Sigma Delta Quantization for Compressed Sensing

C. Sinan Güntürk,[1] Mark Lammers,[2] Alex Powell,[3] Rayan Saab,[4] Özgür Yılmaz [4]
[1]Courant Institute of Mathematical Sciences, New York University, NY, USA.
[2]University of North Carolina, Wilmington, NC, USA.
[3]Vanderbilt University, Nashville, TN, USA.
[4]University of British Columbia, Vancouver BC, Canada.

***Abstract*—Recent results make it clear that the compressed sensing paradigm can be used effectively for dimension reduction. On the other hand, the literature on quantization of compressed sensing measurements is relatively sparse, and mainly focuses on pulse-code-modulation (PCM) type schemes where each measurement is quantized independently using a uniform quantizer, say, of step size δ. The robust recovery result of Candès et al. and Donoho guarantees that in this case, under certain generic conditions on the measurement matrix such as the restricted isometry property, ℓ^1 recovery yields an approximation of the original sparse signal with an accuracy of $O(\delta)$. In this paper, we propose sigma-delta quantization as a more effective alternative to PCM in the compressed sensing setting. We show that if we use an rth order sigma-delta scheme to quantize m compressed sensing measurements of a k-sparse signal in $\mathbb{R}^N$, the reconstruction accuracy can be improved by a factor of $(m/k)^{(r-1/2)\alpha}$ for any $0 < \alpha < 1$ if $m \gtrsim_r k(\log N)^{1/(1-\alpha)}$ (with high probability on the measurement matrix). This is achieved by employing an alternative recovery method via rth-order Sobolev dual frames.**

I. Introduction

Sparse approximations play an increasingly important role in signal and image processing. This relies on the fact that various classes of signals, e.g., audio and images, can be well-approximated by only a few elements of an appropriate basis or frame. For such signals, compressed sensing (CS) [6], [8], [11] provides an efficient sampling theory: high-dimensional signals with sparse approximations can be recovered from significantly fewer measurements than their ambient dimension by means of efficient non-linear reconstruction algorithms, e.g., ℓ^1 minimization or greedy algorithms such as orthogonal matching pursuit (OMP).

One of the main potential applications of CS is analog-to-digital (A/D) conversion, e.g., see [19], [22], yet the literature on how to quantize CS measurements in an efficient and robust way is relatively sparse. In this paper, we will establish a link between frame quantization and compressed sensing quantization. This will enable us to assess the limitations of pulse code modulation (PCM) based quantization methods in the CS setting. More importantly, using techniques from frame quantization theory, we shall construct "noise-shaping" quantizers for CS measurements that yield superior approximations.

II. Background on Compressed Sensing

Let Σ_k^N denote the space of *k-sparse* signals in $\mathbb{R}^N$. Suppose the *measurements* of x are given by $y = \Phi x$ where Φ is an $m \times N$ *measurement matrix* with $m < N$. The goal in CS is to choose Φ such that one can recover every $x \in \Sigma_k^N$ exactly from y. For such a guarantee, it is sufficient to have a matrix Φ that is in general position with $m \geq 2k$, but only along with a recovery algorithm, the so-called ℓ^0 minimization, that is computationally intractable. On the other hand, under stricter conditions on Φ, e.g., the restricted isometry property (RIP), one can recover sparse vectors by means of computationally efficient algorithms, such as ℓ^1 minimization. Furthermore, in this case, the recovery will be stable (if the original signal is not sparse, but compressible) and robust (if the measurements are corrupted by additive noise).

Suppose $\hat{y} = \Phi x + e$ where e is any vector with $\|e\|_2 \leq \epsilon$. Define $\Delta_1^\epsilon : \mathbb{R}^m \mapsto \mathbb{R}^N$ via

$$\Delta_1^\epsilon(\hat{y}) := \arg\min_z \|z\|_1 \quad \text{subject to} \quad \|\Phi z - \hat{y}\|_2 \leq \epsilon.$$

It was shown by Candès et al. [7], and by Donoho [12], that if Φ satisfies an appropriate RIP condition, then $x^\# := \Delta_1^\epsilon(\hat{y})$ satisfies

$$\|x - x^\#\| \lesssim \epsilon. \tag{1}$$

Checking numerically whether a given matrix Φ satisfies the RIP becomes rapidly intractable as the dimension of the problem grows. On the other hand, it was shown that if Φ is a random matrix the entries of which are sampled independently, e.g., from the Gaussian distribution $\mathcal{N}(0, 1/m)$, then Φ satisfies RIP (that guarantees (1) to hold) if $m \gtrsim k \log(\frac{N}{k})$.

III. Compressed Sensing and Quantization

The robust recovery result outlined above justifies that CS is effective for dimension reduction. On the other hand, it is not clear whether CS is in fact "compressive". For this, one needs to convert the measurements into bit-streams, i.e., to quantize the measurements, and compute the relationship between the bit budget and the accuracy of the recovery after quantization.

