Sobolev Duals for Random Frames and $\Sigma\Delta$ Quantization of Compressed Sensing Measurements

C.S. Güntürk[*], M. Lammers[†], A.M. Powell[‡], R. Saab[§], Ö. Yılmaz[§]

October 12, 2010

Abstract

Quantization of compressed sensing measurements is typically justified by the robust recovery results of Candès, Romberg and Tao, and of Donoho. These results guarantee that if a uniform quantizer of step size δ is used to quantize m measurements $y = \Phi x$ of a k-sparse signal $x \in \mathbb{R}^N$, where Φ satisfies the restricted isometry property, then the approximate recovery $x^\#$ via ℓ_1-minimization is within $O(\delta)$ of x. The simplest and commonly assumed approach is to quantize each measurement independently. In this paper, we show that if instead an rth order $\Sigma\Delta$ quantization scheme with the same output alphabet is used to quantize y, then there is an alternative recovery method via Sobolev dual frames which guarantees a reduction of the approximation error by a factor of $(m/k)^{(r-1/2)\alpha}$ for any $0 < \alpha < 1$, if $m \gtrsim_r k(\log N)^{1/(1-\alpha)}$. The result holds with high probability on the initial draw of the measurement matrix Φ from the Gaussian distribution, and uniformly for all k-sparse signals x that satisfy a mild size condition on their supports.

1 Introduction

Compressed sensing is concerned with when and how sparse signals can be recovered exactly or approximately from few linear measurements [9, 11, 15]. Let Φ be an $m \times N$ matrix providing the measurements where $m \ll N$, and let Σ_k^N denote the space of k-sparse signals in $\mathbb{R}^N$, $k < m$. A standard objective, after a suitable change of basis, is that the mapping $x \mapsto y = \Phi x$ be injective on Σ_k^N. Minimal conditions on Φ that offer such a guarantee are well-known (see, e.g. [12]) and require at least that $m \geq 2k$. On the other hand, under stricter conditions on Φ, such as the restricted isometry property (RIP), one can recover sparse vectors from their measurements by numerically efficient methods, such as ℓ^1-minimization. Moreover, the recovery will also be robust when the measurements are corrupted [10], cf. [16]; if $\hat{y} = \Phi x + e$ where e is any vector such that $\|e\|_2 \leq \epsilon$, then the solution $x^\#$ of the optimization problem

$$\min \|z\|_1 \text{ subject to } \|\Phi z - \hat{y}\|_2 \leq \epsilon \tag{1}$$

will satisfy $\|x - x^\#\|_2 \leq C_1\epsilon$ for some constant $C_1 = C_1(\Phi)$.

[*]Courant Institute of Mathematical Sciences, New York University.
[†]University of North Carolina, Wilmington.
[‡]Vanderbilt University.
[§]University of British Columbia.

The price paid for these stronger recovery guarantees is the somewhat smaller range of values available for the dimensional parameters m, k, and N. While there are some explicit (deterministic) constructions of measurement matrices with stable recovery guarantees, best results (widest range of values) have been found via random families of matrices. For example, if the entries of Φ are independently sampled from the Gaussian distribution $\mathcal{N}(0, \frac{1}{m})$, then with high probability, Φ will satisfy the RIP (with a suitable set of parameters) if $m \sim k \log(\frac{N}{k})$. Significant effort has been put on understanding the phase transition behavior of the RIP parameters for other random families, e.g., Bernoulli matrices and random Fourier samplers.

Quantization for compressed sensing measurements

The robust recovery result mentioned above is essential to the practicability of compressed sensing, especially from an analog-to-digital conversion point of view. If a discrete alphabet $\mathcal{A}$, such as $\mathcal{A} = \delta\mathbb{Z}$ for some step size $\delta > 0$, is to be employed to replace each measurement y_j with a quantized measurement $q_j := \hat{y}_j \in \mathcal{A}$, then the temptation, in light of this result, would be to minimize $\|e\|_2 = \|y - q\|_2$ over $q \in \mathcal{A}^m$. This immediately reduces to minimizing $|y_j - q_j|$ for each j, i.e., quantizing each measurement separately to the nearest element of $\mathcal{A}$, which is called memoryless scalar quantization (MSQ), also known as pulse code modulation (PCM).

Since $\|y - q\|_2 \leq \frac{1}{2}\delta\sqrt{m}$, the robust recovery result guarantees that

$$\|x - x^{\#}_{\mathrm{MSQ}}\|_2 \lesssim \delta\sqrt{m}. \tag{2}$$

Note that (2) is somewhat surprising as the reconstruction error bound does not improve by increasing the number of (quantized) measurements; on the contrary, it deteriorates. However, the $\sqrt{m}$ term is an artifact of our choice of normalization for the measurement matrix Φ. In the compressed sensing literature, it is conventional to normalize a (random) measurement matrix Φ so that it has unit-norm columns (in expectation). This is the necessary scaling to achieve isometry, and for random matrices it ensures that $\mathbb{E}\|\Phi x\|^2 = \|x\|^2$ for any x, which then leads to the RIP through concentration of measure and finally to the robust recovery result stated in (1). On the other hand, this normalization imposes an m-dependent dynamic range for the measurements which scales as $1/\sqrt{m}$, hence it is not fair to use the same value δ for the quantizer resolution as m increases. In this paper, we investigate the dependence of the recovery error on the number of quantized measurements where δ is independent of m. A fair assessment of this dependence can be made only if the dynamic range of each measurement is kept constant while increasing the number of measurements. This suggests that the natural normalization in our setting should ensure that the entries of the measurement matrix Φ are independent of m. In the specific case of random matrices, we can achieve this by choosing the entries of Φ standard i.i.d. random variables, e.g. according to $\mathcal{N}(0, 1)$. With this normalization of Φ, the robust recovery result of [10] stated at the beginning now becomes

$$\|\hat{y} - y\|_2 \leq \epsilon \implies \|x - x^{\#}\|_2 \leq \frac{C_1}{\sqrt{m}}\epsilon, \tag{3}$$

which also replaces (2) with

$$\|x - x^{\#}_{\mathrm{MSQ}}\|_2 \lesssim \delta. \tag{4}$$

As expected, this error bound does not deteriorate with m anymore. In this paper, we will adopt this normalization convention and work with the standard Gaussian distribution $\mathcal{N}(0, 1)$ when

quantization is involved, but also use the more typical normalization $\mathcal{N}(0, 1/m)$ for certain concentration estimates that will be derived in Section 3. The transition between these two conventions is of course trivial.

The above analysis of quantization error is based on MSQ, which involves separate (independent) quantization of each measurement. The vast logarithmic reduction of the ambient dimension N would seem to suggest that this strategy 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 MSQ-quantized measurements and variants thereof, e.g., [7,13,18,21,24,36]. In addition to [8], which uses $\Sigma\Delta$ modulation to quantize x *before* the random measurements are made, the only exceptions we are aware of are [27, 31], where the authors model the sparse vectors probabilistically and construct non-uniform scalar quantizers that minimize the quantization error among all memoryless quantizers provided that the sparse vectors obey some probabilistic model and that the recovery is done with the lasso formulation (see [32]) of (1).

On the other hand, it is clear that if (once) the support of the signal is known (recovered), then the m measurements that have been taken are highly redundant compared to the maximum k degrees of freedom that the signal has on its support. At this point, the signal may be considered *oversampled*. However, the error bound (4) does not offer an improvement of reconstruction accuracy, even if additional samples become available. (The RIP parameters of Φ are likely to improve as m increases, but this does not seem to reflect on the implicit constant factor in (4) satisfactorily.) This is contrary to the conventional wisdom in the theory and practice of oversampled quantization in A/D conversion where reconstruction error decreases as the sampling rate increases, especially with the use of quantization algorithms specially geared for the reconstruction procedure. The main goal of this paper is to show how this can be done in the compressed sensing setting as well.

Quantization for oversampled data

Methods of quantization have long been studied for oversampled data conversion. Sigma-delta ($\Sigma\Delta$) quantization (modulation), for instance, is the dominant method of A/D conversion for audio signals and relies heavily on oversampling, see [14, 20, 26]. In this setting, oversampling is typically exploited to employ very coarse quantization (e.g., 1 bit/sample), however, the working principle of $\Sigma\Delta$ quantization is applicable to any quantization alphabet. In fact, it is more natural to consider $\Sigma\Delta$ quantization as a "noise[a] shaping" method, for it seeks a quantized signal (q_j) by a recursive procedure to push the quantization error signal $y - q$ towards an unoccupied portion of the signal spectrum. In the case of bandlimited signals, this would correspond to high frequency bands.

As the canonical example, the standard first-order $\Sigma\Delta$ quantizer with input (y_j) computes a bounded solution (u_j) to the difference equation

$$(\Delta u)_j := u_j - u_{j-1} = y_j - q_j. \tag{5}$$

This can be achieved recursively by choosing, for example,

$$q_j = \arg\min_{p \in \mathcal{A}} |u_{j-1} + y_j - p|. \tag{6}$$

[a]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.

Since the reconstruction of oversampled bandlimited signals can be achieved with a low-pass filter φ that can also be arranged to be well-localized in time, the reconstruction error $\varphi * (y-q) = \Delta\varphi * u$ becomes small due to the smoothness of φ. It turns out that, with this procedure, the reconstruction error is reduced by a factor of the oversampling ratio λ, defined to be the ratio of the actual sampling rate to the bandwidth of φ.

This principle can be iterated to set up higher-order $\Sigma\Delta$ quantization schemes. It is well-known that a reconstruction accuracy of order $O(\lambda^{-r})$ can be achieved (in the supremum norm) if a bounded solution to the equation $\Delta^r u = y - q$ can be found [14] (here, $r \in \mathbb{N}$ is the order of the associated $\Sigma\Delta$ scheme). The boundedness of u is important for practical implementation, but it is also important for the error bound. The implicit constant in this bound depends on r as well as $\|u\|_\infty$. Fine analyses of carefully designed schemes have shown that optimizing the order can even yield exponential accuracy $O(e^{-c\lambda})$ for fixed sized finite alphabets $\mathcal{A}$ (see [20]), which is optimal[b] apart from the value of the constant c.

The above formulation of noise-shaping for oversampled data conversion generalizes naturally to the problem of quantization of arbitrary frame expansions, e.g., [3]. Specifically, we will consider finite frames in $\mathbb{R}^k$. A collection $(e_j)_1^m$ in $\mathbb{R}^k$ is a *frame for* $\mathbb{R}^k$ with frame bounds $0 < A \le B < \infty$ if

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

Suppose that we are given an input signal x and an analysis frame $(e_i)_1^m$ of size m in $\mathbb{R}^k$. We can represent the frame vectors as the rows of a full-rank $m \times k$ matrix E, the sampling operator. The input sequence y to be quantized will simply be the frame coefficients, i.e., $y = Ex$. Similarly, let us consider a corresponding synthesis frame $(f_j)_1^m$. We stack these frame vectors along the columns of a $k \times m$ matrix F, the reconstruction operator, which is then a left inverse of E, i.e., $FE = I$. A quantization algorithm will replace the coefficient sequence y with its quantization $q \in \mathcal{A}^m$, which will then yield an approximate reconstruction $\hat{x}$ using the synthesis frame via $\hat{x} = Fq$. Typically $(y - \mathcal{A}^m) \cap \mathrm{Ker}(F) = \emptyset$, so we have $\hat{x} \neq x$. The reconstruction error is given by

$$x - \hat{x} = F(y - q), \tag{7}$$

and the goal of noise shaping amounts to arranging q in such a way that $y - q$ is close to $\mathrm{Ker}(F)$.

