - y) \in \{\pm 1\}^M$:

$$x^* = \arg \min_{u \in \mathbb{R}^N} \|u\|_1 \text{ s.t. } \begin{pmatrix} \|\Phi u - y\|_{p,w} \leq \epsilon \\ \text{sign}(\Phi u - y) = \sigma \end{pmatrix}.$$ \hspace{1cm} (4)

with $\|\cdot\|_{p,w} \equiv \|\text{diag}(w) \cdot \|_{p,\ell} = w(s, \sigma) \in \mathbb{R}^M$ being the weighting vector such that $w_i = s_i^+(\sigma_i) + s_i^-(-\sigma_i)$.

Similarly to what we discovered in [5], the stability of (4) depends on the good behavior of $\Phi$ in the normed space $\ell^{p,w}_\Phi = (\mathbb{R}^M, \| \cdot \|_{p,\ell})$. In particular, we say that, given a weight vector $w' \in \mathbb{R}^M$, a matrix $\Phi \in \mathbb{R}^{M \times N}$ satisfies the Restricted Isometry Property from $\ell^{p,w}_\Phi$ to $\ell_2$ at order $K \in \mathbb{N}$, radius $0 \leq \delta < 1$ and for a normalization $\mu = \mu(p, M, N) > 0$, if for all $x \in \Sigma_K = \{ u \in \mathbb{R}^N : \|u\|_0 \leq K \}$,

$$(1 - \delta)^{1/2} \|x\| \leq \frac{1}{\mu} \|\Phi x\|_{p,w'} \leq (1 + \delta)^{1/2} \|x\|.$$ \hspace{1cm} (5)

We will write shortly that $\Phi$ is RIP$_{p,w',\ell_2} = (K, \delta, \mu)$. Of course, the common RIP and the RIP$_{p,2}$ [5] are obtained with $w' = 1$ and $p = 2$. We found that a small modification of the Proposition 1 in [5] allows us to check that, with very high (controllable) probability, a Standard Gaussian Random (SGR) matrix $\Phi \in \mathbb{R}^{M \times N}$ with $\Phi_{ij} \sim N(0,1)$ is RIP$_{p,w',\ell_2} = (K, \delta, \mu)$ as soon as

$$M^{2/p} = O(K \log N/K) \quad \text{and} \quad \mu = \mathbb{E}[\|x\|_{p,w'}],$$

for $x \in \mathbb{R}^M$ a SGR vector.

Obviously, if $\Phi$ is RIP$_{p,w',\ell_2}$, then $\Phi' = \text{diag}(w) \Phi$ is RIP$_{p,\ell_2} = (K, \delta, \mu)$, while $\|\Phi u - y\|_{p,w} = \|\Phi' u - \text{diag}(w) y\|_{p,w}$.

Therefore, despite the complementary constraint on the sign $\sigma$, the stability proved in [5] holds for (4).

**Theorem 1.** Let $x \in \mathbb{R}^N$ be a signal with a $K$-term $\ell_1$-approximation error $e_0(K) = K^{-\frac{1}{2}} \|x - x_K\|_1$, for $0 \leq K \leq N$. Let $\Phi$ be a RIP$_{p,w',\ell_2} = (s, \delta, \mu)$ matrix for $s \in \{K, 2K, 3K\}$ and $2 \leq p < \infty$. If $x$ is a feasible point of the constraints in (4), then

$$\|x^* - x\|_2 \leq A_p e_0(K) + B_p \mu^{-1} \epsilon,$$ \hspace{1cm} (6)

for values $A_p(K) = \frac{2(1+C_p - \delta_{2K})}{1-2\delta_{2K} - C_p}$, $B_p(K) = \frac{4\sqrt{1+2\delta_{2K}}}{1-2\delta_{2K} - C_p}$, $C_p = O\left(\sqrt{\delta_{2K} + \delta_{3K}(p-2)}\right)$ as $p \gg 2$ and $C_p = \delta_{3K} + O(p-2)$ as $p \to 2$.$^2$

$^2$The precise definition of $C_p$ is given in [5].
OVERSAMPLING EFFECT

An interesting effect occurs in (6) when the sensing is over-
sampled, that is, when $M/K$ is large enough to allow us to
select high moment $p$ in (4) and still have the RIP $\ell_p^M, \ell_p^N$ of
the SGR sensing matrix $\Phi$. Indeed, we are going to show that
the error term $\mu^{-1} \epsilon$ in (6) decreases as $1/\sqrt{p}$ when $p$
increases, and it is further reduced when the weights $w$ are well
adjusted.

