Distortion-Rate Functions for Quantized Compressive Sensing

Wei Dai, Hoa Vinh Pham, and Olgica Milenkovic
Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign

***Abstract*—We study the average distortion introduced by quantizing compressive sensing measurements. Both uniform quantization and non-uniform quantization are considered. The asymptotic distortion-rate functions are obtained when the measurement matrix belongs to certain random ensembles. A new modification of greedy reconstruction algorithm that accommodates quantization errors is proposed and its performance is evaluated through extensive computer simulations.**

I. Introduction

Compressive sensing (CS) has received significant attention due to its wide applications in medical imaging, biosensing, spectrum monitoring and other areas of signal processing. It is a sampling method that converts unknown input signals, embedded in a high dimensional space, into signals that lie in a space of significantly smaller dimension, using linear measurements. In general, it is an ill-posed problem to recover the unknown signal using measurements embedded in a reduced-dimensional space. Nevertheless, if the input signal is sufficiently sparse, exact reconstruction is possible in polynomial time [1], [2], [3], [4], which is the central result of CS theory. As a result, CS significantly reduces the number of measurements required to acquire an unknown sparse signal.

There exist many different techniques for sparse signal reconstruction. For simplicity, assume that the unknown signal $\mathbf{x} \in \mathbb{R}^N$ is K-sparse, i.e., that there are at most K nonzero entries in $\mathbf{x}$. The naive reconstruction method is to search among all possible signals and find the sparsest one which is consistent with the linear measurements. In general, this method requires only $m = 2K$ random linear measurements. Unfortunately, to find the sparsest signal representation is an NP-hard problem. On the other hand, the work by Donoho and Candès et. al. [1], [2], [3], [4] demonstrated that sparse signal reconstruction is a polynomial time problem provided that more measurements are taken. This is achieved by casting the reconstruction problem as a linear programming problem and solving it using the basis pursuit (BP) method. More recently, the authors proposed the so called subspace pursuit (SP) algorithm in [5] (see the independent work [6] for a modification). Its computational complexity is linear in the signal dimension, and the required number of linear measurements is of the same order as that needed for the BP method. For both the BP and SP algorithms, the reconstruction distortion can also be analyzed when the measurements are subjected to noise with bounded power [7], [5].

For most practical applications, it is reasonable to assume that the measurements are quantized and that therefore one does not have infinite precision observations. With certain assumptions of the quantization rate, it has been shown in [8] that uniform scalar quantization provides near-optimal distortion performance. However, the same quantization technique is not efficient for sparse signals in the sense that a large fraction of quantization regions is not used at all [9]. Both of the above approaches focus on the worst case analysis. On the algorithmic side, a reconstruction algorithm for one bit quantization is proposed and discussed in [10].

As opposed to the approach in [8], [9], we consider the average distortion introduced by quantization, and study the effects of both uniform quantization and non-uniform quantization. The asymptotic distortion rate function is characterized when the measurement matrix is taken from certain random matrix ensembles. We also discuss the more general case of CS in which constraints on the column norms of the measurement matrix are imposed. Finally, we adapt two benchmark CS reconstruction algorithms to accommodate quantization errors and show that the new algorithms offer significant performance improvement over classical CS reconstruction techniques.

This paper is organized as follows. Section II contains a brief overview of CS theory, the BP and SP reconstruction algorithms, and various quantization techniques. In Section III, we analyze the CS distortion rate function and examine the influence of quantization errors on the BP and SP reconstruction algorithms. In Section IV, we describe two modifications of the aforementioned algorithms, suitable for quantized data, that offer significant performance improvements when copmared to standard BP and SP techniques. Simulation results are presented in Section V.

II. Preliminaries

A. Compressive Sensing (CS)

In CS, one encodes a signal $\mathbf{x}$ of dimension N by computing a measurement vector $\mathbf{y}$ of dimension of $m \ll N$ via linear projections, i.e.,

$$\mathbf{y} = \mathbf{\Phi}\mathbf{x},$$

where $\mathbf{\Phi} \in \mathbb{R}^{m\times N}$ is referred to as the *measurement matrix*. In this paper, we assume that $\mathbf{x} \in \mathbb{R}^N$ is exactly K-sparse, i.e., that there are exactly K entries of $\mathbf{x}$ that are nonzero. The reconstruction problem is to recover $\mathbf{x}$ given $\mathbf{y}$ and $\mathbf{\Phi}$.

The BP method is a technique that casts the reconstruction problem as an l_1-regularized optimization problem, i.e.,

$$\min \|\mathbf{x}\|_1 \text{ subject to } \mathbf{y} = \mathbf{\Phi}\mathbf{x}, \tag{1}$$

where $\|\mathbf{x}\|_1 = \sum_{i=1}^{N} |x_i|$ denotes the l_1-norm of the vector $\mathbf{x}$. It is a convex optimization problem and can be solved efficiently by linear programming techniques. The reconstruction complexity equals $O\left(m^2 N^{3/2}\right)$ if the convex optimization problem is solved using interior point methods [11].

The computational complexity of CS reconstruction can be further reduced by the SP algorithm, recently proposed by two research groups independently in [5] and [6]. It is an iterative algorithm drawing on the theory of list decoding. The computational complexity of this algorithm is upper bounded by $O\left(Km(N+K^2)\right)$, which is significantly smaller than that of the BP method when $K \ll N$. See [5] for a detailed performance and complexity analysis of this greedy algorithm.