If the sequence $(f_j)_1^m$ of dual frame vectors 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., [6, 22]. 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 [5, 23] to use special alternative dual frames, called Sobolev dual frames, that are naturally adapted to $\Sigma\Delta$ quantization. It is shown in [5] (see also Section 2) that for any frame E, if a standard rth order $\Sigma\Delta$ quantization algorithm with alphabet $\mathcal{A} = \delta\mathbb{Z}$ is used to compute $q := q_{\Sigma\Delta}$, then with an rth order Sobolev dual frame $F := F_{\mathrm{Sob},r}$ and $\hat{x}_{\Sigma\Delta} := F_{\mathrm{Sob},r} q_{\Sigma\Delta}$, the reconstruction error obeys the bound

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

[b]The optimality remark does not apply to the case of infinite quantization alphabet $\mathcal{A} = \delta\mathbb{Z}$ because depending on the coding algorithm, the (effective) bit-rate can still be unbounded. Indeed, arbitrarily small reconstruction error can be achieved with a (sufficiently large) fixed value of λ and a fixed value of δ by increasing the order r of the $\Sigma\Delta$ modulator. This would not work with a finite alphabet because the modulator will eventually become unstable. In practice, almost all schemes need to use some form of finite quantization alphabet.

where D is the $m \times m$ difference matrix defined by

$$D_{ij} := \begin{cases} 1, & \text{if } i = j, \\ -1, & \text{if } i = j+1, \\ 0, & \text{otherwise,} \end{cases} \tag{9}$$

and $\sigma_{\min}(D^{-r}E)$ stands for the smallest singular value of $D^{-r}E$.

Contributions

For the compressed sensing application that is the subject of this paper, E will simply be a submatrix of the measurement matrix Φ, hence it may have been found by sampling an i.i.d. random variable. Minimum singular values of random matrices with i.i.d. entries have been studied extensively in the mathematical literature. For an $m \times k$ random matrix E with $m \geq k$ and with i.i.d. entries sampled from a sub-Gaussian distribution with zero mean and unit variance,[c] one has

$$\sigma_{\min}(E) \gtrsim \sqrt{m} - \sqrt{k} \tag{10}$$

with high probability [29]. Note that in general $D^{-r}E$ would not have i.i.d. entries. A naive lower bound for $\sigma_{\min}(D^{-r}E)$ would be $\sigma_{\min}(D^{-r})\sigma_{\min}(E)$. However (see Proposition 3.1), $\sigma_{\min}(D^{-r})$ satisfies

$$\sigma_{\min}(D^{-r}) \asymp_r 1, \tag{11}$$

and therefore this naive product bound yields no improvement on the reconstruction error for $\Sigma\Delta$-quantized measurements over the bound (4) for MSQ-quantized ones. In fact, the true behavior of $\sigma_{\min}(D^{-r}E)$ turns out to be drastically different and is described in Theorem A, one of our main results (see also Theorem 3.7).

For simplicity, we shall work with standard i.i.d. Gaussian variables for the entries of E. In analogy with our earlier notation, we define the "oversampling ratio" λ of the frame E by

$$\lambda := \frac{m}{k}. \tag{12}$$

Theorem A. *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 $1 - \exp(-c'm\lambda^{-\alpha})$,*

$$\sigma_{\min}(D^{-r}E) \gtrsim_r \lambda^{\alpha(r-\frac{1}{2})}\sqrt{m}, \tag{13}$$

which yields the reconstruction error bound

$$\|x - \hat{x}_{\Sigma\Delta}\|_2 \lesssim_r \lambda^{-\alpha(r-\frac{1}{2})}\delta. \tag{14}$$

While the kind of decay in this error bound is familiar to $\Sigma\Delta$ modulation, the domain of applicability of this result is rather surprising. Previously, the only setting in which this type of approximation accuracy could be achieved (with or without Sobolev duals) was the case of highly structured frames (e.g. when the frame vectors are found by sampling along a piecewise smooth frame path). Theorem A shows that such an accuracy is obtained even when the analysis frame is a random Gaussian matrix, provided the reconstruction is done via Sobolev duals.

[c] As mentioned earlier, we do not normalize the measurement matrix Φ in the quantization setting.

In the compressed sensing setting, one needs (13) to be uniform for all the frames E that are found by selecting k columns of Φ at a time. The proof of Theorem A extends in a straightforward manner to allow for this using a standard "union bound" argument, provided λ is known to be slightly larger. More precisely, if Φ is an $m \times N$ matrix whose entries are i.i.d. according to $\mathcal{N}(0,1)$, and if $\lambda := m/k \geq c(\log N)^{1/(1-\alpha)}$, then (13) holds for all $E = \Phi_T$ with $\#T \leq k$ with the same type of probability bound (with new constants). This result can be utilized to improve the reconstruction accuracy of a sparse signal x from its $\Sigma\Delta$-quantized compressed sensing measurements if the support T of x is known. This is because if T is known, Φ_T is known, and its Sobolev dual can be found and used in the reconstruction. On the other hand, for most signals, recovering the exact or approximate support is already nearly guaranteed by the robust recovery result shown in (3) together with the stability of the associated $\Sigma\Delta$ quantizer. For example, a simple sufficient condition for full recovery of the support is that all the $|x_j|$ for $j \in T$ be larger than $C\|y - q_{\Sigma\Delta}\|_2$ for a suitable constant C. A precise version of this condition is stated in Theorem B.

In light of all these results, we propose $\Sigma\Delta$ quantization as a more effective alternative of MSQ (independent quantization) for compressed sensing. With high probability on the measurement matrix, a significant improvement of the reconstruction accuracy of sparse signals can be achieved through a two-stage recovery procedure:

1. **Coarse recovery:** ℓ_1-minimization (or any other robust recovery procedure) applied to $q_{\Sigma\Delta}$ yields an initial, "coarse" approximation $x^{\#}$ of x, and in particular, the exact (or approximate) support T of x.

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

Combining all these, our second main theorem follows (also see Theorem 4.2):

Theorem B. *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*

$$\|x - \hat{x}_{\Sigma\Delta}\|_2 \lesssim_r \lambda^{-\alpha(r-\frac{1}{2})}\delta. \tag{15}$$

To put this result in perspective, note that the approximation error given in (15) decays as the "redundancy" $\lambda = \frac{m}{k}$ increases. In fact, by using an arbitrarily high order $\Sigma\Delta$ scheme, we can make this decay faster than any power law (albeit with higher constants). Note that such a decay is not observed in the reconstruction error bound for MSQ given in (4). Of course, one could argue that these upper bounds may not reflect the actual behavior of the error. However, in the setting of frame quantization the performance of MSQ is well investigated. In particular, 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 we obtain $q_{\mathrm{MSQ}}(x)$ by quantizing the entries of $y = Ex$ using MSQ with alphabet $\mathcal{A} = \delta\mathbb{Z}$. Let Δ_{opt} be an optimal decoder. Then, Goyal et al. show in [19] that

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

where $\lambda = m/k$ and the expectation is with respect a probability measure on x that is, for example, absolutely continuous. This lower bound limits the extent to which one can improve the reconstruction by means of alternative reconstruction algorithms from MSQ-quantized compressed sensing

measurements. On the other hand, setting, for example, $\alpha = 3/4$ in Theorem B we observe that if we use a second-order $\Sigma\Delta$ scheme to quantize the measurements, and if we adopt the two-stage recovery procedure proposed above, the resulting approximation will be superior to that produced optimally from MSQ-quantized measurements, provided m/k is sufficiently large. Of course, one might argue that it is contrary to the philosophy of compressed sensing to consider cases where λ is large – we are, after all, using compressed sensing to reduce the sampling rate. However, as we also mentioned above, redundancy is built into the compressed sensing paradigm: we need $\lambda \gtrsim \log N$ for the robust recovery result to hold, and we might as well utilize this redundancy when designing effective quantizers in this setup.

It is possible to imagine more sophisticated and more effective quantization and recovery algorithms for compressed sensing. However using $\Sigma\Delta$ quantization has a number of appealing features:

- It produces **more accurate** approximations than any known quantization scheme in this setting (even when sophisticated recovery algorithms are employed).
- It is **modular** in the sense that if the fine recovery stage is not available or practical to implement, then the standard (coarse) recovery procedure can still be applied as is.
- It is **progressive** in the sense that if new measurements arrive (in any given order), noise shaping can be continued on these measurements as long as the state of the system (r real values for an rth order scheme) has been stored.
- It is **universal** in the sense that it uses no information about the measurement matrix or the signal.

The paper is organized as follows. We review the basics of $\Sigma\Delta$ quantization and Sobolev duals in frame theory in Section 2, followed by the reconstruction error bounds for random Gaussian frames in Section 3. We then present the specifics of our proposed quantization and recovery algorithm for compressed sensing in Section 4. We present our numerical experiments in Section 5 and conclude with extensions to more general settings in Section 6.

2 Background on $\Sigma\Delta$ quantization of frame expansions

$\Sigma\Delta$ quantization

The governing equation of a standard rth order $\Sigma\Delta$ quantization scheme with input $y = (y_j)$ and output $q = (q_j)$ is

$$(\Delta^r u)_j = y_j - q_j, \quad j = 1, 2, \ldots, \tag{16}$$

where the $q_j \in \mathcal{A}$ are chosen according to some *quantization rule*, typically a predetermined function of the input and past state variables $(u_l)_{l<j}$, given by

$$q_j = Q(u_{j-1}, \ldots, u_{j-T}, y_j, \ldots, y_{j-S}). \tag{17}$$

Not all $\Sigma\Delta$ quantization schemes are presented (or implemented) in this canonical form, but they all can be rewritten as such for an appropriate choice of r and u. We shall not be concerned with the specifics of the mapping Q, except that it needs to be so that u is bounded. The smaller the size of the alphabet $\mathcal{A}$ gets relative to r, the harder it is to guarantee this property. The extreme case

is 1-bit quantization, i.e., $|\mathcal{A}| = 2$, which is typically the most challenging setting. We will not be working in this case. In fact, for our purposes, $\mathcal{A}$ will in general have to be sufficiently fine to allow for the recovery of the support of sparse signals. In order to avoid technical difficulties, we shall work with the infinite alphabet $\mathcal{A} = \delta\mathbb{Z}$, but also note that only a finite portion of this alphabet will be used for bounded signals. A standard quantization rule that has this "boundedness" property is given by the greedy rule which minimizes $|u_j|$ given $u_{j-1}, \ldots, u_{j-r}$ and y_j, i.e.,

$$q_j = \arg\min_{a\in\mathcal{A}} \Big| \sum_{i=1}^{r} (-1)^{i-1} \binom{r}{i} u_{j-i} + y_j - a \Big|. \tag{18}$$

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

$$\|u\|_\infty \lesssim \delta \text{ and } \|y - q\|_\infty \lesssim_r \delta. \tag{19}$$

With more stringent quantization rules, the first inequality would also have an r-dependent constant. In fact, it is known that for quantization rules with a 1-bit alphabet, this constant will be as large as $O(r^r)$, e.g., see [14,20]. In this paper, unless otherwise stated, we shall be working with the greedy quantization rule of (18).

The initial conditions of the recursion in (16) can be set arbitrarily, however for the purposes of this paper it will be convenient to set them equal to zero. With $u_{-r+1} = \cdots = u_0 = 0$, and $j = 1, \ldots, m$, the difference equation (16) can be rewritten as a matrix equation

$$D^r u = y - q, \tag{20}$$

where D is as in (9).

As before, we assume E is an $m \times k$ matrix whose rows form the analysis frame and F is a $k \times m$ left inverse of E whose columns form the dual (synthesis) frame. Given any $x \in \mathbb{R}^k$, we set $y = Ex$, and denote its rth order $\Sigma\Delta$ quantization by $q_{\Sigma\Delta}$ and its reconstruction by $\hat{x}_{\Sigma\Delta} := Fq_{\Sigma\Delta}$. Substituting (20) into (7), we obtain the error expression

$$x - \hat{x} = FD^r u. \tag{21}$$

With this expression, $\|x - \hat{x}\|$ can be bounded for any norm $\|\cdot\|$ simply as

$$\|x - \hat{x}\| \leq \|u\|_\infty \sum_{j=1}^{m} \|(FD^r)_j\|. \tag{22}$$

Here $(FD^r)_j$ is the jth column of FD^r. This bound is also valid in infinite dimensions, and in fact has been used extensively in the mathematical treatment of oversampled A/D conversion of bandlimited functions.