We first need to find an estimator for $\epsilon$ in the sensing
model (1). This is done by observing when, with high proba-
bility, $x$ is a feasible point of the constraints in (4). We notice
that $\|\Phi x - y\|_{p,w}^p = \sum_i w_i^p (\sigma_i) \|\Phi x_i - y_i\|^p$. From the
AHRA, each $w_i(\sigma_i) (\|\Phi x_i - y_i\|)$ can be modeled as a uni-
form random variable on $[0,1]$. We can then take $\alpha = 2$ in
Lemma 3 of [5] and determine that

$$\|\Phi x - y\|_{p,w} \leq \epsilon \leq \epsilon_p(M) = (\sqrt{(p+1)} M + \kappa \sqrt{M})^{1/p}$$

holds with probability higher than $1 - e^{-2\kappa^2}$.

Second, we must lower bound the RIP $\ell_p^M, \ell_p^N$ normalization $\mu$. For this purpose, let us assume that the weights $w = G(M) \in \mathbb{R}^M$ have been generated by a particu-
lar weight generator $G(M)$ (for instance as an output of a
Lloyd-Max quantizer) respecting the following property.

Definition 1. The generator $G$ (and by extension $w = G(M)$) has the Converging Moments (CM) property if, for
any $p \geq 1$, there is a $M_0$ such that

$$\rho_p^{\text{min}} \leq M^{-1/p} \|G(M)\|_p \leq \rho_p^{\text{max}}, \quad \forall M \geq M_0,$$

where $\rho_p^{\text{min}} > 0$ and $\rho_p^{\text{max}} > 0$ are, respectively, the biggest and the smallest values such that (7) holds.

The CM property makes sense for instance if all the weights $\{w_i\}_{1 \leq i \leq M}$ are taken (with repetition) inside a finite set of values (of size independent of $M$).

A simple modification of Lemma 1 in [5] shows that, if
$M \geq 2\beta^{-1} (2/\rho_p^{\text{min}})^p$ and if $w$ is CM, then

$$\mu \geq c \rho_p^{\text{min}} \sqrt{p + 1} \left(1 + \beta\right)^{1/p - 1} M^{1/p},$$

with $c = (8\sqrt{2})/(9\sqrt{7})$. Therefore, using this relation, simple calculations show that, with a probability higher
than $1 - e^{-2\kappa^2}$,

$$\mu^{-1} \epsilon_p(M) \leq C (\rho_p^{\text{min}} \sqrt{p + 1})^{-1}.$$  

with $C < 2.17$ as soon as $M > (\frac{\rho_p^{\text{min}}}{\rho_p^{\text{max}}})^2$. Interestingly, this
last inequality reduces to the bound of $\epsilon/\mu$ for the uniform
quantization of bin width $\alpha > 0$ obtained in [5]. Indeed, this
case is equivalent to $w = \frac{2}{\alpha}$, so that $\rho_p^{\text{min}} = \rho_p^{\text{max}} = \frac{2}{\alpha}$. For non-uniform quantization, in addition to the division
of the error by $\sqrt{p + 1}$ due to the oversampled sensing, a new
effect occurs: an error reduction due to $\rho_p^{\text{min}}$. As estimated
in Fig. 1-right with $\rho_p^{\text{min}} \approx M^{-1/p} \|w\|_p$, this value can be
higher for non-uniform quantization than for uniform one.

NUMERICAL IMPLEMENTATION

We have implemented the convex WBPDQ decoder thanks
to the BPDQ toolbox$^3$. Briefly, this toolbox solves the un-
weighted problem $(s^\pm = 1)$ where the optimization con-
straint reduces to a simple $\ell_1$ norm. It proceeds by using
the Douglas Rachford splitting [9]. This iterative methods
combines the proximal operator of the $\ell_1$ norm (a soft-
thresholding) and the orthogonal projector onto a $\ell_1$ ball
of radius $\epsilon$ centered at the measurement vector $y$. We have
adapted this toolbox in order to solve WBPDQ by modifying
this last operation with a projection on the ball associated
to the asymmetric norm in (3). As described in [5], we used
for that a Newton’s method solving the related KKT system
which is reminiscent of the Sequential Quadratic Program-
ing (SQP) method.