A *quantizer* is a map $Q : \mathbb{R}^m \mapsto \mathcal{A}^m$ where $\mathcal{A}$ is a discrete set, typically called the *quantizer alphabet*. Suppose $y = \Phi x$ are the CS measurements of a sparse signal x. We will focus on how to quantize y, which consists of two sub-problems: (i) designing the quantizer Q; (ii) finding an appropriate recovery algorithm Δ_Q, possibly tailored for the underlying quantizer Q. Throughout, we will fix the quantizer alphabet $\mathcal{A} = \delta\mathbb{Z}$, and for various quantizers and recovery methods, we will estimate the reconstruction error $\|x - \Delta_Q(Q(y))\|_2$ as a function of the dimensional parameters m, k, and N.

Remark on normalization of Φ*:* In the CS literature, it is conventional to normalize a random measurement matrix Φ so that it has unit column variance. This scaling ensures that $E\|\Phi x\|^2 = \|x\|^2$ for any input x, which then leads to RIP through concentration of measure, and finally to the robust recovery result given in (1). On the other hand, such a normalization scales the dynamic range of each measurement y_j by $1/\sqrt{m}$. In this paper, we investigate the dependence of the recovery error on the number of quantized measurements with fixed quantizer step size δ. A fair assessment of such a dependence can be made only if the dynamic range of each measurement is kept constant while increasing the number of measurements. This can be ensured by normalizing the measurement matrix Φ such that its entries are independent of m. In the specific case of Gaussian matrices, we can achieve this by choosing the entries of Φ i.i.d. according to $\mathcal{N}(0,1)$. With this normalization of Φ, the robust recovery result of [7], given above, can be modified as

$$\|\hat{y} - y\|_2 \leq \epsilon \implies \|x - x^{\#}\|_2 \lesssim \frac{1}{\sqrt{m}}\epsilon. \tag{2}$$

The transition between these two conventions is of course trivial.

IV. PCM for Compressed Sensing Quantization

Perhaps the most intuitive quantization strategy is to replace each entry y_j of the measurement vector y with q_j that is the nearest element of $\mathcal{A}$ to y_j. The associated quantization scheme is typically called pulse code modulation (PCM). In this case, $\|y - q\| \leq \frac{1}{2}\delta\sqrt{m}$, and thus it follows from (2) that $x^{\#}_{\text{PCM}} = \Delta_1^{\epsilon}(q)$ with $\epsilon = \frac{1}{2}\delta\sqrt{m}$ satisfies

$$\|x - x^{\#}_{\text{PCM}}\|_2 \lesssim \delta. \tag{3}$$

Note that the error bound given above does not improve if we increase the number of measurements m. Of course, the actual error may behave differently as (3) provides only an upper bound on the reconstruction error. However, as seen in Figures 1 and 2, numerical experiments also corroborate that the approximation error $\|x - x^{\#}_{\text{PCM}}\|_2$ does not decay as m increases. This suggests that PCM quantization together with reconstruction via ℓ^1 minimization is not utilizing extra information obtained by collecting more measurements.

There are two ingredients in the above analysis: the quantizer (PCM) and the reconstruction method (ℓ^1 minimization). The vast logarithmic reduction of the ambient dimension N would seem to suggest that PCM quantization is essentially optimal since information appears to be squeezed (compressed) into few uncorrelated measurements. Perhaps for this reason, the existing literature on quantization of compressed sensing measurements focused mainly on alternative reconstruction methods from PCM-quantized measurements and variants thereof, e.g., [4], [9], [13], [17], [20], [23]. (The only exception we are aware of is [5], which uses $\Sigma\Delta$ modulation to quantize x *before* the random measurements are made.) Next, we discuss why PCM is in fact highly suboptimal in this setting, and give a lower bound on the approximation error $\|x - x^{\text{opt}}_{\text{PCM}}\|$ where $x^{\text{opt}}_{\text{PCM}}$ is the optimal (consistent) reconstruction obtained from PCM-quantized CS measurements of x.

Compressed sensing and oversampling*:* Despite the "vast reduction of the ambient dimension", the measurements obtained in CS are in fact not uncorrelated. To see this, let $x \in \Sigma_k^N$ and let T be the support of x. Denote by x_T the vector in $\mathbb{R}^k$ consisting of those entries of x the indices of which are in T (if $|T| < k$, just complete it to a set of size k). Suppose we knew (recovered) T. Then in the CS setting, we have $m > k$ measurements of the k-dimensional signal x_T. *For processing purposes* it is important to remember that the measurement vector y is in fact a *redundant encoding* of x_T. In particular, $y = \Phi_T x_T$ where Φ_T is the $m \times k$ submatrix consisting of the columns of Φ indexed by T. Note, now, that the collection of the rows of Φ_T is a redundant frame (with $m > k$ vectors) for $\mathbb{R}^k$ (of course assuming that Φ is in general position), and the entries of y are the associated frame coefficients.

Our discussion above has an important consequence for quantizer design in the setting of CS: if a quantization scheme is *not* effective for quantizing redundant frame expansions, then it will not be effective in the CS setting. For this reason, in the next section, we turn our attention to quantization methods for oversampled data.