For $r = 1$ and $\|\cdot\| = \|\cdot\|_2$, the sum term on the right hand side of (22) motivated the study of the so-called *frame variation* defined by

$$V(F) := \sum_{j=1}^{m} \|f_j - f_{j+1}\|_2, \tag{23}$$

where (f_j) are the columns of F, and one defines $f_{m+1} = 0$; see [3]. Higher-order frame variations to be used with higher-order $\Sigma\Delta$ schemes are defined similarly; see [2]. Frames (analysis as well as synthesis) that are obtained via uniform sampling a smooth curve in $\mathbb{R}^k$ (so-called *frame path*) are typical in many settings. However, the "frame variation bound" is useful in finite dimensions only when the frame path terminates smoothly. Otherwise, it typically does not provide higher-order reconstruction accuracy (see [2] for an exception). Designing smoothly terminating frames can be technically challenging, e.g., [6].

Sobolev duals

Recently, a more straightforward approach was proposed in [22] for the design of (alternate) duals of finite frames for $\Sigma\Delta$ quantization. Here, one instead considers the operator norm of FD^r on ℓ_2 and the corresponding bound

$$\|x - \hat{x}\|_2 \leq \|FD^r\|_{\text{op}} \|u\|_2. \tag{24}$$

Note that this bound is not available in the infinite dimensional setting of bandlimited functions due to the fact that u is typically not in ℓ_2. It is now natural to minimize $\|FD^r\|_{\text{op}}$ over all dual frames of a given analysis frame E. These frames, introduced in [5], have been called Sobolev duals, in analogy with ℓ_2-type Sobolev (semi)norms.

$\Sigma\Delta$ quantization algorithms are normally designed for analog circuit operation, so they control $\|u\|_\infty$, which would control $\|u\|_2$ only in a suboptimal way. However, it turns out that there are important advantages in working with the ℓ_2 norm in the analysis. The first advantage is that Sobolev duals are readily available by an explicit formula. The solution $F_{\text{sob},r}$ of the optimization problem

$$\min_F \|FD^r\|_{\text{op}} \text{ subject to } FE = I \tag{25}$$

is given by the matrix equation

$$F_{\text{sob},r} D^r = (D^{-r}E)^\dagger, \tag{26}$$

where † stands for the Moore-Penrose inversion operator, which, in our case, is given by $E^\dagger := (E^*E)^{-1}E^*$. Note that for $r = 0$ (i.e., no noise-shaping, or MSQ), one simply obtains $F = E^\dagger$, the canonical dual frame of E.

The second advantage of this approach is its analytic tractability. Plugging (26) into (24), it immediately follows that

$$\|x - \hat{x}\|_2 \leq \|(D^{-r}E)^\dagger\|_{\text{op}} \|u\|_2 = \frac{1}{\sigma_{\min}(D^{-r}E)} \|u\|_2, \tag{27}$$

where $\sigma_{\min}(D^{-r}E)$ stands for the smallest singular value of $D^{-r}E$. There exist highly developed methods for estimating spectral norms and more generally singular values of matrices, especially in the random setting, as we shall employ in this paper.

3 Reconstruction error bound for random frames

In what follows, $\sigma_j(A)$ will denote the jth largest singular value of the matrix A. Similarly, $\lambda_j(B)$ will denote the jth largest eigenvalue of the Hermitian matrix B. Hence, we have $\sigma_j(A) = \sqrt{\lambda_j(A^*A)}$. We will also use the notation $\Sigma(A)$ for the diagonal matrix of singular values of A,

with the convention $(\Sigma(A))_{jj} = \sigma_j(A)$. All matrices in our discussion will be real valued and the Hermitian conjugate reduces to the transpose.

We have seen that the main object of interest for the reconstruction error bound is $\sigma_{\min}(D^{-r}E)$ for a random frame E. Let H be a square matrix. The first observation we make is that when E is i.i.d. Gaussian, the distribution of $\Sigma(HE)$ is the same as the distribution of $\Sigma(\Sigma(H)E)$. To see this, let $U\Sigma(H)V^*$ be the singular value decomposition of H where U and V are unitary matrices. Then $HE = U\Sigma(H)V^*E$. Since the unitary transformation U does not alter singular values, we have $\Sigma(HE) = \Sigma(\Sigma(H)V^*E)$, and because of the unitary invariance of the i.i.d. Gaussian measure, the matrix $\widetilde{E} := V^*E$ has the same distribution as E, hence the claim. Therefore it suffices to study the singular values of $\Sigma(H)E$. In our case, $H = D^{-r}$ and we first need information on the deterministic object $\Sigma(D^{-r})$. The following result will be sufficient for our purposes:

Proposition 3.1. *Let r be any positive integer and D be as in* (9)*. There are positive numerical constants $c_1(r)$ and $c_2(r)$, independent of m, such that*

$$c_1(r)\Big(\frac{m}{j}\Big)^r \leq \sigma_j(D^{-r}) \leq c_2(r)\Big(\frac{m}{j}\Big)^r, \quad j = 1, \dots, m. \tag{28}$$

The somewhat standard proof of this result via the theory of Toeplitz matrices is given in Appendix A.

The remainder of this section is dedicated to proving the main results of the paper. Let E be an $m \times k$ matrix with i.i.d. entries drawn from $\mathcal{N}(0, 1/m)$. In Section 3.1, we show that if m/k is large enough, $\sigma_{\min}(D^{-r}E) \gtrsim (m/k)^{\alpha(r-1/2)}$, with high probability (Theorem 3.7, equivalently Theorem A). To obtain this result, we use a property of Gaussian vectors (Proposition 3.2) and a standard ϵ-net argument to upper bound the norm $\|SE\|_{\ell_2^m \to \ell_2^k}$ with high probability for an arbitrary positive diagonal matrix S (Lemma 3.3). Next, we again use properties of Gaussian vectors (Proposition 3.4 and Corollary 3.5) and an ϵ-net argument to lower bound the least singular value of SE, with high probability, under mild conditions on S (Theorem 3.6). By a careful choice of parameters in the above results and for $S = \Sigma(D^{-r})$, we then obtain Theorem 3.7.

In Section 3.2, we generalize Theorem 3.7 to the compressed sensing setting. In particular, we use a standard union bound to show that if m/k is large enough, then with high probability, $\sigma_{\min}(D^{-r}E) \gtrsim (m/k)^{\alpha(r-1/2)}$, for any $m \times k$ sub-matrix E, of an $m \times N$ random matrix Φ whose entries are drawn i.i.d. from $\mathcal{N}(0, 1/m)$ (Theorem 3.8).

3.1 Lower bound for $\sigma_{\min}(D^{-r}E)$

In light of the above discussion, the distribution of $\sigma_{\min}(D^{-r}E)$ is the same as that of

$$\inf_{\|x\|_2=1} \|\Sigma(D^{-r})Ex\|_2. \tag{29}$$

We replace $\Sigma(D^{-r})$ with an arbitrary diagonal matrix S with $S_{jj} =: s_j > 0$. The first two results will concern upper bounds for the norm of independent but non-identically distributed Gaussian vectors. They are rather standard, but we include them for the definiteness of our discussion when they will be used later.

Proposition 3.2. *Let $\xi \sim \mathcal{N}(0, \frac{1}{m}\mathrm{I}_m)$. For any $\Theta > 1$,*

$$\mathbb{P}\left(\sum_{j=1}^m s_j^2\xi_j^2 > \Theta\|s\|_\infty^2\right) \leq \Theta^{m/2}e^{-(\Theta-1)m/2}. \tag{30}$$

Proof. Since $s_j \le \|s\|_\infty$ for all j, we have

$$\mathbb{P}\left(\sum_{j=1}^m s_j^2\xi_j^2 > \Theta\|s\|_\infty^2\right) \le \mathbb{P}\left(\sum_{j=1}^m \xi_j^2 > \Theta\right). \tag{31}$$

This bound is the (standard) Gaussian measure of the complement of a sphere of radius $\sqrt{m\Theta}$ and can be estimated very accurately. We use a simple approach via

$$\begin{aligned}\mathbb{P}\left(\sum_{j=1}^m \xi_j^2 > \Theta\right) &\le \min_{\lambda\ge 0} \int_{\mathbb{R}^m} e^{-(\Theta-\sum_{j=1}^m x_j^2)\lambda/2} \prod_{j=1}^m e^{-mx_j^2/2} \frac{\mathrm{d}x_j}{\sqrt{2\pi/m}} \\ &= \min_{\lambda\ge 0} e^{-\lambda\Theta/2}(1-\lambda/m)^{-m/2} \\ &= \Theta^{m/2}e^{-(\Theta-1)m/2}, \end{aligned} \tag{32}$$

where in the last step we set $\lambda = m(1-\Theta^{-1})$. □

Lemma 3.3. *Let E be an $m\times k$ random matrix whose entries are i.i.d. $\mathcal{N}(0,\frac{1}{m})$. For any $\Theta > 1$, consider the event*

$$\mathcal{E} := \left\{\|SE\|_{\ell_2^k\to\ell_2^m} \le 2\sqrt{\Theta}\|s\|_\infty\right\}.$$

Then

$$\mathbb{P}(\mathcal{E}^c) \le 5^k\Theta^{m/2}e^{-(\Theta-1)m/2}.$$

Proof. We follow the same approach as in [1]. The maximum number of ρ-distinguishable points on the unit sphere in $\mathbb{R}^k$ is at most $(\frac{2}{\rho}+1)^k$. (This follows by a volume argument[d] as in e.g., [25, p.487].) Fix a maximal set Q of $\frac{1}{2}$-distinguishable points of the unit sphere in $\mathbb{R}^k$ with $\#Q \le 5^k$. Since Q is maximal, it is a $\frac{1}{2}$-net for the unit sphere. For each $q\in Q$, consider $\xi_j = (Eq)_j$, $j=1,\dots,m$. Then $\xi\sim\mathcal{N}(0,\frac{1}{m}\mathrm{I}_m)$. As before, we have

$$\|SEq\|_2^2 = \sum_{j=1}^m s_j^2\xi_j^2.$$

Let $\mathcal{E}(Q)$ be the event $\left\{\|SEq\|_2 \le \sqrt{\Theta}\|s\|_\infty,\ \ \forall q\in Q\right\}$. Then, by Proposition 3.2, we have the union bound

$$\mathbb{P}(\mathcal{E}(Q)^c) \le 5^k\Theta^{m/2}e^{-(\Theta-1)m/2}. \tag{33}$$

Assume the event $\mathcal{E}(Q)$, and let $M = \|SE\|_{\ell_2^k\to\ell_2^m}$. For each $\|x\|_2 = 1$, there is $q\in Q$ with $\|q-x\|_2 \le 1/2$, hence

$$\|SEx\|_2 \le \|SEq\|_2 + \|SE(x-q)\|_2 \le \sqrt{\Theta}\|s\|_\infty + \frac{M}{2}.$$

Taking the supremum over all x on the unit sphere, we obtain

$$M \le \sqrt{\Theta}\|s\|_\infty + \frac{M}{2},$$

i.e., $\|SE\|_{\ell_2^k\to\ell_2^m} \le 2\sqrt{\Theta}\|s\|_\infty$. Therefore $\mathcal{E}(Q)\subset\mathcal{E}$, and the result follows. □

[d] Balls with radii $\rho/2$ and centers at a ρ-distinguishable set of points on the unit sphere are mutually disjoint and are all contained in the ball of radius $1+\rho/2$ centered at the origin. Hence there can be at most $(1+\rho/2)^k/(\rho/2)^k$ of them.

The following estimate concerns a lower bound for the Euclidean norm of $(s_1\xi_1, \dots, s_m\xi_m)$. It is not sharp when the s_j are identical, but it will be useful for our problem where $s_j = \sigma_j(D^{-r})$ obey a power law (see Corollary 3.5).