For completeness, we briefly describe the SP algorithm. For an index set $T \subset \{1, 2, \cdots, N\}$, let $\mathbf{\Phi}_T$ be the "truncated matrix" consisting of the columns of $\mathbf{\Phi}$ indexed by T, and let $\mathrm{span}\left(\mathbf{\Phi}_T\right)$ be the subspace in $\mathbb{R}^m$ spanned by the columns of $\mathbf{\Phi}_T$. Suppose that $\mathbf{\Phi}_T^* \mathbf{\Phi}_T$ is invertible. For any given $\mathbf{y} \in \mathbb{R}^m$, the projection of $\mathbf{y}$ onto $\mathrm{span}\left(\mathbf{\Phi}_T\right)$ is defined as

$$\begin{aligned} \mathbf{y}_p &= \mathrm{proj}\left(\mathbf{y}, \mathbf{\Phi}_T\right) \\ &:= \mathbf{\Phi}_T \left(\mathbf{\Phi}_T^* \mathbf{\Phi}_T\right)^{-1} \mathbf{\Phi}_T^* \mathbf{y}. \end{aligned} \tag{2}$$

The corresponding projection residue vector $\mathbf{y}_r$ and projection coefficient vector $\mathbf{x}_p$ are defined as

$$\mathbf{y}_r = \mathrm{resid}\left(\mathbf{y}, \mathbf{\Phi}_T\right) := \mathbf{y} - \mathbf{y}_p, \tag{3}$$

and

$$\begin{aligned} \mathbf{x}_p &= \mathrm{pcoeff}\left(\mathbf{y}, \mathbf{\Phi}_T\right) \\ &:= \left(\mathbf{\Phi}_T^* \mathbf{\Phi}_T\right)^{-1} \mathbf{\Phi}_T^* \mathbf{y}. \end{aligned} \tag{4}$$

Then the SP algorithm is summarized in Algorithm 1.

Algorithm 1 The Subspace Pursuit (SP) Algorithm

Input: K, $\mathbf{\Phi}$, $\mathbf{y}$

Initialization: Let $T^0 = \{K$ indices of the largest magnitude entries in $\mathbf{\Phi}^* \mathbf{y}\}$ and $\mathbf{y}_r^0 = \mathrm{resid}\left(\mathbf{y}, \mathbf{\Phi}_{\hat{T}^0}\right)$.

Iteration: At the ℓ^{th} iteration, go through the following steps.

1) $\tilde{T}^\ell = T^{\ell-1} \bigcup \{K$ indices of the largest magnitude entries in $\mathbf{\Phi}^* \mathbf{y}_r^{\ell-1}\}$.
2) Let $\mathbf{x}_p = \mathrm{pcoeff}\left(\mathbf{y}, \mathbf{\Phi}_{\tilde{T}^\ell}\right)$ and $T^\ell = \{K$ indices of the largest magnitude entries in $\mathbf{x}_p\}$.
3) $\mathbf{y}_r^\ell = \mathrm{resid}\left(\mathbf{y}, \mathbf{\Phi}_{T^\ell}\right)$.
4) If $\left\|\mathbf{y}_r^\ell\right\|_2 > \left\|\mathbf{y}_r^{\ell-1}\right\|_2$, let $T^\ell = T^{\ell-1}$ and quit the iteration.

Output: The vector $\hat{\mathbf{x}}$ satisfying $\hat{\mathbf{x}}_{\{1,\cdots,N\}-T^\ell} = \mathbf{0}$ and $\hat{\mathbf{x}}_{T^\ell} = \mathrm{pcoeff}\left(\mathbf{y}, \mathbf{\Phi}_{T^\ell}\right)$.

A sufficient condition for both BP and SP algorithms to perform exact reconstruction is the so called restricted isometry property (RIP) [2], formally defined as follows.

Definition 1 (RIP)**.** A matrix $\mathbf{\Phi} \in \mathbb{R}^{m \times N}$ is said to satisfy the Restricted Isometry Property (RIP) with parameters (K, δ) for $K \leq m$, $0 \leq \delta \leq 1$, if for all index sets $I \subset \{1, \cdots, N\}$ such that $|I| \leq K$ and for all $\mathbf{q} \in \mathbb{R}^{|I|}$, one has

$$(1-\delta)\|\mathbf{q}\|_2^2 \leq \|\mathbf{\Phi}_I \mathbf{q}\|_2^2 \leq (1+\delta)\|\mathbf{q}\|_2^2.$$

The RIP constant is defined as the infimum of all parameters δ for which the RIP holds, i.e.,

$$\begin{aligned} \delta_K := \inf \Big\{ \delta : \ & (1-\delta)\|\mathbf{q}\|_2^2 \leq \|\mathbf{\Phi}_I \mathbf{q}\|_2^2 \leq (1+\delta)\|\mathbf{q}\|_2^2, \\ & \forall\, |I| \leq K,\ \forall \mathbf{q} \in \mathbb{R}^{|I|} \Big\}. \end{aligned}$$

Most known families of matrices satisfying the RIP property with optimal or near-optimal performance guarantees are random, and include Gaussian random matrices with i.i.d. $\mathcal{N}\left(0, \frac{1}{m}\right)$ entries, where $m \geq O\left(K \log N\right)$.