We end this section by substantiating our claim that PCM is a highly suboptimal quantization strategy for CS even if one uses recovery algorithms other than those based on ℓ^1 minimization. The following theorem of Goyal et al. [14] illustrates the limitations of PCM as a frame quantization strategy, which, in the light of the discussion above, immediately translates to the CS setting as it gives a lower bound on the approximation error *even if the support of sparse signal* x *is known*.

Theorem IV.1 (Goyal et al.)**.** *Let E be an $m \times k$ real matrix, and let K be a bounded set in $\mathbb{R}^k$. For $x \in K$, suppose that we obtain $q_{PCM}(x)$ by quantizing $y = Ex$ using PCM with alphabet $\mathcal{A} = \delta\mathbb{Z}$. Let Δ_{opt} be an optimal decoder. Then*

$$\left[\mathbb{E}\,\|x - \Delta_{opt}(q_{PCM}(x))\|_2^2\right]^{1/2} \gtrsim \lambda^{-1}\delta$$

where the "oversampling rate" $\lambda := m/k$ and the expectation is with respect a probability measure on K that is, for example, absolutely continuous.

V. Finite Frames and Quantization

A collection $\{e_j\}_1^m$ in $\mathbb{R}^k$ is a *frame for* $\mathbb{R}^k$ with frame bounds $0 < A \leq B < \infty$ if

$$\forall x \in \mathbb{R}^k, \quad A\|x\|_2^2 \leq \sum_{j=1}^{m} |\langle x, e_j\rangle|^2 \leq B\|x\|_2^2.$$

It is easy to see that $\{e_j\}_1^m$ is a frame for $\mathbb{R}^k$ if and only if the $m \times k$ matrix E whose jth row is e_j^T is full-rank.[1]

Let E be a full-rank $m \times k$ matrix and F be any left inverse of E. Then the collection of the columns of F, which is also a

[1]We call E and E^T the *analysis* and *synthesis* matrices of the frame $\{e_j\}_1^m$, respectively.

frame for $\mathbb{R}^k$, is said to be a *dual* of the frame consisting of the rows of E. For $x \in \mathbb{R}^k$, let $y = Ex$ be the frame coefficient vector of x, let $q \in \mathcal{A}^m$ be a quantization of y, and let $\hat{x} := Fq$ be a (linear) reconstruction of x from its quantized frame coefficients. (Here, we restrict our attention to such linear reconstruction methods; see, e.g., [14] for further discussion.) Ideally one would wish to choose q such that $y-q \in \text{Ker}(F)$. Typically, however, this is not possible, and thus $\hat{x} \neq x$, i.e., quantization is inherently lossy. One the other hand, since $m > k$, $\text{Ker}(F)$ is an $m-k$ dimensional subspace, and we can at least hope to choose q such that $y-q$ is "close" to $\text{Ker}(F)$, i.e., employ a "noise-shaping[2]" quantization method. Note that here we have two design goals: (i) choose a good quantizer, and (ii) choose a good dual frame (i.e., a good left inverse F). Next, we discuss $\Sigma\Delta$ schemes, which are known to provide efficient quantizers for redundant representations in the settings of oversampled bandlimited functions, e.g., [10], [15], [21], and general frame expansions, e.g., [1], [18].

$\Sigma\Delta$ *schemes for frame quantization*

An rth-order $\Sigma\Delta$ scheme (with the standard "greedy" quantization rule) quantizes a vector y to q by running the recursion

$$(\Delta^r u)_j = y_j - q_j, \tag{4}$$
$$q_j = \arg\min_{a \in \mathcal{A}} \Big| \sum_{i=1}^{r} (-1)^{i-1} \binom{r}{i} u_{j-i} + y_j - a \Big|,$$

with $u_{-(r-1)} = \cdots = u_0 = 0$ (here $r \in \mathbb{N}$). It is easy to check that with this rule, one has $|u_j| \leq 2^{-1}\delta$ and $|y_j - q_j| \leq 2^{r-1}\delta$. In turn, if $\|y\|_\infty < C$, then one needs only $L := 2\lceil \frac{C}{\delta} \rceil + 2^r + 1$ levels. In this case, the associated quantizer is said to be $\log_2 L$-bit, and we have

$$\|u\|_\infty \lesssim \delta \text{ and } \|y - q\|_\infty \lesssim_r \delta. \tag{5}$$

Recently, it was shown that $\Sigma\Delta$ schemes can be effectively used for quantization of arbitrary frame expansions, e.g., [1]. Let E be a full-rank $m \times k$ matrix and let F be any left inverse of E. If the sequence $(f_j)_1^m$ of dual frame vectors (i.e., the columns of F) were known to vary smoothly in j (including smooth termination into null vector), then $\Sigma\Delta$ quantization could be employed without much alteration, e.g., [3], [18]. However, this need not be the case for many examples of frames (together with their canonical duals) that are used in practice. For this reason, it has recently been proposed in [2] to use special alternative dual frames, called Sobolev dual frames, that are naturally adapted to $\Sigma\Delta$ quantization. Among all left inverses of E, the rth-order Sobolev dual $F_{\text{sob},r}$ minimizes the operator norm of FD^r on ℓ^2 where D is the $m \times m$ difference matrix defined by

$$D_{ij} := \begin{cases} 1, & \text{if } i = j, \\ -1, & \text{if } i = j+1, \\ 0. & \text{otherwise,} \end{cases} \tag{6}$$

[2]The quantization error is often modeled as white noise in signal processing, hence the terminology. However our treatment of quantization error in this paper is entirely deterministic.