Proposition 3.4. *Let $\xi \sim \mathcal{N}(0, \frac{1}{m}\mathrm{I}_m)$. For any $\gamma > 0$,*

$$\mathbb{P}\left(\sum_{j=1}^m s_j^2\xi_j^2 < \gamma\right) \le \min_{1\le L\le m}\left(\frac{e\gamma m}{L}\right)^{L/2}(s_1 s_2\cdots s_L)^{-1}. \tag{34}$$

Proof. For any $t \ge 0$ and any integer $L \in \{1, \dots, m\}$, we have

$$\begin{aligned}
\mathbb{P}\left(\sum_{j=1}^m s_j^2\xi_j^2 < \gamma\right) &\le \int_{\mathbb{R}^m} e^{(\gamma - \sum_{j=1}^m s_j^2x_j^2)t/2}\prod_{j=1}^m e^{-mx_j^2/2}\frac{\mathrm{d}x_j}{\sqrt{2\pi/m}} \\
&= e^{t\gamma/2}\prod_{j=1}^m\int_{\mathbb{R}} e^{-x_j^2(m+ts_j^2)/2}\frac{\mathrm{d}x_j}{\sqrt{2\pi/m}} \\
&= e^{t\gamma/2}\prod_{j=1}^m(1+ts_j^2/m)^{-1/2} \\
&\le e^{t\gamma/2}\prod_{j=1}^L(ts_j^2/m)^{-1/2} \\
&\le e^{t\gamma/2}(m/t)^{L/2}(s_1s_2\cdots s_L)^{-1}.
\end{aligned} \tag{35}$$

For any L, we can set $t = L/\gamma$, which is the critical point of the function $t \mapsto e^{t\gamma}t^{-L}$. Since L is arbitrary, the result follows. □

Corollary 3.5. *Let $\xi \sim \mathcal{N}(0, \frac{1}{m}\mathrm{I}_m)$, r be a positive integer, and $c_1 > 0$ be such that*

$$s_j \ge c_1\left(\frac{m}{j}\right)^r, \qquad j = 1, \dots, m. \tag{36}$$

Then for any $\Lambda \ge 1$ and $m \ge \Lambda$,

$$\mathbb{P}\left(\sum_{j=1}^m s_j^2\xi_j^2 < c_1^2\Lambda^{2r-1}\right) < (60m/\Lambda)^{r/2}e^{-m(r-1/2)/\Lambda}. \tag{37}$$

Proof. By rescaling s_j, we can assume $c_1 = 1$. For any $L \in \{1, \dots, m\}$, we have

$$(s_1s_2\cdots s_L)^{-1} \le \frac{(L!)^r}{m^{rL}} < (8L)^{r/2}\left(\frac{L^r}{e^rm^r}\right)^L,$$

where we have used the coarse estimate $L! < e^{1/12L}(2\pi L)^{1/2}(L/e)^L < (8L)^{1/2}(L/e)^L$. Setting $\gamma = \Lambda^{2r-1}$ in Proposition 3.4, we obtain

$$\mathbb{P}\left(\sum_{j=1}^m s_j^2\xi_j^2 < \Lambda^{2r-1}\right) < (8L)^{r/2}\left[\left(\frac{\Lambda L}{em}\right)^L\right]^{r-1/2}. \tag{38}$$

We set $L = \lfloor \frac{m}{\Lambda} \rfloor$. Since $1 \le \Lambda \le m$, it is guaranteed that $1 \le L \le m$. Since $\Lambda L \le m$, we get

$$\left(\frac{\Lambda L}{em}\right)^L \le e^{-L} < e^{1-\frac{m}{\Lambda}}$$

Plugging this in (38) and using $8e^2 < 60$, we find

$$\mathbb{P}\left(\sum_{j=1}^m s_j^2 \xi_j^2 < \Lambda^{2r-1}\right) \quad < \quad (60m/\Lambda)^{r/2} e^{-m(r-1/2)/\Lambda}. \tag{39}$$

□

Theorem 3.6. *Let E be an $m \times k$ random matrix whose entries are i.i.d. $\mathcal{N}(0, \frac{1}{m})$, r be a positive integer, and assume that the entries s_j of the diagonal matrix S satisfy*

$$c_1 \left(\frac{m}{j}\right)^r \le s_j \le c_2 m^r, \qquad j = 1, \dots, m. \tag{40}$$

Let $\Lambda \ge 1$ be any number and assume $m \ge \Lambda$. Consider the event

$$\mathcal{F} := \left\{ \|SEx\|_2 \ge \frac{1}{2} c_1 \Lambda^{r-1/2} \|x\|_2, \ \forall x \in \mathbb{R}^k \right\}.$$

Then

$$\mathbb{P}\left(\mathcal{F}^c\right) \le 5^k e^{-m/2} + 8^r \left(17 c_2/c_1\right)^k \Lambda^{k/2} \left(\frac{m}{\Lambda}\right)^{r(k+1/2)} e^{-m(r-1/2)\Lambda}.$$

Proof. Consider a ρ-net $\widetilde{Q}$ of the unit sphere of $\mathbb{R}^k$ with $\#\widetilde{Q} \le \left(\frac{2}{\rho} + 1\right)^k$ where the value of $\rho < 1$ will be chosen later. Let $\widetilde{\mathcal{E}}(\widetilde{Q})$ be the event $\left\{ \|SEq\|_2 \ge c_1 \Lambda^{r-1/2}, \ \ \forall q \in \widetilde{Q} \right\}$. By Corollary 3.5, we know that

$$\mathbb{P}\left(\widetilde{\mathcal{E}}(\widetilde{Q})^c\right) \le \left(\frac{2}{\rho} + 1\right)^k \left(\frac{60m}{\Lambda}\right)^{r/2} e^{-m(r-1/2)/\Lambda}. \tag{41}$$

Let $\mathcal{E}$ be the event in Lemma 3.3 with $\Theta = 4$. Let E be any given matrix in the event $\mathcal{E} \cap \widetilde{\mathcal{E}}(\widetilde{Q})$. For each $\|x\|_2 = 1$, there is $q \in \widetilde{Q}$ with $\|q - x\|_2 \le \rho$, hence by Lemma 3.3, we have

$$\|SE(x-q)\|_2 \le 4\|s\|_\infty \|x - q\|_2 \le 4 c_2 m^r \rho.$$

Choose

$$\rho = \frac{c_1 \Lambda^{r-1/2}}{8 c_2 m^r} = \frac{c_1}{8 c_2 \sqrt{\Lambda}} \left(\frac{\Lambda}{m}\right)^r.$$

Hence

$$\|SEx\|_2 \ge \|SEq\|_2 - \|SE(x-q)\|_2 \ge c_1 \Lambda^{r-1/2} - 4 c_2 m^r \rho = \frac{1}{2} c_1 \Lambda^{r-1/2}.$$

This shows that $\mathcal{E} \cap \widetilde{\mathcal{E}}(\widetilde{Q}) \subset \mathcal{F}$. Clearly, $\rho \le 1/8$ by our choice of parameters and hence $\frac{2}{\rho} + 1 \le \frac{17}{8\rho}$. Using the probability bounds of Lemma 3.3 and (41), we have

$$\begin{aligned} \mathbb{P}\left(\mathcal{F}^c\right) &\le 5^k 4^{m/2} e^{-3m/2} + \left(\frac{17}{8\rho}\right)^k \left(\frac{60m}{\Lambda}\right)^{r/2} e^{-m(r-1/2)/\Lambda} \\ &\le 5^k e^{-m/2} + 8^r (17 c_2/c_1)^k \Lambda^{k/2} \left(\frac{m}{\Lambda}\right)^{r(k+1/2)} e^{-m(r-1/2)/\Lambda}, \end{aligned} \tag{42}$$

where we have used $2 < e$ and $\sqrt{60} < 8$ for simplification. □

The following theorem is now a direct corollary of the above estimate.

Theorem 3.7. *Let E be an $m \times k$ random matrix whose entries are i.i.d. $\mathcal{N}(0, \frac{1}{m})$, r be a positive integer, D be the difference matrix defined in (9), and the constant $c_1 = c_1(r)$ be as in Proposition 3.1. Let $0 < \alpha < 1$ be any number. Assume that*

$$\lambda := \frac{m}{k} \geq c_3 (\log m)^{1/(1-\alpha)}, \tag{43}$$

where $c_3 = c_3(r)$ is an appropriate constant. Then

$$\mathbb{P}\left(\sigma_{\min}(D^{-r}E) \geq c_1 \lambda^{\alpha(r-1/2)}\right) \geq 1 - 2e^{-c_4 m^{1-\alpha} k^{\alpha}} \tag{44}$$

for some constant $c_4 = c_4(r) > 0$.

Proof. Set $\Lambda = \lambda^{\alpha}$ in Theorem 3.6 and $S = \Sigma(D^{-r})$. We only need to show that

$$\max\left[5^k e^{-m/2}, 8^r (17c_2/c_1)^k \Lambda^{k/2} \left(\frac{m}{\Lambda}\right)^{r(k+1/2)} e^{-m(r-1/2)/\Lambda}\right] \leq e^{-c_4 m^{1-\alpha} k^{\alpha}}.$$

It suffices to show that

$$k \log 5 - m/2 \leq -c_4 m^{1-\alpha} k^{\alpha}$$

and

$$r \log 8 + k \log(17c_2/c_1) + \frac{1}{2} k \log \Lambda + r(k + \frac{1}{2}) \log(m/\Lambda) - (r - \frac{1}{2}) \frac{m}{\Lambda} \leq -c_4 m^{1-\alpha} k^{\alpha}.$$

The first inequality is easily seen to hold if $\lambda \geq \frac{\log 5}{\frac{1}{2} - c_4}$. For the second inequality, first notice that $m/\Lambda = m^{1-\alpha} k^{\alpha}$. Since $k + 1/2 \asymp k$, and $r - 1/2 \asymp r$, it is easily seen that we only need to check that

$$k \log m \leq c_5 \frac{m}{\Lambda}$$

for a sufficiently small c_5. This follows from our assumption on λ by setting $c_5 = 1/c_3^{1-\alpha}$. □

Remark. By replacing E in Theorem 3.7 with $\sqrt{m}E$, we obtain Theorem A.

3.2 Implication for compressed sensing matrices

Theorem 3.8. *Let r, D, $c_1(r)$ be as in Theorem 3.7 and Φ be an $m \times N$ random matrix whose entries are i.i.d. $\mathcal{N}(0, \frac{1}{m})$. Let $0 < \alpha < 1$ be any number and assume that*

$$\lambda := \frac{m}{k} \geq c_6 (\log N)^{1/(1-\alpha)}, \tag{45}$$

where $c_6 = c_6(r)$ is an appropriate constant. Then with probability at least $1 - 2e^{-c_7 m \lambda^{-\alpha}}$ for some $c_7 = c_7(r) > 0$, every $m \times k$ submatrix E of Φ satisfies

$$\sigma_{\min}(D^{-r}E) \geq c_1 \lambda^{\alpha(r-1/2)}. \tag{46}$$

Proof. We will choose $c_7 = c_4/2$, where c_4 is as in Theorem 3.7. The proof will follow immediately by a union bound once we show that

$$\binom{N}{k} \leq e^{\frac{1}{2}c_4 m^{1-\alpha}k^{\alpha}}.$$

Since $\binom{N}{k} \leq N^k$, it suffices to show that

$$k \log N \leq \frac{c_4}{2} m^{1-\alpha}k^{\alpha}.$$

Both this condition and the hypothesis of Theorem 3.7 will be satisfied if we choose

$$c_6 = \max(c_3, (2/c_4)^{1/(1-\alpha)}).$$

□

Remark. If Φ is a Gaussian matrix with entries i.i.d. $\mathcal{N}(0,1)$ rather than $\mathcal{N}(0, \frac{1}{m})$, Theorem 3.8 applied to $\frac{1}{\sqrt{m}}\Phi$ implies that every $m \times k$ submatrix E of Φ satisfies

$$\sigma_{\min}(D^{-r}E) \geq c_1 \lambda^{\alpha(r-1/2)}\sqrt{m}. \tag{47}$$

4 $\Sigma\Delta$ quantization of compressed sensing measurements

In this section we will assume that the conditions of Theorem 3.8 are satisfied for some $0 < \alpha < 1$ and r, and the measurement matrix Φ that is drawn from $\mathcal{N}(0,1)$ yields (47). For definiteness, we also assume that Φ admits the robust recovery constant $C_1 = 10$, i.e., the solution $x^{\#}$ of the program (1) satisfies

$$\|\hat{y} - y\|_2 \leq \epsilon \implies \|x - x^{\#}\|_2 \leq \frac{10}{\sqrt{m}}\epsilon.$$

Note that C_1 depends only on the RIP constants of Φ and is well-behaved in our setting. For more details and the admissibility of this value for C_1, see [10].