B. Scalar Quantization

Let $\mathcal{C} \subset \mathbb{R}$ be a finite discrete set, referred to as a codebook. A quantization is a mapping from $\mathbb{R}$ to the codebook $\mathcal{C}$ such that

$$\begin{aligned} \mathfrak{q} : \ & \mathbb{R} \to \mathcal{C} \\ & y \mapsto \omega \in \mathcal{C} \text{ iff } y \in \mathcal{R}_\omega, \end{aligned} \tag{5}$$

where ω is referred to as a *level* and $\mathcal{R}_\omega$ is the *quantization region* corresponding to level ω. The performance of a quantizer is often described by its distortion-rate function defined as follows. The distortion measure is assumed to be the squared Euclidean distance. For a random source Y, the distortion associated with a quantizer $\mathfrak{q}$ is $D_{\mathfrak{q}} := \mathrm{E}\left[(Y - \mathfrak{q}(Y))^2\right]$. For a given codebook $\mathcal{C}$, the optimal quantization level that minimizes the Euclidean distortion measure is given by

$$\mathfrak{q}^*(Y) = \arg \min_{\omega \in \mathcal{C}} \ (Y - \omega)^2.$$

The distortion associated with this codebook $\mathcal{C}$ equals

$$D(\mathcal{C}) := \mathrm{E}\left[(Y - \mathfrak{q}^*(Y))^2\right].$$

Let $R := \log_2 |\mathcal{C}|$ be the rate of the codebook $\mathcal{C}$. For a given code rate R, the distortion rate function is given by

$$D^*(R) := \inf_{\mathcal{C}:\ |\mathcal{C}| \leq 2^R} D(\mathcal{C}).$$

Necessary conditions for optimal quantizer design can be found in [12]. In this paper, we assume that the random variable Y does not have mass points and that $\omega_1 < \omega_2 < \cdots < \omega_{2^R}$. Let the quantization regions be

$$\mathcal{R}_k = \begin{cases} (-\infty, t_1) & k = 1 \\ [t_{i-1}, t_i) & k = 2, 3, \cdots, 2^{R_{\mathrm{fb}}} \end{cases}, \tag{6}$$

where $t_{2^R} = +\infty$, and $t_1, \cdots, t_{2^R - 1} \in \mathbb{R}$ satisfy $\omega_k \leq t_i \leq \omega_{k+1}$. Note that for simplicity, we replaced the symbol $\mathcal{R}_\omega$ in (5) with $\mathcal{R}_k$. We adhere to this notation throughout the remainder of this paper. An optimal quantizer satisfies the following conditions:

1) If the optimal quantizer has levels ω_{k-1} and ω_k, then the threshold that minimizes the mean square error (MSE) is
$$t_i = \frac{1}{2}\left(\omega_k + \omega_{k+1}\right). \tag{7}$$
2) If the optimal quantizer has thresholds t_{k-1} and t_k, then the level that minimizes the MSE is
$$\omega_k = \mathrm{E}\left[Y | \mathcal{R}_k\right]. \tag{8}$$

Lloyd's algorithm [12] for quantizer codebook design is based on the above necessary conditions for an optimal quantizer. Lloyd's algorithm starts with an initial codebook, and then in each iteration, computes the thresholds t_is according to (7) and updates the codebook via (8). Although Lloyd's algorithm does not guarantee global optimality, it produces locally optimal codebooks.

As a low-complexity alternative, a uniform quantizer is most widely used in practice. A uniform quantizer is associated with a "uniform codebook" $\mathcal{C}_u = \{\omega_1 < \omega_2 < \cdots < \omega_M\}$ for which $\omega_i - \omega_{i-1} = \omega_j - \omega_{j-1}$ for all $1 < i \neq j \leq 2^R$. For a fixed code rate R, the distortion rate function of a uniform quantizer is defined as

$$D_u^*(R) := \inf_{\mathcal{C}_u:\, |\mathcal{C}_u| \leq 2^R} D(\mathcal{C}_u).$$

For a given probability density function, the exact asymptotic distortion rate function can be quantified exactly. Denote the probability density function of the source Y by $p(y)$. It was shown by Zador [13] that

$$\lim_{R\to\infty} 2^{2R} D^*(R) = \frac{1}{12}\left(\int p^{1/3}(x)\, dx\right)^3.$$

If the source is Gaussian with variance σ^2, then the corresponding asymptotic distortion rate function becomes

$$\lim_{R\to\infty} 2^{2R} D^*(R) = \frac{\sigma^2 \pi \sqrt{3}}{2}. \tag{9}$$

The distortion rate function of a uniform quantizer was described in [14, Theorem 6]. For Gaussian random sources with variance σ^2, one has

$$\lim_{R\to\infty} \frac{2^{2R}}{R} D_u^*(R) = \frac{4}{3}\sigma^2. \tag{10}$$

C. Scalar Quantization of CS Measurements

We study the effect of quantization on CS measurements. For simplicity, we assume a scalar quantization scheme: to each entry of $\mathbf{Y}$, say Y_i, one applies the same quantization procedure

$$\begin{aligned} \mathfrak{q}:\ & \mathbb{R} \to \mathcal{C} \subset \mathbb{R} \\ & Y_i \mapsto \hat{Y}_i = \arg\min_{\omega\in\mathcal{C}} |Y_i - \omega|^2. \end{aligned}$$

Similar to the traditional distortion-rate function for a scalar, we define the distortion rate function for quantization of the measurement vector $\mathbf{Y}$ by

$$D_{\mathbf{Y}}^*(R) := \inf_{\mathcal{C}:\, |\mathcal{C}| \leq 2^R} \mathrm{E}_{\mathbf{Y}}\left[\sum_{i=1}^m \min_{\omega\in\mathcal{C}} |Y_i - \omega|^2\right]. \tag{11}$$

When only uniform quantization is taken into consideration, the corresponding distortion-rate function is defined by

$$D_{\mathbf{Y},u}^*(R) := \inf_{\mathcal{C}_u:\, |\mathcal{C}_u| \leq 2^R} \mathrm{E}_{\mathbf{y}}\left[\sum_{i=1}^m \min_{\omega\in\mathcal{C}_u} |Y_i - \omega|^2\right], \tag{12}$$

where $\mathcal{C}_u$ denotes a uniform codebook. We are particularly interested in the total distortion of the form (11) and (12), because the CS reconstruction distortion is determined by the total distortion in the measurements rather than the distortion of each individual measurement.