In fact, $F_{\text{sob},r}$ is given by the explicit formula

$$F_{\text{sob},r} = (D^{-r}E)^\dagger D^{-r}. \tag{7}$$

It is shown in [2] that if the rth order $\Sigma\Delta$ quantization algorithm, as in (4), is used to quantize $y = Ex$ to $q_{\Sigma\Delta} := q$, then with $\hat{x}_{\Sigma\Delta} := F_{\text{Sob},r} q_{\Sigma\Delta}$, the reconstruction error obeys the bound

$$\|x - \hat{x}_{\Sigma\Delta}\|_2 \lesssim_r \frac{\delta\sqrt{m}}{\sigma_{\min}(D^{-r}E)}. \tag{8}$$

where $\sigma_{\min}(D^{-r}E)$ stands for the smallest singular value of $D^{-r}E$. Moreover, in [2] it was also shown that if the columns of E vary smoothly, then (8) implies that

$$\|x - \hat{x}_{\Sigma\Delta}\|_2 \lesssim \lambda^{-r} \tag{9}$$

where $\lambda = \frac{m}{k}$ is the oversampling ratio of E. Note that, for sufficiently high λ, this shows that $\Sigma\Delta$ schemes of order 2 or higher outperform the optimal accuracy of PCM, which is $O(\lambda^{-1})$, at least when the analysis frame is smooth.

The following rather surprising result [16] shows that an error bound analogous to (9) also holds when E is a random Gaussian matrix with high probability.

Theorem V.1. *Let E be an $m \times k$ random matrix whose entries are i.i.d. $\mathcal{N}(0,1)$. For any $\alpha \in (0,1)$, if $\lambda \geq c(\log m)^{1/(1-\alpha)}$, then with probability at least $1 - \exp(-c' m \lambda^{-\alpha})$,*

$$\sigma_{\min}(D^{-r}E) \gtrsim_r \lambda^{\alpha(r-\frac{1}{2})}\sqrt{m}, \tag{10}$$

which yields the reconstruction error bound

$$\|x - \hat{x}_{\Sigma\Delta}\|_2 \lesssim_r \lambda^{-\alpha(r-\frac{1}{2})}\delta. \tag{11}$$

VI. $\Sigma\Delta$ Quantization for Compressed Sensing

We now return to the quantizer design problem for CS measurements. Let Φ be an $m \times N$ random Gaussian matrix with independent entries drawn from $\mathcal{N}(0,1)$, and let $x \in \Sigma_k^N$. Suppose $q_{\Sigma\Delta}$ is obtained by quantizing $y = \Phi x$ using an rth-order $\Sigma\Delta$ quantization algorithm with alphabet $\mathcal{A} = \delta\mathbb{Z}$. In this section, we will show that an accurate reconstruction of x from the $\Sigma\Delta$-quantized CS measurement vector $q_{\Sigma\Delta}$ can be obtained via a two-stage procedure:

(i) **Coarse recovery:** ℓ^1-minimization (or any other robust recovery procedure) applied to $q_{\Sigma\Delta}$ yields a "coarse" approximation $x^\#$ of x, and in particular, the exact (or approximate) support T of x.

(ii) **Fine recovery:** Sobolev dual of the frame Φ_T applied to $q_{\Sigma\Delta}$ yields a finer approximation $\hat{x}_{\Sigma\Delta}$ of x.

Next, we analyze each step of our two-stage approach.

A. Coarse recovery

Our first goal is to recover the support T of x. For this purpose we shall use a coarse approximation of x given by $x' = \Delta_1^\epsilon(q_{\Sigma\Delta})$ with $\epsilon := 2^{r-1}\delta\sqrt{m}$. By (2) we know that

$$\|x - x'\|_2 \leq \eta := C \cdot 2^{r-1}\delta$$

where C is the "robust recovery constant" associated with Φ. The simplest attempt to recover T from x' is to pick the

positions of its k largest entries. Clearly, this attempt can fail if some entry $x_j \neq 0$ is smaller in magnitude than η for then it is possible that $x'_j = 0$ and therefore j is not picked. On the other hand, if the smallest nonzero entry of x is strictly bigger than $\sqrt{2}\eta$ in magnitude, then this method always succeeds. In fact, by a careful analysis the constant $\sqrt{2}$ can be made arbitrarily close to 1 by picking more than k positions. The following proposition [16] gives a precise condition on how well this can be done.