Note again that our choice of normalization for the measurement matrix Φ is different from the compressed sensing convention. As mentioned in the Introduction, it is more appropriate to work with a measurement matrix $\Phi \sim \mathcal{N}(0,1)$ in order to be able to use a quantizer alphabet that does not depend on m. For this reason, in the remainder of the paper, Φ shall denote an $m \times N$ matrix whose entries are i.i.d. from $\mathcal{N}(0,1)$.

Let $q := q_{\Sigma\Delta}$ be output of the standard greedy rth order $\Sigma\Delta$ quantizer with the alphabet $\mathcal{A} = \delta\mathbb{Z}$ and input y. As stated in Section 2, we know that $\|y - q\|_\infty \leq 2^{r-1}\delta$ and therefore $\|y - q\|_2 \leq 2^{r-1}\delta\sqrt{m}$.

4.1 Coarse recovery and recovery of support

Our first goal is to recover the support T of x. Note that support recovery in various compressed sensing contexts has received some attention lately (e.g., [17], [34], [35]). However, for this paper, the results we present in this section are sufficient and more appropriate, given our choice of decoder.

To estimate the support T, we shall use a coarse approximation of x. Let

$$x' := \arg\min \|z\|_1 \text{ subject to } \|\Phi z - q\|_2 \leq \epsilon := 2^{r-1}\delta\sqrt{m}. \tag{48}$$

By the robust recovery result (for our choice of normalization for Φ), we know that

$$\|x - x'\|_2 \leq \eta := 5 \cdot 2^r \delta.$$

The simplest attempt to recover T from x' is to pick the positions of its k largest entries. This attempt can fail if some entry of x_j on T is smaller than η for then it is possible that $x'_j = 0$ and therefore j is not picked. On the other hand, it is easy to see that if the smallest nonzero entry of x is strictly bigger than 2η in magnitude, then this method always succeeds. (Since $\|x - x'\|_\infty \leq \eta$, the entries of x' are bigger than η on T and less than η on T^c.) The constant 2 can be replaced with $\sqrt{2}$ by a more careful analysis, and can be pushed arbitrarily close to 1 by picking more than k positions. The proposition below gives a precise condition on how well this can be done. We also provide a bound on how much of x can potentially be missed if no lower bound on $|x_j|$ is available for $j \in T$.

Proposition 4.1. *Let $\|x - x'\|_{\ell_2^N} \leq \eta$, $T = \operatorname{supp} x$ 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'.*

(i) *$\|x_{T \setminus T'}\|_2 \leq \beta\eta$ where $\beta \leq \left(1 + \frac{k}{k'}\right)^{1/2}$.*

(ii) *If $|x_j| > \gamma\eta$ for all $j \in T$, where $\gamma := \left(1 + \frac{1}{k'-k+1}\right)^{1/2}$, then $T' \supset T$.*

Proof. (i) We have

$$\sum_{j \in T} |x_j - x'_j|^2 + \sum_{j \in T^c} |x'_j|^2 = \|x - x'\|_2^2 \leq \eta^2. \tag{49}$$

In particular, this implies

$$\sum_{j \in T \setminus T'} |x_j - x'_j|^2 + \sum_{j \in T' \setminus T} |x'_j|^2 \leq \eta^2. \tag{50}$$

Suppose $T \setminus T' \neq \emptyset$. Then $T' \setminus T$ is also nonempty. In fact, we have

$$|T' \setminus T| = |T \setminus T'| + k' - k.$$

Now, observe that

$$\frac{1}{|T \setminus T'|} \sum_{j \in T \setminus T'} |x'_j|^2 \leq \max_{j \in T \setminus T'} |x'_j|^2 \leq \min_{j \in T' \setminus T} |x'_j|^2 \leq \frac{1}{|T' \setminus T|} \sum_{j \in T' \setminus T} |x'_j|^2,$$

which, together with (50) implies

$$\|x_{T \setminus T'}\|_2 \leq \|x'_{T \setminus T'}\|_2 + \|(x - x')_{T \setminus T'}\|_2 \leq \|x'_{T \setminus T'}\|_2 + \sqrt{\eta^2 - \frac{|T' \setminus T|}{|T \setminus T'|} \|x'_{T \setminus T'}\|_2^2}.$$

It is easy to check that for any $A > 0$, and any $0 \leq t \leq \eta/\sqrt{A}$,

$$t + \sqrt{\eta^2 - At^2} \leq \left(1 + \frac{1}{A}\right)^{1/2} \eta. \tag{51}$$

The result follows by setting $A = |T' \setminus T|/|T \setminus T'|$ and noticing that $A \geq k'/k$.

(ii) Let $z_1 \geq \cdots \geq z_N$ be the decreasing rearrangement of $|x'_1|, \ldots, |x'_N|$. We have

$$\sum_{j \in T} |x'_j|^2 \leq \sum_{i=1}^{k} z_i^2$$

so

$$\sum_{j \in T^c} |x'_j|^2 \geq \sum_{i=k+1}^{N} z_i^2 \geq \sum_{i=k+1}^{k'+1} z_i^2 \geq (k' - k + 1) z_{k'+1}^2.$$

Hence by (49) we have

$$\max_{j \in T} |x_j - x'_j|^2 + (k' - k + 1) z_{k'+1}^2 \leq \eta^2.$$

Since $|x'_j| \geq |x_j| - |x_j - x'_j|$, the above inequality now implies

$$\min_{j \in T} |x'_j| \geq \min_{j \in T} |x_j| - \max_{j \in T} |x_j - x'_j| \geq \min_{j \in T} |x_j| - \sqrt{\eta^2 - (k' - k + 1) z_{k'+1}^2}.$$

Now, another application of (51) with $A = k' - k + 1$ yields

$$-\sqrt{\eta^2 - (k' - k + 1) z_{k'+1}^2} \geq z_{k'+1} - \gamma\eta$$

and therefore

$$\min_{j \in T} |x'_j| \geq \min_{j \in T} |x_j| + z_{k'+1} - \gamma\eta > z_{k'+1} = \max_{j \in T'^c} |x'_j|.$$

It is then clear that $T \subset T'$ because if $T'^c \cap T \neq \emptyset$, the inequality

$$\max_{j \in T'^c} |x'_j| \geq \max_{j \in T'^c \cap T} |x'_j| \geq \min_{j \in T} |x'_j|$$

would give us a contradiction. □

Note that if the k' largest entries of x' are picked with $k' > k$, then one would need to work with T' for the fine recovery stage, and therefore the starting assumptions on Φ have to be modified for k'. For simplicity we shall stick to $k' = k$ and consequently $\gamma = \sqrt{2}$.

4.2 Fine recovery

Once T is found, the rth order Sobolev dual frame $F := F_{\mathrm{Sob},r}$ of $E = \Phi_T$ is computed and we set $\hat{x}_{\Sigma\Delta} = Fq$. We now restate and prove Theorem B.

Theorem 4.2. *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*

$$\|x - \hat{x}_{\Sigma\Delta}\|_2 \lesssim_r \lambda^{-\alpha(r-\frac{1}{2})}\delta. \tag{52}$$

Proof. Suppose that $\lambda \geq c(\log N)^{1/(1-\alpha)}$ with $c = c_6$ as in the proof of Theorem 3.8. Let $q_{\Sigma\Delta}$ be obtained by quantizing $y := \Phi x$ via an rth order $\Sigma\Delta$ scheme with alphabet $\mathcal{A} = \delta\mathbb{Z}$ and with the quantization rule as in (18), and let u be the associated state sequence as in (16). Define $x^{\#}$ as the solution of the program

$$\min \|z\|_1 \text{ subject to } \|\Phi z - q_{\Sigma\Delta}\|_2 \leq \epsilon.$$

Suppose that Φ admits the robust recovery constant C_1, e.g. $C_1 = 10$. Hence the solution $x^{\#}$ of the program (3) satisfies $\|x - x^{\#}\|_2 \leq C_1\epsilon/\sqrt{m}$ for every x in Σ_k^N provided that $\|y - q_{\Sigma\Delta}\|_2 \leq \epsilon$. As discussed in Section 2, in this case we have $\|y - q_{\Sigma\Delta}\|_2 \leq 2^{r-1}\delta\sqrt{m}$ which implies

$$\|x - x^{\#}\|_2 \leq C_1 2^{r-1}\delta.$$

Assume that

$$\min_{j\in T} |x_j| \geq C_1 \cdot 2^{r-1/2}\delta =: C\delta. \tag{53}$$

Then, Proposition 4.1 (with $\gamma = \sqrt{2}$ and $\eta = C_1 2^{r-1}$) shows that T', the support of the k largest entries of $x^{\#}$, is identical to the support T of x. Finally, set

$$\hat{x}_{\Sigma\Delta} = F_{\text{sob},r} q_{\Sigma\Delta}$$

where $F_{\text{sob},r}$ is the rth order Sobolev dual of Φ_T. Using the fact that $\|u\|_2 \leq 2^{-1}\delta\sqrt{m}$ (see Section 2) together with the conclusion of Theorem 3.8 and the error bound (27), we conclude that

$$\|x - \hat{x}_{\Sigma\Delta}\|_2 \leq \frac{\|u\|_2}{\sqrt{m}\,\sigma_{\min}(D^{-r}E)} \leq \frac{\lambda^{-\alpha(r-1/2)}}{2c_1}\delta. \tag{54}$$

The normalizing $\sqrt{m}$ factor appears in the first inequality in (54) because Theorem 3.8 is stated for matrices with $\mathcal{N}(0, \frac{1}{m})$ entries. Note that the RIP and therefore the robust recovery will hold with probability $1 - \exp(c''m)$, and our Sobolev dual reconstruction error bound will hold with probability $1 - \exp(-c_7 m\lambda^{-\alpha})$. Here c_1 and c_7 are as in the proof of Theorem 3.8.

□

Remark. In the concrete case $C_1 = 10$, suppose we have

$$\min_{j\in T} |x_j| \geq \sqrt{2}\eta = 5 \cdot 2^{r+1/2}\delta. \tag{55}$$

If MSQ is used as the quantization method, then the best error guarantee we have that holds uniformly on T would be

$$\|x - x^{\#}_{\text{MSQ}}\|_\infty \leq \|x - x^{\#}_{\text{MSQ}}\|_2 \leq 5\delta.$$

It can be argued that the approximately recovered entries of $x^{\#}_{\text{MSQ}}$ are meaningful only when the minimum nonzero entry of x is at least as large as the maximum uncertainty in $x^{\#}_{\text{MSQ}}$, which is only known to be bounded by 5δ. Hence, in some sense the size condition (55) is natural (modulo the factor $2^{r+1/2}$).

4.3 Quantizer choice and rate-distortion issues

Suppose that we are given a CS problem with fixed dimensions m, N, and k, which also fixes λ. Furthermore suppose $x \in \Sigma_k$ satisfies

$$A \leq |x_j| \leq \rho \quad \text{for all } j \in T. \tag{56}$$

Our ultimate goal is to determine the quantizer (among the infinite family of $\Sigma\Delta$ quantizers of arbitrary order as well as MSQ) that minimizes the resulting approximation error for the given problem dimensions for a fixed bit-budget. A complete analysis of this problem is beyond the scope of this paper. However, below we will show that even a first-order $\Sigma\Delta$ quantizer is significantly superior to MSQ as long as the quantizer step size δ is sufficiently small and λ satisfies the condition of Theorem 4.2 for $r = 1$.