D. Subgaussian Random Variables

Definition 2. A random variable X is said to be *Subgaussian* if there exist positive constants c_1 and c_2 such that

$$\Pr(|X| > x) \leq c_1 e^{-c_2 x^2} \quad \forall x > 0.$$

One property of Subgaussian distributions is that they have a well defined moment generating function. Note that the Gaussian and Bernoulli distributions are special cases of the Subgaussian distribution.

III. Distortion Analysis

A. Distortion of Measurements

We consider the following two CS scenarios.

Assumptions I:

1) Let $\mathbf{\Phi} = \frac{1}{\sqrt{m}}\mathbf{A} \in \mathbb{R}^{m\times N}$, where the entries of $\mathbf{A}$ are i.i.d. Subgaussian random variables with zero mean and unit variance.
2) Let $\mathbf{X} \in \mathbb{R}^N$ be exactly K-sparse, that is, the signal $\mathbf{X}$ has exactly K nonzero entries. We assume that the nonzeros entries of $\mathbf{X}$ are i.i.d. Subgaussian random variables with zero mean and unit variance.
3) The quantization code $\mathcal{C}$ is designed offline and fixed when the measurements are taken.

Assumptions II:

1) Let $\mathbf{\Phi} \in \mathbb{R}^{m\times N}$ be such that $\frac{1}{N}\sum_{j=1}^N \|\boldsymbol{\varphi}_j\|_2^2 = 1$, where $\boldsymbol{\varphi}_j$ is the j^{th} column of the matrix $\mathbf{\Phi}$.
2) Assume that there are exactly K nonzero entries in $\mathbf{X} \in \mathbb{R}^n$, and that the nonzeros entries of $\mathbf{X}$ are i.i.d. standard Gaussian random variables with zero mean and unit variance.
3) The quantization code $\mathcal{C}$ is designed offline and fixed when the measurements are taken.

The asysmptotic distortion rate function of the measurements under the first scenario is characterized in Theorem 3.

Theorem 3. *Suppose that Assumptions I hold. Then*

$$\lim_{R\to\infty} \lim_{(K,m,N)\to\infty} \frac{2^{2R}}{K} D_{\mathbf{Y}}^*(R) = \frac{\pi\sqrt{3}}{2}, \tag{13}$$

and there exist constants $0 < c_1 < c_2$ such that

$$\lim_{R\to\infty} \lim_{(K,m,N)\to\infty} \frac{2^{2R}}{KR} D_{\mathbf{Y},u}^*(R) = \frac{4}{3}. \tag{14}$$

Remark 4. According to Theorem 3, if the quantization rate R is sufficiently large, the distortion of the optimal non-uniform quantizer is approximately $1/R$ of that of the optimal uniform quantizer.

Proof: Let $T = \{1 \leq j \leq N :\ X_j \neq 0\}$ be the support set of $\mathbf{x}$. It is easy to show that for all $1 \leq i \leq m$ and $T \subset \{1, \cdots, N\}$ such that $|T| = K$,

$$\mathrm{E}\left[\sum_{j\in T} A_{i,j} X_j\right] = 0$$

and

$$\mathrm{E}\left[\left(\sum_{j\in T} A_{i,j}X_j\right)^2\right] = K.$$

According to the Central Limit Theorem, the distribution of $\frac{1}{\sqrt{K}}\sum_{j\in T} A_{i,j}X_j$ converges weakly to the standard Gaussian distribution as $K\to\infty$. As a result, the distribution of $\sqrt{\frac{m}{K}}Y_i$ converges weakly to the standard Gaussian distribution as $K, m, N \to \infty$.

If we apply a scalar quantizer with 2^R levels to the random variable $\sqrt{\frac{m}{K}}Y_i$, then it holds that

$$\lim_{R\to\infty} \lim_{(K,m,N)\to\infty} 2^{2R} D^*(R) = \frac{\pi\sqrt{3}}{2}. \tag{15}$$

Note that

$$\begin{aligned}
&\frac{1}{K}\mathrm{E}\left[\left\|\hat{\mathbf{Y}} - \mathbf{Y}\right\|_2^2\right] \\
&= \frac{1}{m}\frac{m}{K}\mathrm{E}\left[\sum_{i=1}^{m}\left(\hat{Y}_i - Y_i\right)^2\right] \\
&= \frac{1}{m}\sum_{i=1}^{m}\mathrm{E}\left[\left(\sqrt{\frac{m}{K}}\hat{Y}_i - \sqrt{\frac{m}{K}}Y_i\right)^2\right] \\
&= \mathrm{E}\left[\left(\sqrt{\frac{m}{K}}\hat{Y}_i - \sqrt{\frac{m}{K}}Y_i\right)^2\right],
\end{aligned} \tag{16}$$

where the last line represents the distortion of quantizing $\sqrt{\frac{m}{K}}Y_i$. Combining (**??**) and (16) proves the result of (13).

Consider a uniform quantizer with codebook $\mathcal{C}_u$, such that $|\mathcal{C}_u| = 2^R$, and apply this uniform quantizer to the random variable $\sqrt{\frac{m}{K}}Y_i$. Let $K, m, N \to \infty$. Note that the distribution of $\sqrt{\frac{m}{K}}Y_i$ converges weakly to the standard Gaussian distribution. Applying the result in (10) proves the result claimed in (14). ∎

For the scenario described by Assumptions II, lower bounds on the distortion rate function are described in Theorem 5.