Proposition VI.1. *Let $\|x - x'\|_{\ell_2^N} \leq \eta$, $T = \mathrm{supp}(x)$ and $k = |T|$. For any $k' \in \{k, \ldots, N-1\}$, let T' be the support of (any of) the k' largest entries of x'. If $|x_j| > \gamma\eta$ for all $j \in T$, where $\gamma := \left(1 + \frac{1}{k'-k+1}\right)^{1/2}$, then $T' \supset T$.*

B. Fine recovery

Suppose x satisfies the size condition specified in Proposition VI.1 with $\gamma = \sqrt{2}$ (in which case, we can recover the support of x perfectly). Once the support T of x is found, $E = \Phi_T$ will simply be an $m \times k$ sub-matrix of the measurement matrix Φ and thus Theorem V.1 shows that for all such x, $\|x - \hat{x}_{\Sigma\Delta}\|_2$ satisfies (11) with high probability. To obtain a uniform error bound (which, with high probability, holds for all $x \in \Sigma_k^N$ that satisfy the size condition of Proposition VI.1), we need (10) to hold uniformly for all the frames $E = \Phi_T$ where $T \subset \{1, \ldots, N\}$ with $\#T{=}k$. Indeed, such a result holds as the proof of Theorem V.1 extends in a straightforward manner using a standard "union bound" argument, provided λ is known to be slightly larger. Consequently, we obtain our main result [16].

Theorem VI.2. *Let Φ be an $m \times N$ matrix whose entries are i.i.d. according to $\mathcal{N}(0,1)$. Suppose $\alpha \in (0,1)$ and $\lambda := m/k \geq c(\log N)^{1/(1-\alpha)}$ where $c = c(r,\alpha)$. Then there are two constants c' and C that depend only on r such that with probability at least $1 - \exp(-c'm\lambda^{-\alpha})$ on the draw of Φ, the following holds: For every $x \in \Sigma_k^N$ such that $\min_{j\in\mathrm{supp}(x)} |x_j| \geq C\delta$, the reconstruction $\hat{x}_{\Sigma\Delta}$ satisfies*

$$\|x - \hat{x}_{\Sigma\Delta}\|_2 \lesssim_r \lambda^{-\alpha(r-\frac{1}{2})}\delta. \tag{12}$$

VII. Rate-Distortion Issues

Above, we showed that $\Sigma\Delta$ schemes produce better approximation error when compared to PCM if both quantizers have the same infinite alphabet $\mathcal{A} = \delta\mathbb{Z}$. Of course, such infinite quantizers are not practical, and have to be replaced with finite ones. To that end, suppose the sparse signals of interest lie in some appropriate bounded set $K := \{x \in \Sigma_k^N : A \leq |x_j| \leq \rho,\ \forall j \in T\}$, with $A \ll \rho$. Suppose, instead of $\mathcal{A} = \delta\mathbb{Z}$ we use a B_r-bit uniform quantizer with the largest allowable step-size, say δ_r, for our support recovery result in Proposition VI.1 to hold. Here, we choose B_r such that the associated $\Sigma\Delta$ quantizer does not overload, i.e., $B_r = \log_2 L$ where $L = 2\lceil \frac{C_K}{\delta} \rceil + 2^r + 1$ with $C_K = \sup_{x\in K} \|\Phi x\|_\infty$, as seen in (5). Then the approximation error (the distortion) $\mathcal{D}_{\Sigma\Delta}$ incurred after the fine recovery stage via Sobolev duals satisfies the bound

$$\mathcal{D}_{\Sigma\Delta} \lesssim_r \lambda^{-\alpha(r-1/2)}\delta_r \approx \frac{\lambda^{-\alpha(r-1/2)}A}{2^{r+1/2}}. \tag{13}$$

A similar calculation for the PCM encoder with the same step size δ_r and the standard ℓ^1 decoder results in the necessity for roughly the same number of bits B_r as the $\Sigma\Delta$ encoder, but provides only the distortion bound

$$\mathcal{D}_{\mathrm{PCM}} \lesssim \delta_r \approx \frac{A}{2^{r+1/2}}. \tag{14}$$

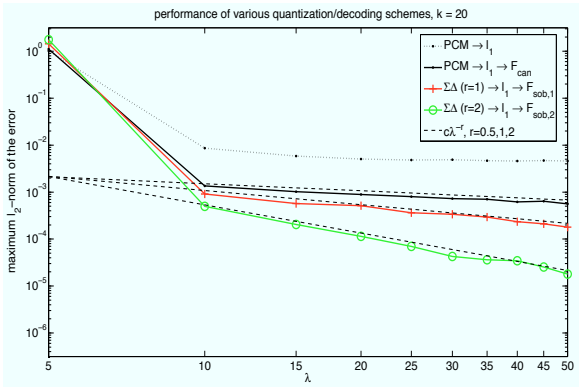
A comparison of (13) and (14) makes it clear that $\Sigma\Delta$ schemes are superior to PCM when $\lambda > 1$ is sufficiently large. The details of this discussion is given in [16].