For usefulness of our results, we need δ, the quantizer step size, to satisfy

$$\delta \leq \frac{A}{10\sqrt{2}} \tag{57}$$

where, as before, we assumed $C_1 = 10$. Fix some δ that satisfies (57). Next, we need to determine finite alphabets $\mathcal{A}_1 = \{\pm j\delta : \ j = 0, 1, \ldots, 2^{B_1}\}$ and $\mathcal{A}_2 = \{\pm j\delta : \ j = 0, 1, \ldots, 2^{B_2}\}$ that ensure that the first-order $\Sigma\Delta$ quantizer implemented with $\mathcal{A}_1$ and MSQ implemented with $\mathcal{A}_2$ do not overload whenever $y = \Phi x$ is quantized with the respective scheme. In the case of the first-order $\Sigma\Delta$ quantizer, we then require B_1 to satisfy $2^{B_1}\delta \geq \|y\|_\infty$, and in the case of MSQ, we need $2^{B_2}\delta \geq \|y\|_\infty - \delta/2$. Therefore, in order to estimate B_1 and B_2 as a function of the problem dimensions, we need to bound $\|y\|_\infty$ efficiently.

An improved bound for the dynamic range

If we use the RIP, then Φ does not expand the ℓ_2-norm of k-sparse vectors by more than a factor of $2\sqrt{m}$ (note our choice of normalization for Φ), and therefore it follows that

$$\|y\|_\infty \leq \|y\|_2 \leq 2\sqrt{m}\|x\|_2 \leq 2\rho\sqrt{mk},$$

which is a restatement of the inequality

$$\|E\|_{\ell_\infty^k \to \ell_\infty^m} \leq \sqrt{k}\|E\|_{\ell_2^k \to \ell_2^m}$$

that holds for any $m \times k$ matrix E. However, it can be argued that the (∞, ∞)-norm of a random matrix should typically be smaller. In fact, if E were drawn from the Bernoulli model, i.e., $E_{ij} \sim \pm 1$, then we would have

$$\|E\|_{\ell_\infty^k \to \ell_\infty^m} = k = \lambda^{-1/2}\sqrt{mk},$$

as can easily be seen from the general formula

$$\|E\|_{\ell_\infty^k \to \ell_\infty^m} = \max_{1 \leq i \leq m} \sum_{j=1}^{k} |E_{ij}|. \tag{58}$$

Using simple concentration inequalities for Gaussian random variables, it turns out that for the range of aspect ratio $\lambda = m/k$ and probability of encountering a matrix Φ that we are interested in, we have $\|E\|_{\ell_\infty^k \to \ell_\infty^m} \leq \lambda^{-\alpha/2}\sqrt{mk}$ for every $m \times k$ submatrix E of Φ. We start with the following estimate:

Proposition 4.3. *Let $\xi_1, \dots, \xi_k$ i.i.d. standard Gaussian variables. Then, for any $\Theta > 1$,*

$$\mathbb{P}\left(\sum_{j=1}^{k} |\xi_j| > \Theta\right) \leq 2^k e^{-\Theta^2/(2k)}. \tag{59}$$

Proof.

$$\begin{aligned}
\mathbb{P}\left(\sum_{j=1}^{k} |\xi_j| > \Theta\right) &\leq \min_{t\geq 0} \int_{\mathbb{R}^k} e^{-\left(\Theta - \sum_{j=1}^k |x_j|\right)t} \prod_{j=1}^{k} e^{-x_j^2/2} \frac{\mathrm{d}x_j}{\sqrt{2\pi}} \\
&= \min_{t\geq 0} e^{-\Theta t} \left(e^{t^2/2} \int_{\mathbb{R}} e^{-\frac{1}{2}(|x|-t)^2} \frac{\mathrm{d}x}{\sqrt{2\pi}}\right)^k \\
&= \min_{t\geq 0} e^{-\Theta t} \left(2e^{t^2/2} \int_{0}^{\infty} e^{-\frac{1}{2}(x-t)^2} \frac{\mathrm{d}x}{\sqrt{2\pi}}\right)^k \\
&\leq 2^k \min_{t\geq 0} e^{-\Theta t + kt^2/2} \\
&= 2^k e^{-\Theta^2/(2k)}.
\end{aligned} \tag{60}$$

where in the last step we set $t = \Theta/k$. □

Proposition 4.4. *Let Φ be an $m \times N$ random matrix whose entries are i.i.d. $\mathcal{N}(0,1)$. Let $0 < \alpha < 1$ be any number and assume that*

$$\lambda := \frac{m}{k} \geq c_1 (\log N)^{1/(1-\alpha)}, \tag{61}$$

where c_1 is an appropriate constant. Then with probability at least $1 - e^{-c_2 m^{1-\alpha} k^\alpha}$ for some $c_2 > 0$, every $m \times k$ submatrix E of Φ satisfies

$$\|E\|_{\ell_\infty^k \to \ell_\infty^m} \leq \lambda^{-\alpha/2} \sqrt{mk}. \tag{62}$$

Proof. Proposition 4.3 straightforwardly implies that

$$\mathbb{P}\left(\{\exists T \text{ such that } |T| = k \text{ and } \|\Phi_T\|_{\ell_\infty^k \to \ell_\infty^m} > \Theta\}\right) \leq \binom{N}{k} m 2^k e^{-\Theta^2/(2k)}. \tag{63}$$

Let $\Theta = \lambda^{-\alpha/2}\sqrt{mk}$. It remains to show that

$$k \log N + k \log 2 + \log m + c_2 m^{1-\alpha} k^\alpha \leq \frac{\Theta^2}{2k}.$$

If c_1 in (61) is sufficiently large and c_2 is sufficiently small, then the expression on the left hand side is bounded by $k\lambda^{1-\alpha}/2 = \Theta^2/(2k)$. □

Without loss of generality, we may now assume that Φ also satisfies the conclusion of Proposition 4.4. Hence we have an improved bound on the range of y given by

$$\|y\|_\infty \leq \rho \lambda^{-\alpha/2} \sqrt{mk} = \rho \lambda^{(1-\alpha)/2} k. \tag{64}$$

Comparison of $\Sigma\Delta$ and MSQ

Let $\mathcal{A}_1, \mathcal{A}_2, B_1, B_2$ be as in the discussion below (57). Using (64), we see that B_1 needs to satisfy

$$2^{B_1}\delta \geq \rho\lambda^{(1-\alpha)/2}k, \tag{65}$$

so that the first-order $\Sigma\Delta$ quantizer is not overloaded when implemented with $\mathcal{A}_1$. Similarly, in the case MSQ, we need B_2 satisfy

$$2^{B_2}\delta \geq \rho\lambda^{(1-\alpha)/2}k - \delta/2. \tag{66}$$

In a practical setting it is natural to assume that $\rho \gg \delta$, and consequently we have $B_1 \approx B_2$, i.e., we may as well assume $\mathcal{A}_1 = \mathcal{A}_2$. Thus, to compare the two quantization schemes, we need to compare only the associated approximation errors. Based on Theorem 4.2, the approximation error (the distortion) $\mathcal{D}_{\Sigma\Delta}$ incurred after the fine recovery stage via Sobolev duals satisfies the bound

$$\mathcal{D}_{\Sigma\Delta} \leq \frac{\pi}{2}\lambda^{-\alpha/2}\delta. \tag{67}$$

which follows from (54) after setting $c_1 = c_1(1) = 1/\pi$ (see (84)). A similar calculation for the MSQ encoder with the standard ℓ_1 decoder results in

$$\mathcal{D}_{\text{MSQ}} \leq 5\delta. \tag{68}$$

Interpretation

The analysis above requires that both MSQ and $\Sigma\Delta$ encoders utilize high-resolution quantizers. In this setting, the benefit of using a first-order $\Sigma\Delta$ encoder is obvious upon comparing (67) and (68). It is important, though, to note that the above comparison makes sense only for a fixed value of λ that is built in to the compressed sensing setup and is sufficiently large to satisfy the condition of Theorem 4.2. If additional bits were available, it is more economical to put those in use to reduce the quantizer step size δ rather than for collecting more measurements.

From Theorem 4.2 it is clear that $\Sigma\Delta$ quantizers of order $r > 1$ could improve the distortion further if λ is sufficiently large. Specifically, given λ there is an optimal order $r(\lambda)$ that minimizes the associated distortion after taking into account all r-dependencies of the numerical constants. An analysis to determine the optimal order as a function of λ, and the associated minimum distortion, is beyond the scope of this paper. However, numerical experiments that we present in the next section suggest that with modest values of λ, $\Sigma\Delta$ schemes of order $r = 2$ and $r = 3$ indeed produce approximations with significantly smaller distortion.

5 Numerical experiments

In order to test the accuracy of Theorem 3.7, our first numerical experiment concerns the minimum singular value of $D^{-r}E$ as a function of $\lambda = m/k$. In Figure 1, we plot the worst case (the largest) value, among 1000 realizations, of $1/\sigma_{min}(D^{-r}E)$ for the range $1 \leq \lambda \leq 25$, where we have kept $k = 50$. As predicted by this theorem, we find that the negative slope in the log-log scale is roughly equal to $r - 1/2$, albeit slightly less, which seems in agreement with the presence of our control parameter α. As for the size of the r-dependent constants, the function $5^r\lambda^{-r+1/2}$ seems to be a

reasonably close numerical fit, which also explains why we observe the separation of the individual curves after $\lambda > 5$.

Our next experiment involves the full quantization algorithm for compressed sensing including the "recovery of support" and "fine recovery" stages. To that end, we first generate a 1000×2000 matrix Φ, where the entries of Φ are drawn i.i.d. according to $\mathcal{N}(0,1)$. To examine the performance of the proposed scheme as the redundancy λ increases in comparison to the performance of the standard MSQ quantization, we run a set of experiments: In each experiment we fix the sparsity $k \in \{10, 20\}$, and we generate k-sparse signals x 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 MSQ, as well as the 1st and 2nd order $\Sigma\Delta$ quantizers, defined via (16) and (18) (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 MSQ quantization) is displayed in Figure 2 (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 $F_{\text{sob},r}$. Defining $F_{\text{sob},0} := (\Phi_{\widetilde{T}}^{(m)})^\dagger$, in each case, our final estimate of the signal is obtained via the fine recovery stage as $\hat{x}_{\widetilde{T}} = F_{\text{sob},r}q$, $\hat{x}_{\widetilde{T}^c} = 0$. Note that this way, we obtain an alternative reconstruction also in the case of MSQ. 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 2. For our final 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 in Figure 3.

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

- $\Sigma\Delta$ schemes outperform the coarse reconstruction obtained from MSQ quantized measurements significantly 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 B which are 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 MSQ (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 MSQ quantized measurements (with the ℓ_1 decoder only) where the approximation error does not seem to depend on λ. A rigourous analysis of this behaviour is an open problem.

6 Remarks on extensions

6.1 Other noise-shaping matrices

In the above approach, the particular quantization scheme that we use can be identified with its "noise-shaping matrix", which is D^r in the case of an rth order $\Sigma\Delta$ scheme and the identity matrix in the case of MSQ.

The results we obtained above are valid for the aforementioned noise-shaping matrices. However, our techniques are fairly general and our estimates can be modified to investigate the accuracy obtained using an arbitrary quantization scheme with the associated invertible noise-shaping matrix H. In particular, the estimates depend solely on the distribution of the singular values of H. Of course, in this case, we also need change our "fine recovery" stage and use the "H-dual" of the corresponding frame E, which we define via

$$F_H H = (HE)^\dagger. \tag{69}$$

As an example, consider an rth order *high-pass* $\Sigma\Delta$ scheme whose noise shaping matrix is H^r where H is defined via

$$H_{ij} := \begin{cases} 1, & \text{if } i = j \text{ or if } i = j+1, \\ 0, & \text{otherwise.} \end{cases} \tag{70}$$

It is easy to check that the singular values of H are identical to those of D. It follows that all the results presented in this paper are valid also if the compressed measurements are quantized via an an rth order high-pass $\Sigma\Delta$ scheme, provided the reconstruction is done using the H^r-duals instead of the rth order Sobolev duals. Note that such a result for high-pass $\Sigma\Delta$ schemes is not known to hold in the case of structured frames.