Theorem 5. *Suppose that Assumptions II hold. Then*

$$\liminf_{R\to\infty} \liminf_{(K,m,N)\to\infty} \frac{2^{2R}}{K} D^*_{\mathbf{y}}(R) \geq \frac{\pi\sqrt{3}}{2}, \tag{17}$$

and there exists a constant $c > 0$ such that

$$\liminf_{R\to\infty} \liminf_{(K,m,N)\to\infty} \frac{2^{2R}}{KR} D^*_{\mathbf{y},u}(R) \geq \frac{4}{3}. \tag{18}$$

Proof: Given Assumptions II, each Y_i, $1 \leq i \leq m$, is a linear combination of Gaussian random variables, and therefore each Y_i is a Gaussian random variable ifself. For a given i and a given T, the mean and the variance of Y_i are $\mathrm{E}[Y_i] = 0$ and $\sigma^2_{i,T} = \mathrm{E}\left[Y_i^2\right] = \sum_{j\in T}\varphi^2_{i,j}$, respectively. The variance depends on the row index i and the support set T. We calculate the average variance across all rows and all support sets as

$$\begin{aligned}
\bar{\sigma}^2 &= \frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{\binom{N}{K}}\sum_{T}\sum_{j\in T}\varphi^2_{i,j}\right) \\
&= \frac{1}{m}\frac{1}{\binom{N}{K}}\sum_{T}\sum_{j\in T}\left(\sum_{i=1}^{m}\varphi^2_{i,j}\right) \\
&\overset{(a)}{=} \frac{1}{m}\frac{1}{\binom{N}{K}}\sum_{j=1}^{N}\left(\sum_{T:\, j\in T}\|\boldsymbol{\varphi}_j\|_2^2\right) \\
&\overset{(b)}{=} \frac{1}{m}\frac{1}{\binom{N}{K}}\sum_{j=1}^{N}\binom{N-1}{K-1}\|\boldsymbol{\varphi}_j\|_2^2 \\
&\overset{(c)}{=} \frac{K}{m}\frac{1}{N}\sum_{j=1}^{N}\|\boldsymbol{\varphi}_j\|_2^2 \\
&\overset{(d)}{=} \frac{K}{m},
\end{aligned} \tag{19}$$

where

- (a) is obtained by exchanging the sums over T and j,
- (b) holds because for any given $1 \leq j \leq N$, there are $\binom{N-1}{K-1}$ many subsets T containing the index j,
- (c) is due to the fact that $\binom{N-1}{K-1}/\binom{N}{K} = K/N$,
- (d) follows from Assumption II-1).

Suppose that one deals with the ideal case: the support set T is known before taking measurements; and for different values of i and T, we are allowed to use different quantizers. Given i and T, we apply the optimal quantizer for the standard Gaussian random variable to $\sqrt{\frac{m}{K}}y_i$, so that the corresponding distortion rate function satisfies

$$\lim_{R\to\infty} 2^{2R} D^{i,T}(R) = \frac{\pi\left(\frac{m}{K}\sigma^2_{i,T}\right)}{2}\sqrt{3}.$$

Taking the average over all i and all T gives

$$\begin{aligned}
&\frac{1}{m}\sum_{i=1}^{m}\frac{1}{\binom{T}{K}}\sum_{T}\left(\lim_{R\to\infty} 2^{2R} D^{i,T}(R)\right) \\
&= \frac{1}{m}\sum_{i=1}^{m}\frac{1}{\binom{T}{K}}\sum_{T}\left(\frac{\pi\left(\frac{m}{K}\sigma^2_{i,T}\right)}{2}\sqrt{3}\right) \\
&= \frac{\pi}{2}\sqrt{3},
\end{aligned}$$

where the last equality follows from (19).

However, the support set T is unknown before taking the measurements. Furthermore, the same quantizer has to be employed for different choices of i and T according to Assumptions II. Thus, for every R, i and T, $\mathrm{E}\left[\left|Y_i - \hat{Y}_i\right|^2\right] \geq$

$D^{i,T}(R)$. As a result

$$\begin{aligned}
&\liminf_{R\to\infty}\frac{2^{2R}}{K}\mathrm{E}_T\left[\mathrm{E}_Y\left[\left\|\hat{\mathbf{Y}}-\mathbf{Y}\right\|_2^2\right]\right]\\
&=\liminf_{R\to\infty}\frac{2^{2R}}{\binom{N}{T}}\sum_T\frac{1}{m}\sum_{i=1}^m\frac{m}{K}\mathrm{E}_Y\left[(\hat{y}_i-y_i)^2\right]\\
&\geq\liminf_{R\to\infty}\frac{2^{2R}}{\binom{N}{T}}\sum_T\frac{1}{m}\sum_{i=1}^m D^{i,T}(R)\\
&=\frac{\pi}{2}\sqrt{3}.
\end{aligned}$$

Since the above derivation is valid for all K, m and N, the claim in (17) is proved.

The result in (10) for uniform quantizers can be proved using similar arguments. ∎

Remark 6. Our work is based on the fundamental assumption that the sparsity level K is known in advance and that the statistics of the sparse vector $\mathbf{x}$ is specified. Very frequently, however, this is not the case in practice. If we relax Assumptions I and II further by assuming that K is sufficently large, it will often be the case that the statistics of the measurement Y_i is well approximated by a Gaussian distribution. Here, note that different Y_i variables may have different variances and these variances are generally unknown in advance. The problem of statistical unmatch has been previously addressed in quantization theory [15, Chapter 8]. Particularly, non-uniform quantizations with slightly under-estimated variance perform better than those with over-estimated variance [15, Chapter 8.6].

