SIGMA-DELTA QUANTIZATION OF SUB-GAUSSIAN FRAME EXPANSIONS AND ITS APPLICATION TO COMPRESSED SENSING

FELIX KRAHMER, RAYAN SAAB, AND ÖZGÜR YILMAZ

ABSTRACT. Suppose that the collection $\{e_i\}_{i=1}^m$ forms a frame for $\mathbb{R}^k$, where each entry of the vector e_i is a sub-Gaussian random variable. We consider expansions in such a frame, which are then quantized using a Sigma-Delta scheme. We show that an arbitrary signal in $\mathbb{R}^k$ can be recovered from its quantized frame coefficients up to an error which decays root-exponentially in the oversampling rate m/k. Here the quantization scheme is assumed to be chosen appropriately depending on the oversampling rate and the quantization alphabet can be coarse. The result holds with high probability on the draw of the frame uniformly for all signals. The crux of the argument is a bound on the extreme singular values of the product of a deterministic matrix and a sub-Gaussian frame. For fine quantization alphabets, we leverage this bound to show polynomial error decay in the context of compressed sensing. Our results extend previous results for structured deterministic frame expansions and Gaussian compressed sensing measurements. compressed sensing, quantization, random frames, root-exponential accuracy, Sigma-Delta, sub-Gaussian matrices
2010 Math Subject Classification: 94A12, 94A20, 41A25, 15B52

1. INTRODUCTION

1.1. **Main problem.** In this paper we address the problem of digitizing, or quantizing, generalized linear measurements of finite dimensional signals. In this setting a signal is a vector $x \in \mathbb{R}^N$, and the acquired measurements are inner products of x with elements from a collection of m *measurement vectors* in $\mathbb{R}^N$. This generalized linear measurement model has received much attention lately, both in the *frame theory* literature where one considers $m \geq N$, e.g., [1], and in the *compressed sensing* literature where $m \ll N$, e.g., [2]. For concreteness, let $\{a_i\}_{i=1}^m \subset \mathbb{R}^N$ denote the measurement vectors. The generalized linear measurements are given by $y_i = \langle a_i, x \rangle, \ i \in \{1, ..., m\}$, and can be organized as a vector y, given by $y = Ax$. Note that here and throughout the paper, we always consider column vectors. Our goal is to *quantize* the measurements, i.e., to map the components y_i of y to elements of a fixed finite set so that the measurements can be stored and transmitted digitally. A natural requirement for such maps is that they allow for accurate recovery of the underlying signal. As the domain of the forward map is typically an uncountable set and its range is finite, an exact inversion is, in

general, not possible. Accordingly we seek an approximate inverse, which
we shall refer to as a *reconstruction scheme* or *decoder*. More precisely, let
$\mathcal{A} \subset \mathbb{R}$ be a finite set, which we shall refer to as the quantization alphabet,
and consider a set $\mathcal{X}$ of signals, which is compact in $\mathbb{R}^N$. Then a *quantization
scheme* is a map
$$\mathcal{Q} : A\mathcal{X} \to \mathcal{A}^m$$
and a *reconstruction scheme* is of the form
$$\Delta : \mathcal{A}^m \to \mathbb{R}^N.$$
For Δ to be an approximate inverse of $\mathcal{Q}$, we need that the reconstruction
error
$$\epsilon(x) := \|x - \Delta(\mathcal{Q}(Ax))\|_2$$
viewed as a function of x is as small as possible in some appropriate norm.
Typical choices are $\|\epsilon\|_{L^\infty(\mathcal{X})}$, the worst case error, and $\|\epsilon\|_{L^2(\mathcal{X})}^2$, the mean
square error.

   We are interested in quantization and reconstruction schemes that yield
fast error decay as the number of measurements increases. Next, we give a
brief overview of the literature on quantization in both frame quantization
and compressed sensing settings and explain how these two settings are
intimately connected. Below, we use different notations for the two settings
to make it easier to distinguish between them.