VIII. Numerical Experiments

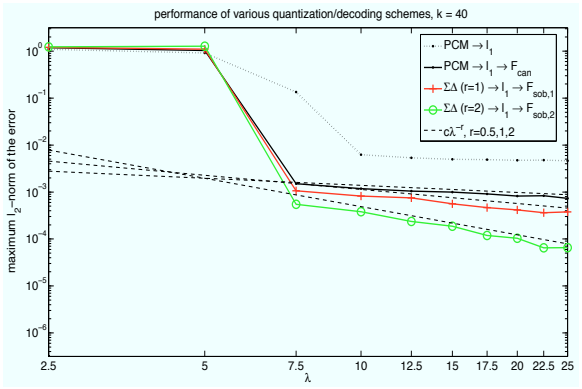
To examine the performance of the proposed scheme(s) as the redundancy λ increases in comparison to the performance of the standard PCM quantization, we run numerical experiments. First, we generate a 1000×2000 matrix Φ, where the entries of Φ are drawn i.i.d. according to $\mathcal{N}(0,1)$. In each experiment we fix the sparsity $k \in \{20, 40\}$, and we generate k-sparse signals $x \in \mathbb{R}^{2000}$ with the non-zero entries of each signal supported on a random set T, but with magnitude $1/\sqrt{k}$. This ensures that $\|x\|_2 = 1$. Next, for $m \in \{100, 200, ..., 1000\}$ we generate the measurements $y = \Phi^{(m)}x$, where $\Phi^{(m)}$ is comprised of the first m rows of Φ. We then quantize y using PCM, as well as the 1st and 2nd order $\Sigma\Delta$ quantizers, defined via (4) (in all cases the quantizer step size is $\delta = 10^{-2}$). For each of these quantized measurements q, we perform the coarse recovery stage, i.e., we solve the associated ℓ^1 minimization problem to recover a coarse estimate of x as well as an estimate $\widetilde{T}$ of the support T. The approximation error obtained using the coarse estimate (with PCM quantization) is displayed in Figure 1 (see the dotted curve). Next, we implement the fine recovery stage of our algorithm. In particular, we use the estimated support set $\widetilde{T}$ and generate the associated dual $F_{\mathrm{sob},r}$. Defining $F_{\mathrm{sob},0} := (\Phi_{\widetilde{T}}^{(m)})^\dagger$, in each case, our final estimate of the signal is obtained via the fine recovery stage as $\hat{x}_{\widetilde{T}} = F_{\mathrm{sob},r}q$ and $\hat{x}_{\widetilde{T}^c} = 0$. Note that this way, we obtain an alternative reconstruction also in the case of PCM. We repeat this experiment 100 times for each (k, m) pair and plot the maximum of the resulting errors $\|x - \tilde{x}\|_2$ as a function of λ in Figure 1. For our second experiment, we choose the entries of x_T i.i.d. from $\mathcal{N}(0,1)$, and use a quantizer step size $\delta = 10^{-4}$. Otherwise, the experimental setup is identical to the previous one. The maximum of the resulting errors $\|x - \tilde{x}\|_2$ as a function of λ is reported Figure 2.

The main observations that we obtain from these experiments are as follows:

- $\Sigma\Delta$ schemes outperform the coarse reconstruction obtained from PCM quantized measurements even when $r = 1$ and even for small values of λ.
- For the $\Sigma\Delta$ reconstruction error, the negative slope in the log-log scale is roughly equal to r. This outperforms the (best case) predictions of Theorem VI.2 which are

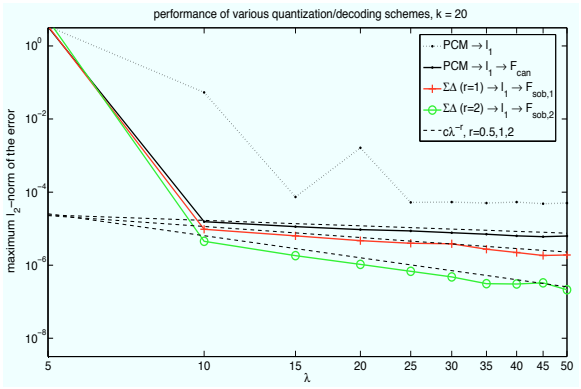


(a)

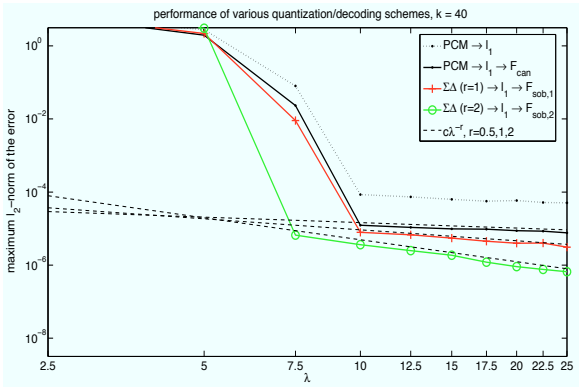


(b)

Fig. 1. The worst case performance of the proposed $\Sigma\Delta$ quantization and reconstruction schemes for various values of k. For this experiment the non-zero entries of x are constant and $\delta = 0.01$.



(a)



(b)

Fig. 2. The worst case performance of the proposed $\Sigma\Delta$ quantization and reconstruction schemes for various values of k. For this experiment the non-zero entries of x are i.i.d. $\mathcal{N}(0,1)$ and $\delta = 10^{-4}$.