B. Reconstruction Distortion

It is well known from CS literature that the reconstruction distortion is dependent on the distortion in the measurements. Consider the quantized compressive sensing scenario, where

$$\hat{\mathbf{Y}}=\mathfrak{q}(\mathbf{Y})=\mathbf{\Phi}\mathbf{X}+\mathbf{E},$$

and where $\mathbf{E}\in\mathbb{R}^m$ denotes the quantization error. Let $\hat{\mathbf{X}}$ be the reconstructed signal based on the noisy measurements $\hat{\mathbf{Y}}$. Then the reconstruction distortion is defined as $\left\|\mathbf{X}-\hat{\mathbf{X}}\right\|_2^2$. For the BP method, the reconstruction distortion is upper bounded by (see [7])

$$\left\|\mathbf{X}-\hat{\mathbf{X}}\right\|_2^2\leq c_{lp}^2\left\|\mathbf{E}\right\|_2^2,$$

where

$$c_{bp}^2=\frac{2/\sqrt{3}}{\sqrt{1-\delta_{4K}}-\frac{1}{\sqrt{3}}\sqrt{1+\delta_{4K}}}.$$

A similar upper bound on the reconstruction distortion is derived in [5] for the SP algorithm, and is of the form

$$\left\|\mathbf{X}-\hat{\mathbf{X}}\right\|_2^2\leq c_{sp}^2\left\|\mathbf{E}\right\|_2^2,$$

where

$$c_{sp}=\frac{1+\delta_{3K}+\delta_{3K}^2}{\delta_{3K}\left(1-\delta_{3K}\right)}.$$

Consider Assumptions I. For the optimal scalar quantizer, the reconstruction distortion can be upper bounded by

$$\begin{aligned}
&\lim_{R\to\infty}\lim_{(K,m,N)\to\infty}\frac{2^{2R}}{K}\mathrm{E}\left[\left\|\mathbf{X}-\hat{\mathbf{X}}\right\|_2^2\right]\\
&\leq\begin{cases}c_{sp}^2\frac{\pi\sqrt{3}}{2} & \text{for the SP algorithm}\\ c_{bp}^2\frac{\pi\sqrt{3}}{2} & \text{for the BP algorithm}\end{cases}.
\end{aligned}\tag{20}$$

For the optimal uniform quantizer, an upper bound for the reconstruction distortion is given by

$$\begin{aligned}
&\lim_{R\to\infty}\lim_{(K,m,N)\to\infty}\frac{2^{2R}}{KR}\mathrm{E}\left[\left\|\mathbf{X}-\hat{\mathbf{X}}\right\|_2^2\right]\\
&\leq\begin{cases}c_{sp}^2\frac{4}{3} & \text{for the SP algorithm}\\ c_{bp}^2\frac{4}{3} & \text{for the BP algorithm}\end{cases}.
\end{aligned}\tag{21}$$

The upper bounds on the reconstruction distortion increase for the scenario described in Assumptions II. The upper bounds for optimal quantizers and optimal uniform quantizers are lower bounded by (20) and (21), respectively.

IV. Reconstruction Algorithms for Quantized CS

We present next modifications of BP and SP algorithms that take into account quantization effects.

To describe these algorithms, we find the following notation useful. Let $\hat{\mathbf{Y}}$ be the quantized measurement vector. From $\hat{\mathbf{Y}}$, the corresponding quantization regions for each coordinate of $\hat{\mathbf{Y}}$ can be easily identified. Let $\hat{Y}_i\in\mathcal{R}_{k_i}$, where $\hat{Y}_i$ is the i^{th} entry of $\hat{\mathbf{Y}}$ and $\mathcal{R}_{k_i}$ is the corresponding quantization region as given by (6). We represent the Cartesian product set of the $\mathcal{R}_{k_i}$s by $\boldsymbol{\mathcal{R}}$: a vector $\mathbf{y}$ is in $\boldsymbol{\mathcal{R}}$ if and only if $\mathbf{y}\in\mathbb{R}^m$ and $y_i\in\mathcal{R}_{k_i}$ for all $i=1,2,\cdots,m$.

Similar to the standard BP method, the reconstruction problem can be now casted as

$$\min\left\|\mathbf{x}\right\|_1\ \text{subject to}\ \mathbf{\Phi}\mathbf{x}\in\boldsymbol{\mathcal{R}},\tag{22}$$

which again can be efficiently solved by linear programming techniques.

In order to adapt the SP algorithm to the quantization scenario at hand, we describe first a geometric interpretation of the projection operation in the SP algorithm. Given $\mathbf{y}\in\mathbb{R}^m$ and $\mathbf{\Phi}_T\in\mathbb{R}^{m\times|T|}$, the projection operation in (2) is used to find a linear combination of the columns of $\mathbf{\Phi}_t$ that best approximates $\mathbf{y}$ (in the l_2-norm), that is,

$$\min_{\mathbf{x}\in\mathbb{R}^k}\left\|\mathbf{y}-\mathbf{\Phi}_t\mathbf{x}\right\|_2^2.\tag{23}$$

Let $\mathbf{x}^*$ be the solution of the quadratic optimization problem (23). Then functions (2-4) are equivalent to $\mathrm{proj}(\mathbf{y},\mathbf{\Phi}_t)=\mathbf{\Phi}_t\mathbf{x}^*$, $\mathrm{resid}(\mathbf{y},\mathbf{\Phi}_t)=\mathbf{y}-\mathbf{\Phi}_t\mathbf{x}^*$ and $\mathrm{pcoeff}(\mathbf{y},\mathbf{\Phi}_t)=\mathbf{x}^*$.

The modified SP algorithm is based on the above geometric interpretation. For quantized compressive sensing, one does not know the exact value of $\mathbf{Y}$. The only information available is that $\mathbf{Y}\in\boldsymbol{\mathcal{R}}$. Following the geometric interpretation of the projection operation, one may intuitively use

$$\min_{\mathbf{x}\in\mathbb{R}^k,\mathbf{y}\in\boldsymbol{\mathcal{R}}}\left\|\mathbf{y}-\mathbf{\Phi}_t\mathbf{x}\right\|_2^2\tag{24}$$

to replace the optimization problem in (23). But there exists a problem associated with this approach. Note that

$$\|\mathbf{y}-\mathbf{\Phi}_t\mathbf{x}\|_2^2=\left\|[\mathbf{I}\ -\mathbf{\Phi}_t]\begin{bmatrix}\mathbf{y}\\ \mathbf{x}\end{bmatrix}\right\|_2^2$$

and the matrix $[\mathbf{I}\ -\mathbf{\Phi}_t]$ does not have full column rank. Consequently, the quadratic optimization problem (24) may not have a unique solution. To solve this difficulty, we use the following definition.