6.2 Measurement noise and compressible signals

One of the natural questions 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 natural question is how to extend this theory to include the case when the underlying signals are not necessarily strictly sparse, but nevertheless still "compressible".

Suppose $x \in \mathbb{R}^N$ is not sparse, but compressible in the usual sense (e.g. as in [10]), and let $y = \Phi x + e$, where e stands for additive measurement noise. The *coarse recovery stage* inherits the stability and robustness properties of ℓ_1 decoding for compressed sensing, therefore the accuracy of this first reconstruction depends on the best k-term approximation error for x, and the deviation of Φx from the quantized signal 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 MSQ, hence the reconstruction error at the coarse recovery stage would also be comparable. In the *fine recovery stage*, however, the difference between $\sigma_{\max}(F_H H)$ and $\sigma_{\max}(F_H)$ plays a critical role. In the particular case of $H = D^r$ and $F_H = F_{\text{sob},r}$, the Sobolev duals we use in the reconstruction are tailored to reduce the effect of the quantization error introduced by an rth order $\Sigma\Delta$ quantizer. This is reflected in the fact that as λ increases, the kernel of the reconstruction operator $F_{\text{sob},r}$ contains a larger portion of high-pass sequences (like the quantization error of $\Sigma\Delta$ modulation), and is quantified by the bound $\sigma_{\max}(F_{\text{sob},r} D^r) \lesssim \lambda^{-(r-1/2)} m^{-1/2}$ (see Theorem A, (26) and (27)). Consequently, obtaining more measurements increases λ, and even though $\|y - q\|_2$ increases as well, the reconstruction error due to quantization decreases. 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 $F_{\mathrm{sob},r}$. In fact, for any dual F, the relation $FE = I$ implies $\sigma_{\max}(F) \geq 1/\sigma_{\max}(E) \gtrsim m^{-1/2}$ already and in the case of measurement noise, it is not possible to do better than the canonical dual $E^\dagger$ on average. 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.

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 H given by

$$H_{ij} := \begin{cases} 1, & \text{if } i = j, \\ -\mu & \text{if } i = j+1, \\ 0, & \text{otherwise,} \end{cases} \tag{71}$$

where $\mu \in (0, 1)$. Our preliminary numerical experiments (see Figure 4) 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. We note that the parameter μ above can be adjusted based on how compressible the signals of interest are and what the expected noise level is.

Acknowledgments

The authors would like to thank Ronald DeVore and Vivek Goyal for valuable discussions. We thank the American Institute of Mathematics and Banff International Research Station for hosting two meetings where this work was initiated. This work was supported in part by: National Science Foundation Grant CCF-0515187 (Güntürk), Alfred P. Sloan Research Fellowship (Güntürk), National Science Foundation Grant DMS-0811086 (Powell), a Pacific Century Graduate Scholarship from the Province of British Columbia through the Ministry of Advanced Education (Saab), a UGF award from the UBC (Saab), and a Natural Sciences and Engineering Research Council of Canada Discovery Grant (Yılmaz).

A Singular values of D^{-r}

It will be more convenient to work with the singular values of D^r. Note that because of our convention of descending ordering of singular values, we have

$$\sigma_j(D^{-r}) = \frac{1}{\sigma_{m+1-j}(D^r)}, \quad j = 1, \ldots, m. \tag{72}$$

For $r = 1$, an explicit formula is available [30, 33]. Indeed, we have

$$\sigma_j(D) = 2\cos\left(\frac{\pi j}{2m+1}\right), \quad j = 1, \ldots, m, \tag{73}$$

which implies

$$\sigma_j(D^{-1}) = \frac{1}{2\sin\left(\frac{\pi(j-1/2)}{2(m+1/2)}\right)}, \quad j = 1, \ldots, m. \tag{74}$$

The first observation is that $\sigma_j(D^r)$ and $(\sigma_j(D))^r$ are different, because D and D^* do not commute. However, this becomes insignificant as $m \to \infty$. In fact, the asymptotic distribution of $(\sigma_j(D^r))_{j=1}^m$ as $m \to \infty$ is rather easy to find using standard results in the theory of Toeplitz matrices: D is a banded Toeplitz matrix whose symbol is $f(\theta) = 1 - e^{i\theta}$, hence the symbol of D^r is $(1 - e^{i\theta})^r$. It then follows by Parter's extension of Szegö's theorem [28] that for any continuous function ψ, we have

$$\lim_{m\to\infty} \frac{1}{m} \sum_{j=1}^{m} \psi(\sigma_j(D^r)) \quad = \quad \frac{1}{2\pi} \int_{-\pi}^{\pi} \psi(|f(\theta)|^r)\, d\theta. \tag{75}$$

We have $|f(\theta)| = 2\sin|\theta|/2$ for $|\theta| \le \pi$, hence the distribution of $(\sigma_j(D^r))_{j=1}^m$ is asymptotically the same as that of $2^r \sin^r(\pi j/2m)$, and consequently, we can think of $\sigma_j(D^{-r})$ roughly as $\big(2^r \sin^r(\pi j/2m)\big)^{-1}$. Moreover, we know that $\|D^r\|_{\mathrm{op}} \le \|D\|_{\mathrm{op}}^r \le 2^r$, hence $\sigma_{\min}(D^{-r}) \ge 2^{-r}$.

When combined with known results on the rate of convergence to the limiting distribution in Szegö's theorem, the above asymptotics could be turned into an estimate of the kind given in Proposition 3.1, perhaps with some loss of precision. Here we shall provide a more direct approach which is not asymptotic, and works for all $m > 4r$. The underlying observation is that D and D^* almost commute: $D^*D - DD^*$ has only two nonzero entries, at $(1,1)$ and (m,m). Based on this observation, we show below that $D^{*r}D^r$ is then a perturbation of $(D^*D)^r$ of rank at most $2r$.

Proposition A.1. *Let $C^{(r)} = D^{*r}D^r - (D^*D)^r$ where we assume $m \ge 2r$. Define*

$$I_r := \{1, \ldots, r\} \times \{1, \ldots, r\} \cup \{m - r + 1, \ldots, m\} \times \{m - r + 1, \ldots, m\}.$$

Then $C^{(r)}_{i,j} = 0$ for all $(i,j) \in I_r^c$. Therefore, $\mathrm{rank}(C^{(r)}) \le 2r$.

Proof. Define the set $\mathcal{C}_r$ of all "r-cornered" matrices as

$$\mathcal{C}_r = \{M : M_{i,j} = 0 \text{ if } (i,j) \in I_r^c\},$$

and the set $\mathcal{B}_r$ of all "r-banded" matrices as

$$\mathcal{B}_r = \{M : M_{i,j} = 0 \text{ if } |i - j| > r\}.$$

Both sets are closed under matrix addition. It is also easy to check the following facts (for the admissible range of values for r and s):

(i) If $B \in \mathcal{B}_r$ and $C \in \mathcal{C}_s$, then $BC \in \mathcal{C}_{r+s}$ and $CB \in \mathcal{C}_{r+s}$.

(ii) If $B \in \mathcal{B}_r$ and $\tilde{B} \in \mathcal{B}_s$, then $B\tilde{B} \in \mathcal{B}_{r+s}$.

(iii) If $C \in \mathcal{C}_r$ and $\tilde{C} \in \mathcal{C}_s$, then $C\tilde{C} \in \mathcal{C}_{\max(r,s)}$.

(iv) If $C \in \mathcal{C}_r$, then $D^*CD \in \mathcal{C}_{r+1}$.

Note that $DD^*, D^*D \in \mathcal{B}_1$ and the commutator $[D^*, D] =: \Gamma_1 \in \mathcal{C}_1$. Define

$$\Gamma_r := (D^*D)^r - (DD^*)^r = (DD^* + \Gamma_1)^r - (DD^*)^r.$$

We expand out the first term (noting the non-commutativity), cancel $(DD^*)^r$ and see that every term that remains is a product of r terms (counting each DD^* as one term) each of which is either in $\mathcal{B}_1$ or in $\mathcal{C}_1$. Repeated applications of (i), (ii), and (iii) yield $\Gamma_r \in \mathcal{C}_r$.

We will now show by induction on r that $C^{(r)} \in \mathcal{C}_r$ for all r such that $2r \leq m$. The cases $r = 0$ and $r = 1$ hold trivially. Assume the statement holds for a given value of r. Since

$$C^{(r+1)} = D^*(C^{(r)} + \Gamma_r)D$$

and $\Gamma_r \in \mathcal{C}_r$, property (iv) above now shows that $C^{(r+1)} \in \mathcal{C}_{r+1}$. □

The next result, originally due to Weyl (see, e.g., [4, Thm III.2.1]), will now allow us to estimate the eigenvalues of $D^{*r}D^r$ using the eigenvalues of $(D^*D)^r$:

Theorem A.2 (Weyl)**.** *Let B and C be $m \times m$ Hermitian matrices where C has rank at most p and $m > 2p$. Then*

$$\lambda_{j+p}(B) \leq \lambda_j(B + C), \qquad j = 1, \ldots, m - p, \tag{76}$$

$$\lambda_{j-p}(B) \geq \lambda_j(B + C), \qquad j = p + 1, \ldots, m, \tag{77}$$

where we assume eigenvalues are in descending order.

We are now fully equipped to prove Proposition 3.1.

Proof of Proposition 3.1. We set $p = 2r$, $B = (D^*D)^r$, and $C = C^{(r)} = D^{*r}D^r - (D^*D)^r$ in Weyl's theorem. By Proposition A.1, C has rank at most $2r$. Hence, we have the relations

$$\lambda_{j+2r}((D^*D)^r) \leq \lambda_j(D^{*r}D^r), \qquad j = 1, \ldots, m - 2r, \tag{78}$$

$$\lambda_{j-2r}((D^*D)^r) \geq \lambda_j(D^{*r}D^r), \qquad j = 2r + 1, \ldots, m. \tag{79}$$

Since $\lambda_j((D^*D)^r) = \lambda_j(D^*D)^r$, this corresponds to

$$\sigma_{j+2r}(D)^r \leq \sigma_j(D^r), \qquad j = 1, \ldots, m - 2r, \tag{80}$$

$$\sigma_{j-2r}(D)^r \geq \sigma_j(D^r), \qquad j = 2r + 1, \ldots, m. \tag{81}$$

For the remaining values of j, we will simply use the largest and smallest singular values of D^r as upper and lower bounds. However, note that

$$\sigma_1(D^r) = \|D^r\|_{\mathrm{op}} \leq \|D\|_{\mathrm{op}}^r = (\sigma_1(D))^r$$

and similarly

$$\sigma_m(D^r) = \|D^{-r}\|_{\mathrm{op}}^{-1} \geq \|D^{-1}\|_{\mathrm{op}}^{-r} = (\sigma_m(D))^r.$$

Hence (80) and (81) can be rewritten together as

$$\sigma_{\min(j+2r,m)}(D)^r \leq \sigma_j(D^r) \leq \sigma_{\max(j-2r,1)}(D)^r, \qquad j = 1, \ldots, m. \tag{82}$$

Inverting these relations via (72), we obtain

$$\sigma_{\min(j+2r,m)}(D^{-1})^r \leq \sigma_j(D^{-r}) \leq \sigma_{\max(j-2r,1)}(D^{-1})^r, \quad j = 1, \ldots, m. \tag{83}$$

Finally, to demonstrate the desired bounds of Proposition 3.1, we rewrite (74) via the inequality $2x/\pi \leq \sin x \leq x$ for $0 \leq x \leq \pi/2$ as

$$\frac{m+1/2}{\pi(j-1/2)} \leq \sigma_j(D^{-1}) \leq \frac{m+1/2}{2(j-1/2)}, \tag{84}$$

and observe that $\min(j+2r, m) \asymp_r j$ and $\max(j-2r, 1) \asymp_r j$ for $j = 1, \ldots, m$. □

Remark. The constants $c_1(r)$ and $c_2(r)$ that one obtains from the above argument would be significantly exaggerated. This is primarily due to the fact that Proposition 3.1 is not stated in the tightest possible form. The advantage of this form is the simplicity of the subsequent analysis in Section 3.1. Our estimates of $\sigma_{\min}(D^{-r}E)$ would become significantly more accurate if the asymptotic distribution of $\sigma_j(D^{-r})$ is incorporated into our proofs in Section 3.1. However, the main disadvantage would be that the estimates would then hold only for all sufficiently large m.