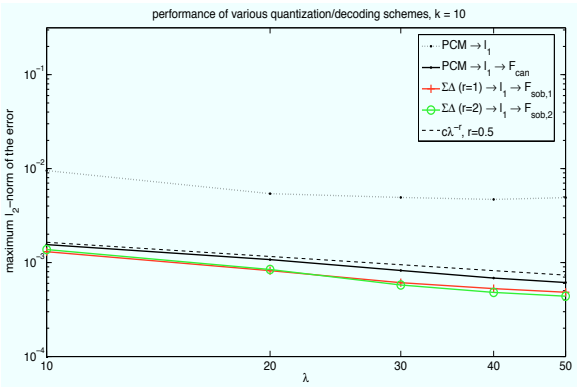
obtained through the operator norm bound and suggests the presence of further cancellation due to the statistical nature of the $\Sigma\Delta$ state variable u, similar to the white noise hypothesis.

- When a fine recovery stage is employed in the case of PCM (using the Moore-Penrose pseudoinverse of the submatrix of Φ that corresponds to the estimated support of x), the approximation is consistently improved (when compared to the coarse recovery). Moreover, the associated approximation error is observed to be of order $O(\lambda^{-1/2})$, in contrast with the error corresponding to the coarse recovery from PCM quantized measurements (with the ℓ^1 decoder only) where the approximation error does not seem to depend on λ. A rigorous analysis of this behaviour will be given in a separate manuscript.

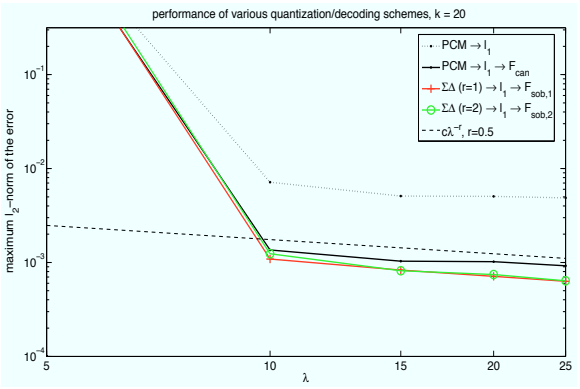
IX. Remarks on measurement noise and compressible signals

One natural question is whether the quantization methods developed in this paper are effective in the presence of measurement noise in addition to the error introduced during the quantization process. Another important issue is how to extend this theory to include the case when the underlying signals are not necessarily strictly sparse, but still "compressible".

Suppose $x \in \mathbb{R}^N$ is not sparse, but compressible in the usual sense (e.g. as in [7]), and let $\hat{y} = \Phi x + e$, where e stands for additive measurement noise. The *coarse recovery stage* inherits the stability and robustness properties of ℓ^1 decoding. Consequently, the accuracy of this first reconstruction depends on the best k-term approximation error for x, and $\Phi x - q$ which comprises of the measurement noise e and the quantization error $y - q$. Up to constant factors, the quantization error for any (stable) $\Sigma\Delta$ quantizer is comparable to that of PCM, hence the approximation accuracy at the coarse recovery stage would also be comparable. In the *fine recovery stage*, however, the difference between $\sigma_{\max}(F_{\text{sob},r}D^r)$ and $\sigma_{\max}(F_{\text{sob},r})$ plays a critical role. The Sobolev duals are tailored to reduce the effect of the quantization error introduced by an rth order $\Sigma\Delta$ quantizer. In particular, obtaining more measurements decreases the reconstruction error due to quantization even though $\|y - q\|_2$ increases. At the same time, obtaining more measurements would also increase the size of the external noise e, as well as the "aliasing error" that is the result of the "off-support" entries of x. However, this noise+error term is not counteracted by the action of $F_{\text{sob},r}$. In this case, depending on the size of the noise term, the fine recovery stage may not improve the total reconstruction error even though the "quantizer error" is still reduced.



(a)



(b)

Fig. 3. The worst case performance of the proposed leaky $\Sigma\Delta$ quantization and reconstruction schemes. In each case, noisy CS measurements $\hat{y} = \Phi x + e$ of 100 different k-sparse vectors x were quantized using the proposed leaky scheme with step size $\delta = 0.01$ and leakage parameter $\mu = 0.8$, and reconstructed using the associated H-duals. The measurement noise e was drawn independently from $\mathcal{N}(0, \sigma^2)$ with $\sigma = 5 \cdot 10^{-4}$.

One possible remedy for this problem is to construct alternative quantization schemes with associated "noise-shaping matrices" that balance the above discussed trade-off between the quantization error and the error that is introduced by other factors. This is a delicate procedure, and it will be investigated thoroughly in future work. However, a first such construction can be made by using "leaky" $\Sigma\Delta$ schemes with noise-shaping matrices H (instead of D) given by

$$H_{ij} := \begin{cases} 1, & \text{if } i = j, \\ -\mu & \text{if } i = j+1, \\ 0, & \text{otherwise,} \end{cases} \tag{15}$$

where $\mu \in (0, 1)$. We again adopt the two stage recovery approach. However, in this case, instead of Sobolev duals, we use the "H-dual" of the corresponding frame E, which we define via $F_H H = (H^{-1}E)^\dagger$. Our preliminary numerical experiments (see Figure 3) suggest that this approach can be used to improve the accuracy of the approximation further in the fine recovery stage in this more general setting. Note that the parameter μ above can be adjusted based on the expected noise level and how compressible the signals of interest are.