Definition 7. For given $\mathbf{\Phi}_t\in\mathbb{R}^{m\times k}$, $\hat{\mathbf{Y}}$ and $\boldsymbol{\mathcal{R}}$, define

$$\mathcal{O}:=\left\{(\mathbf{x},\mathbf{y})\in\mathbb{R}^k\times\overline{\boldsymbol{\mathcal{R}}}:\right.$$
$$\left.\|\mathbf{y}-\mathbf{\Phi}_t\mathbf{x}\|_2\le\|\mathbf{y}'-\mathbf{\Phi}_t\mathbf{x}'\|_2\ \forall(\mathbf{x}',\mathbf{y}')\in\mathbb{R}^k\times\boldsymbol{\mathcal{R}}\right\},$$

where $\overline{\boldsymbol{\mathcal{R}}}$ is the closure of $\overline{\boldsymbol{\mathcal{R}}}$, and

$$(\mathbf{x}^*,\mathbf{y}^*)=\underset{(\mathbf{x},\mathbf{y})\in\mathcal{O}}{\arg\min}\left\|\mathbf{y}-\hat{\mathbf{Y}}\right\|_2.$$

It can be verified that $\mathbf{x}^*$ and $\mathbf{y}^*$ are well defined.

Based on Definition 7, we redefine the projection operation for the modified SP algorithm as follows. For given measurements $\hat{\mathbf{Y}}\in\mathbb{R}^m$, codebook $\mathcal{C}$ and $\mathbf{\Phi}_T\in\mathbb{R}^{m\times|T|}$, we obtain $\boldsymbol{\mathcal{R}}$ and then compute $\mathbf{x}^*$ and $\mathbf{y}^*$. Then the modified projection functions are defined by

$$\mathbf{y}_p^{(q)}=\mathrm{proj}^{(q)}\left(\hat{\mathbf{Y}},\mathbf{\Phi}_T\right):=\mathbf{\Phi}_T\mathbf{x}^*, \tag{25}$$

$$\mathbf{y}_r^{(q)}=\mathrm{resid}^{(q)}\left(\hat{\mathbf{Y}},\mathbf{\Phi}_T\right):=\mathbf{y}^*-\mathbf{\Phi}_T\mathbf{x}^*, \tag{26}$$

and

$$\mathbf{x}_p^{(q)}=\mathrm{pcoeff}^{(q)}\left(\hat{\mathbf{Y}},\mathbf{\Phi}_T\right):=\mathbf{x}^*, \tag{27}$$

where the superscript (q) emphasizes that these definitions are for the quantized case. Finally, we replace the resid and pcoeff functions in Algorithm 1 with the new functions $\mathrm{resid}^{(q)}$ and $\mathrm{pcoeff}^{(q)}$. This gives the modified SP algorithm for reconstruction from quantized measurements.

Remark 8. Both the modified BP algorithm in (22) and the modified SP algorithm work for vector quantization as well.

Remark 9. As discussed in [7], [5], it is often the case that the energy of quantization error, or an upper bound on the error energy, is known before reconstruction. This case can be coped with by replacing the quantization region $\boldsymbol{\mathcal{R}}$ with the l_2-ball $\left\{\mathbf{y}:\ \left\|\mathbf{y}-\hat{\mathbf{Y}}\right\|\le c\right\}$ where $c>0$ is the error energy. On the other hand, the subspace $\boldsymbol{\mathcal{R}}$ used in this paper provides finer information about $\mathbf{Y}$ than the l_2-ball and therefore allows for better reconstruction performance.

V. Empirical Results

We performed extensive computer simulations in order to compare the performance of different quantizers and different reconstruction algorithms empirically. The parameters used in our simulations were $m=128$, $N=256$ and $K=6$. Given these parameters, we generated realizations of $m\times N$ sampling matrices from the standard i.i.d. Gaussian ensemble and normalize the columns to have unit l_2-norm. We also selected a support set T of size $|T|=K$ uniformly at random,

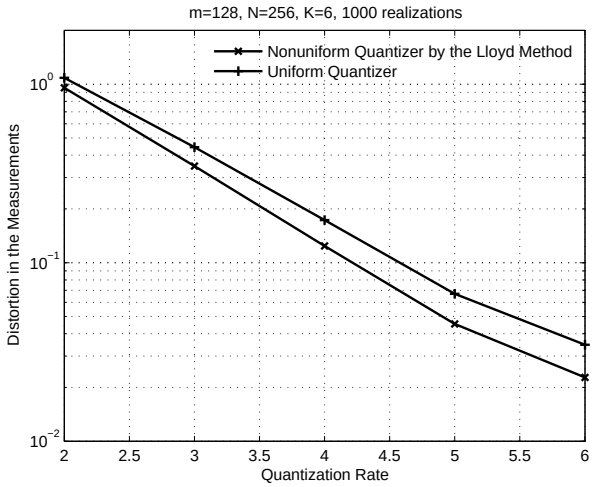


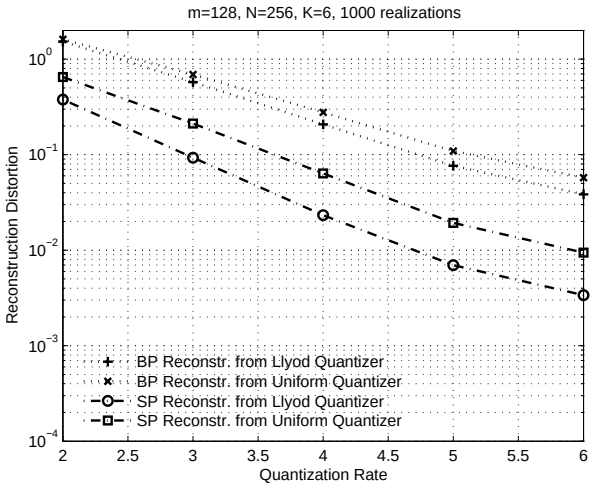
Figure 1: Distortion in the measurements.