References

[1] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin. A simple proof of the restricted isometry property for random matrices. *Constr. Approx.*, 28(3):253–263, 2008.

[2] J.J. Benedetto, A.M. Powell, and Ö. Yılmaz. Second order sigma–delta ($\Sigma\Delta$) quantization of finite frame expansions. *Appl. Comput. Harmon. Anal*, 20:126–148, 2006.

[3] 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.

[4] R. Bhatia. *Matrix analysis*, volume 169 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1997.

[5] J. Blum, M. Lammers, A.M. Powell, and Ö. Yılmaz. Sobolev duals in frame theory and Sigma-Delta quantization. *J. Fourier Anal. and Appl.*, 16(3):365–381, 2010.

[6] 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.

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

[8] P. Boufounos and R.G. Baraniuk. Sigma delta quantization for compressive sensing. In *Wavelets XII*, edited by D. Van De Ville, V.K. Goyal and M. Papadakis, Proceedings of SPIE Vol. 6701 (SPIE, Bellingham, WA, 2007). Article CID 670104.

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

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

[11] 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.

[12] A. Cohen, W. Dahmen, and R. DeVore. Compressed sensing and best k-term approximation. *J. Amer. Math. Soc.*, 22(1):211–231, 2009.

[13] W. Dai, H.V. Pham, and O. Milenkovic. Quantized compressive sensing. `arXiv:0901.0749v2 [cs.IT]`, 2009.

[14] 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.

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

[16] 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.

[17] A.K. Fletcher, S. Rangan, and V.K. Goyal. Necessary and Sufficient Conditions for Sparsity Pattern Recovery. *IEEE Trans. Inform. Theory*, 55(2009):5758–5772, 2009.

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

[19] 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.

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

[21] L. Jacques, D.K. Hammond, and M.J. Fadili. Dequantizing compressed sensing: When oversampling and non-gaussian constraints combine. `arXiv:0902.2367v4 [math.OC]`, 2009.

[22] 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.

[23] M.C. Lammers, A.M. Powell, and Ö. Yılmaz. On quantization of finite frame expansions: sigma-delta schemes of arbitrary order. *Proceedings of SPIE*, 6701:670108, 2007.

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

[25] G.G. Lorentz, M. von Golitschek, and Y. Makovoz. *Constructive approximation*, volume 304 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1996. Advanced problems.

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

[27] R.J. Pai. *Nonadaptive lossy encoding of sparse signals*. PhD thesis, Massachusetts Institute of Technology, 2006.

[28] S.V. Parter. On the distribution of the singular values of Toeplitz matrices. *Linear Algebra Appl.*, 80:115–130, 1986.

[29] M. Rudelson and R. Vershynin. Smallest singular value of a random rectangular matrix. *Comm. Pure Appl. Math.*, 62(12):1595–1739, 2009.

[30] G. Strang. The discrete cosine transform. *SIAM Review*, pages 135–147, 1999.

[31] J.Z. Sun and V.K. Goyal. Optimal quantization of random measurements in compressed sensing. In *IEEE International Symposium on Information Theory, 2009. ISIT 2009*, pages 6–10, 2009.

[32] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B Met.*, 58(1):267–288, 1996.

[33] J. von Neumann. Distribution of the ratio of the mean square successive difference to the variance. *Ann. Math. Statistics*, 12(4):367–395, 1941.

[34] M.J. Wainwright. Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE Trans. Inform. Theory*, 55(12):5728–5741, 2009.

[35] M.J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using l 1-constrained quadratic programming (Lasso). *IEEE Trans. Inform. Theory*, 55(5):2183–2202, 2009.

[36] A. Zymnis, S. Boyd, and E. Candes. Compressed sensing with quantized measurements. *IEEE Signal Proc. Let.*, 17(2):149, 2010.

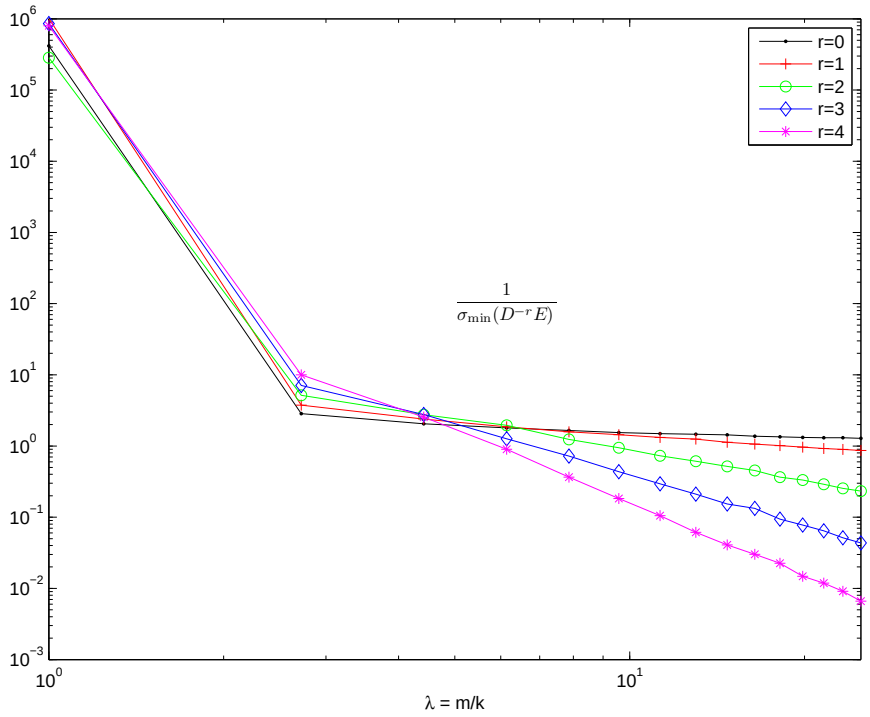
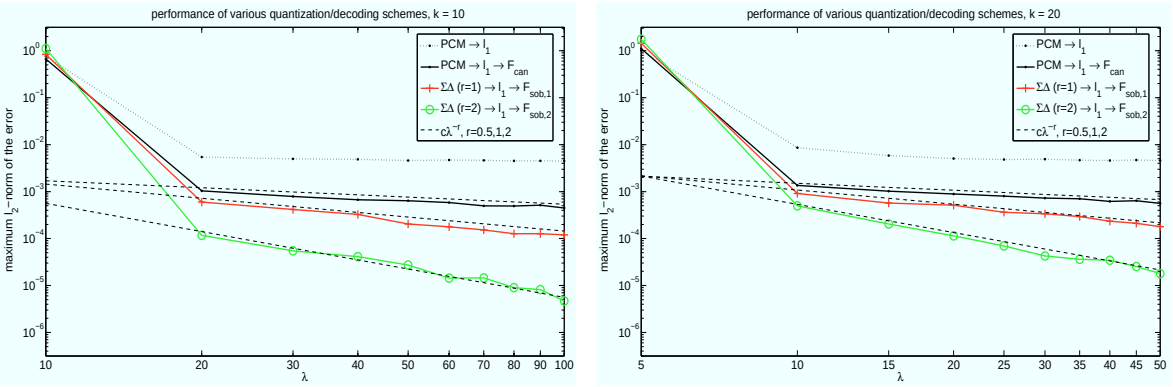
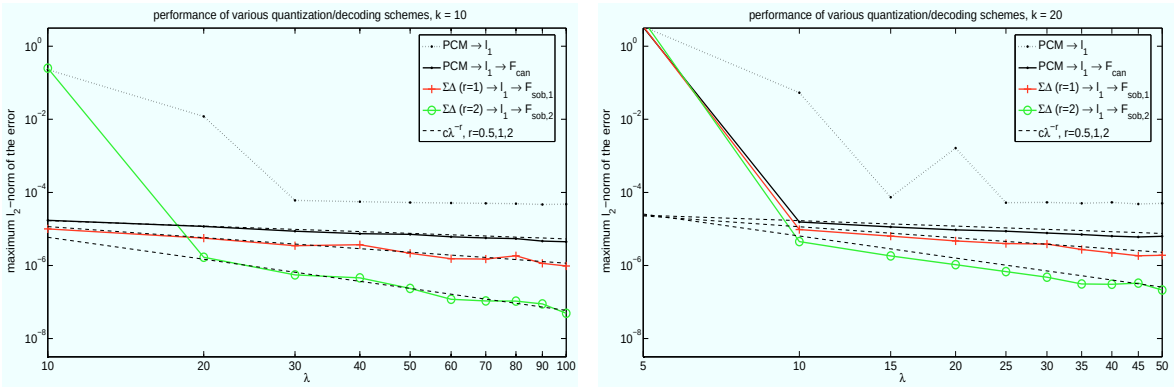


Figure 1: Numerical behavior (in log-log scale) of $1/\sigma_{min}(D^{-r}E)$ as a function of $\lambda = m/k$, for $r = 0, 1, 2, 3, 4$. In this figure, $k = 50$ and $1 \leq \lambda \leq 25$. For each problem size, the largest value of $1/\sigma_{min}(D^{-r}E)$ among 1000 realizations of a random $m \times k$ matrix E sampled from the Gaussian ensemble $\mathcal{N}(0, \frac{1}{m}I_m)$ was recorded.



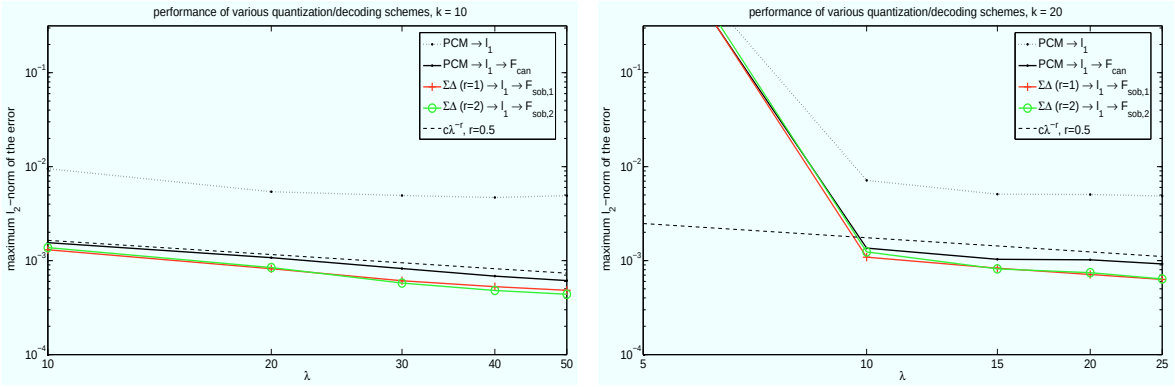
(a) (b)

Figure 2: The worst case performance of the proposed ΣΔ quantization and reconstruction schemes for $k = 10$ and $k = 20$. For this experiment the non-zero entries of x are constant and $\delta = 0.01$.



(a) (b)

Figure 3: The worst case performance of the proposed ΣΔ quantization and reconstruction schemes for for $k = 10$ and $k = 20$. For this experiment the non-zero entries of x are i.i.d. $\mathcal{N}(0, 1)$ and $\delta = 10^{-4}$.



(a) (b)

Figure 4: The worst case performance of the proposed ΣΔ quantization and reconstruction schemes (with general duals) for $k = 10$ and $k = 20$. For this experiment the non-zero entries of x are constant and $\delta = 0.01$.