ACKNOWLEDGMENT

The authors would like to thank Ronald DeVore for valuable discussions, and the American Institute of Mathematics and the Banff International Research Station for hosting two meetings where this work was initiated.

REFERENCES

[1] J.J. Benedetto, A.M. Powell, and Ö. Yılmaz. Sigma-delta ($\Sigma\Delta$) quantization and finite frames. *IEEE Trans. Inform. Theory*, 52(5):1990–2005, May 2006.

[2] J. Blum, M. Lammers, A.M. Powell, and Ö. Yılmaz. Sobolev duals in frame theory and Sigma-Delta quantization. *J. Fourier Anal. Appl.* Accepted.

[3] B.G. Bodmann, V.I. Paulsen, and S.A. Abdulbaki. Smooth Frame-Path Termination for Higher Order Sigma-Delta Quantization. *J. Fourier Anal. Appl.*, 13(3):285–307, 2007.

[4] P. Boufounos and R.G. Baraniuk. 1-bit compressive sensing. In *42nd annual Conference on Information Sciences and Systems (CISS)*, pages 19–21.

[5] P. Boufounos and R.G. Baraniuk. Sigma delta quantization for compressive sensing. In *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, volume 6701, page 4. Citeseer, 2007.

[6] E.J. Candès. Compressive sampling. In *International Congress of Mathematicians. Vol. III*, pages 1433–1452. Eur. Math. Soc., Zürich, 2006.

[7] E.J. Candès, J. Romberg, and T. Tao. Signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math.*, 59(8):1207–1223, 2005.

[8] E.J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inform. Theory*, 52(2):489–509, 2006.

[9] W. Dai, H.V. Pham, and O. Milenkovic. Quantized compressive sensing. `arXiv:0901.0749 [cs.IT]`, 2009.

[10] I. Daubechies and R. DeVore. Approximating a bandlimited function using very coarsely quantized data: A family of stable sigma-delta modulators of arbitrary order. *Ann. of Math.*, 158(2):679–710, 2003.

[11] D.L. Donoho. Compressed sensing. *IEEE Trans. Inform. Theory*, 52(4):1289–1306, 2006.

[12] D.L. Donoho. For most large underdetermined systems of equations, the minimal l1-norm near-solution approximates the sparsest near-solution. *Comm. Pure Appl. Math.*, 59(7):907–934, 2006.

[13] V.K. Goyal, A.K. Fletcher, and S. Rangan. Compressive sampling and lossy compression. *IEEE Signal Processing Magazine*, 25(2):48–56, 2008.

[14] V.K. Goyal, M. Vetterli, and N.T. Thao. Quantized overcomplete expansions in $\mathbb{R}^N$: analysis, synthesis, and algorithms. *IEEE Trans. Inform. Theory*, 44(1):16–31, 1998.

[15] C.S. Güntürk. One-bit sigma-delta quantization with exponential accuracy. *Comm. Pure Appl. Math.*, 56(11):1608–1630, 2003.

[16] C.S. Güntürk, A.M. Powell, R. Saab, and Ö Yılmaz. Sobolev duals for random frames and Sigma-Delta quantization of compressed sensing measurements. `arXiv:1002.0182v1 [cs.IT]`, 2010.

[17] L. Jacques, D.K. Hammond, and M.J. Fadili. Dequantizing compressed sensing: When oversampling and non-gaussian constraints combine. `arXiv:0902.2367 [math.OC]`, 2009.

[18] M. Lammers, A.M. Powell, and Ö. Yılmaz. Alternative dual frames for digital-to-analog conversion in Sigma-Delta quantization. *Adv. Comput. Math.*, 32(1):73–102, 2010.

[19] J. Laska, S. Kirolos, Y. Massoud, R. Baraniuk, A. Gilbert, M. Iwen, and M. Strauss. Random sampling for analog-to-information conversion of wideband signals. In *Proc. IEEE Dallas Circuits and Systems Workshop (DCAS)*, 2006.

[20] J.N. Laska, P.T. Boufounos, M.A. Davenport, and R.G. Baraniuk. Democracy in action: Quantization, saturation, and compressive sensing. *Preprint*, 2009.

[21] S.R. Norsworthy, R.Schreier, and G.C. Temes, editors. *Delta-Sigma Data Converters*. IEEE Press, 1997.

[22] J.A. Tropp, J.N. Laska, M.F. Duarte, J.K. Romberg, and R.G. Baraniuk. Beyond Nyquist: Efficient sampling of sparse, bandlimited signals. *IEEE Trans. Inform. Theory*, 2009.

[23] A. Zymnis, S. Boyd, and E.J. Candès. Compressed sensing with quantized measurements. 2009. Submitted.