7. EXPERIMENTS

Given the quantized sensing model (1), the purpose of
this section is to observe how the WBPDQ reconstruction
quality evolves for different set of parameters $N$, $K$, $M$, $R$ and moments $p$. In all our experiments, a spread-
spectrum sensing matrix has been used [10]. More precisely,
$\Phi = RF \text{diag}(h)$, where $R \in \mathbb{R}^M \times N$ is a restriction ma-
trix picking $M$ values uniformly at random in $\{1, \cdots, N\}$,$F \in \mathbb{R}^N \times N$ is the DCT basis, and $h \in \mathbb{R}^N$ is a random $\pm 1$
Bernoulli sequence. This sensing matrix is not proved to be
as optimal as the SGR matrix (or Random Gaussian Ensemble).
However, we observed that the measurement vector $\Phi x$
still follows a Gaussian distribution, and the multiplication
$\Phi u$ or $\Phi^T v$ is very fast ($O(N \log N)$). Moreover, it is a
tight frame with $\Phi \Phi^T = I$ which makes the Douglas Rach-
ford (DR) faster [5]. For all experiments, 300 iterations were
sufficient to observe a convergence of the DR algorithm.

Sparse 1-D signals: For this experiment, we have randomly
generated $K$-sparse signals in $\mathbb{R}^N$ by picking uniformly at
random their support in $\{1, \cdots, N\}$, their $K$ non-zero val-
ues following a Normal distribution. Each signal has been

$^3$http://wiki.epfl.ch/bpdq
sensed with the sensing model (1) according to two quantization scenario. An optimal uniform quantizer and an non-uniform Lloyd-Max quantizer both tuned to inputs with Gaussian distributions. For the first quantizer, the Basis Pursuit Denoiser (BPDN) has reconstructed the signals, while for the second, the WBPDQ decoder has been selected with weights computed as in Sec. 4. The parameter $N$, $K$ and $R$ have been set to 1024, 16 and 4 bits respectively, while $M/K \in [8, 64]$ and $p \in [2, 10]$. Fig. 3-left shows the averaged reconstruction quality evolution (in SNR) over 100 trials with respect to $M/K$ for different $p$. For comparison, the BPDN quality curve is shown with a dashed line. It can be clearly observed that as soon as the oversampling factor $M/K$ is sufficiently high, taking a $p$ higher than 2 yields significant improvement with up to 4 dB at the highest oversampling for $p = 10$. However, Fig. 3-right, which displays the averaged SNR gain relatively to WBPDQ with $p = 2$, confirms that the moment $p$ leading to the highest gain for a given ratio oversampling ratio increases with $M/K$, while the performance decays if higher moments are selected.

**Compressible 2-D image:** In this second experiment, we have challenged the WBPDQ decoder on the reconstruction of a 512 × 512 image ($N = 262,144$) compressible in a Daubechies 7/9 wavelet basis and sensed with (1) for $M = \frac{3}{4} N$ and 4 bits per measurements. The reconstruction results are shown in Fig. 2. Despite a weak visual improvement between the WBPDQ ($p = 6$, on Lloyd-Max quantized measurements) and the BPDN reconstructions (on uniformly quantized measurements), the PSNR gain for WBPDQ is about 0.8 dB. This is not a large value which is mainly due to the compressible nature of the image; the reduction of the quantization error is masked by the importance of the compressible error in (6). However, we measured that the WBPDQ solution satisfies the quantization consistency, while for the BPDN solution, $\|\Phi x^* - y\|_{\infty, (2/\alpha)1} = 1.963$.

8. CONCLUSION

The objective of this work was to show how to improve the reconstruction of sparse or compressible signals sensed through non-uniformly quantized compressive measurements. We have shown that a weighted asymmetric $\ell_p$ norm allows us to incorporate the quantization consistency of the measurements in the signal decoding, that is, in the WBPDQ program. Finally, theoretical indications and numerical experiments confirm the possibility to reduce the quantization impact in the reconstructed signals.

9. REFERENCES