generated the entries supported by T from the standard i.i.d. Gaussian distribution and set all other entries to zero. We let quantization rates vary from 2 to 6 bits. For each quantization rate, we used Lloyd's algorithm (Section II-B) to obtain a nonuniform quantizer and employed brute-force search to find the optimal uniform quantizer. To test different quantizers and reconstruction algorithms, we randomly generated $\mathbf{\Phi}$ and $\mathbf{x}$ independently thousand times. For each realization, we calculated the measurements $\mathbf{Y}$, the quantized measurements $\hat{\mathbf{Y}}$ and the reconstructed signals $\hat{\mathbf{X}}$.

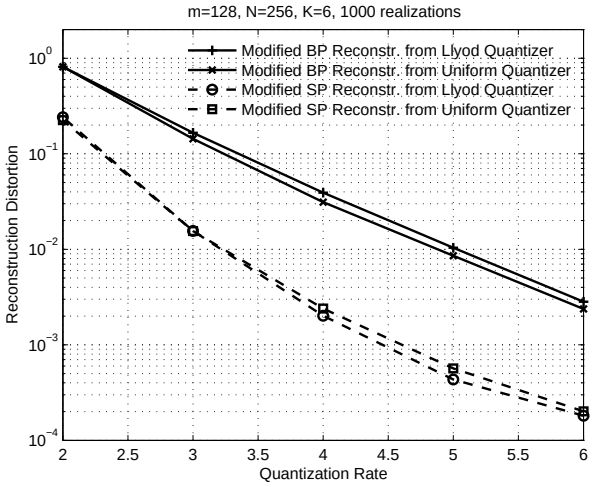
Fig. 1 compares uniform and uniform quantizers with respect to measurement distortion. Though the quantization rates in our experiments are relatively small, the simulation results are consistent with the asymptotic results in Theorem 3: nonuniform quantization is better than uniform quantization and the gain increases with the quantization rate. Fig. 2a compares the reconstruction distortion of the standard BP and SP algorithms. The comparison of the modified algorithms is given in Fig. 2. The modified algorithms reduce the reconstruction distortion significantly. When the quantization rate is 6 bits, the reconstruction distortion of the modified algorithms is roughly only one tenth of that of standard algorithms. Furthermore, for both standard and modified algorithms, the reconstruction distortion given by SP algorithms is much smaller than that of BP methods. Note that the computational complexity of the SP algorithms is also smaller than that of BP methods, which shows clear advantages for using SP algorithms in conjuction with quantized CS data. An interesting phenomenon happens for the case of the modified BP method: although the nonuniform quantization gives smaller measurement distortion, the corresponding reconstruction distortion is actually slightly larger than that of uniform quantization. We do not have solid analytical argument to completely explain this somewhat counterintuitive fact.

References

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



(a) By standard reconstruction algorithms



(b) By modified reconstruction algorithms

Figure 2: Distortion in the reconstruction signals.

[2] E. Candès and T. Tao, "Decoding by linear programming," *Information Theory, IEEE Transactions on*, vol. 51, no. 12, pp. 4203–4215, 2005.

[3] E. Candès, M. Rudelson, T. Tao, and R. Vershynin, "Error correction via linear programming," in *IEEE Symposium on Foundations of Computer Science (FOCS)*, pp. 295 – 308, 2005.

[4] E. J. Candès and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?," *IEEE Trans. Inform. Theory*, vol. 52, no. 12, pp. 5406–5425, 2006.

[5] W. Dai and O. Milenkovic, "Subspace pursuit for compressive sensing signal reconstruction," *IEEE Trans. Inform. Theory*, submitted, 2008.

[6] D. Needell and J. A. Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples," *Appl. Comp. Harmonic Anal.*, accepted, 2008.

[7] E. J. Candès, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Comm. Pure Appl. Math.*, vol. 59, no. 8, pp. 1207–1223, 2006.

[8] E. Candès and J. Romberg, "Encoding the ℓ_p ball from limited measurements," *Data Compression Conference*, pp. 33–42, March 2006.

[9] P. Boufounos and R. Baraniuk, "Quantization of sparse representations," *Preprint*, 2008.

[10] P. Boufounos and R. G. Baraniuk, "1-bit compressive sensing," in *Conf. on Info. Sciences and Systems (CISS)*, (Princeton, NJ), pp. 16–21, March 2008.

[11] I. E. Nesterov, A. Nemirovskii, and Y. Nesterov, *Interior-Point Polynomial Algorithms in Convex Programming*. SIAM, 1994.

[12] S. Lloyd, "Least squares quantization in pcm," *Information Theory, IEEE Transactions on*, vol. 28, pp. 129–137, Mar 1982.

[13] P. Zador, *Development and evaluation of procedures for quantizing multivariate distributions*. PhD thesis, Stanford University, Stanford, CA, 1964.

[14] D. Hui and D. Neuhoff, "Asymptotic analysis of optimal fixed-rate uniform scalar quantization," *IEEE Trans. Inform. Theory*, vol. 47, pp. 957–977, Mar 2001.

[15] K. Sayood, *Introduction to Data Compression*. Morgan Kaufmann, 3rd edition ed., 2005.