1.2. **Finite frame setting.** Let $\{e_i\}_{i=1}^m$ be a frame for $\mathbb{R}^k$, i.e., $m \geq k$, and
assume that the matrix E whose ith row is e_i^T has rank k, thus the map
$x \mapsto Ex$ is injective. Accordingly, one can reconstruct any $x \in \mathbb{R}^k$ exactly
from the frame coefficients $y = Ex$ using, e.g., any left inverse F of E. As
we explained above, the frame coefficients y can be considered as generalized
measurements of x and our goal is to quantize y such that the approximation
error is guaranteed to decrease as the number of measurements, i.e., m,
increases.

1.2.1. *Memoryless scalar quantization.* The most naive (and intuitive) quan-
tization method is rounding off every frame coefficient to the nearest element
of the quantizer alphabet $\mathcal{A}$. This scheme is generally called memoryless
scalar quantization (MSQ) and yields (nearly) optimal performance when
$m = k$ and E is an orthonormal basis. However, as redundancy increases
(say, we keep k fixed and increase m) MSQ becomes highly suboptimal.
In particular, one can prove that the (expected) approximation error via
MSQ with a given fixed alphabet $\mathcal{A}$ can never decay faster than $(m/k)^{-1}$
regardless of the reconstruction scheme that is used and regardless of the
underlying E [3]. This is significantly inferior compared to another family
of quantization schemes, called $\Sigma\Delta$ quantizers, where one can have an ap-
proximation error that decays like $(m/k)^{-s}$ for any integer $s > 0$, provided
one uses an appropiate order quantizer.

1.2.2. *Sigma-Delta quantization.* Despite its use in the engineering community since the 1960's as an alternative quantization scheme for digitizing band-limited signals (see, e.g., [4]), a rigorous mathematical analysis of $\Sigma\Delta$ quantization was not done until the work of Daubechies and Devore [5]. Since then, the mathematical literature on $\Sigma\Delta$ quantization has grown rapidly.

Early work on the mathematical theory of $\Sigma\Delta$ quantization has focused on understanding the reconstruction accuracy as a function of *oversampling rate* in the context of *bandlimited functions*, i.e., functions with compactly supported Fourier transform. Daubechies and DeVore constructed in their seminal paper [5] *stable* rth-order $\Sigma\Delta$ schemes with a one-bit alphabet. Furthermore, they proved that when such an rth-order scheme is used to quantize an *oversampled* bandlimited function, the resulting approximation error is bounded by $C_r\lambda^{-r}$ where $\lambda > 1$ is the *oversampling ratio* and C_r depends on the fine properties of the underlying stable $\Sigma\Delta$ schemes. For a given oversampling rate λ one can then optimize the order r to minimize the associated *worst-case* approximation error, which, in the case of the stable $\Sigma\Delta$ family of Daubechies and DeVore, yields that the approximation error is of order $O(\lambda^{-c\log\lambda})$.

In [6], Güntürk constructed an alternative infinite family of $\Sigma\Delta$ quantizers of arbitrary order—refined later by Deift et al. [7]—and showed that using these new quantizers one can do significantly better. Specifically, using such schemes (see Section 2.2) in the bandlimited setting, one obtains an approximation error of order $O(2^{-c\lambda})$ where $c = 0.07653$ in [6] and $c \approx 0.102$ in [7]. In short, *when quantizing bounded bandlimited functions*, one gets exponential accuracy in the oversampling rate λ by using these $\Sigma\Delta$ schemes. In other words, one can refine the approximation by increasing the oversampling rate λ, i.e., by collecting more measurements, *exponentially* in the number of measurements without changing the quantizer resolution. Exponential error decay rates are known to be optimal [8, 6]; lower bounds for the constants c for arbitrary coarse quantization schemes are derived in [9]. In contrast, it again follows from [3] that independently quantizing the sample values at best yields linear decay of the average approximation error.

Motivated by the observation above that suggests that $\Sigma\Delta$ quantizers utilize the redundancy of the underlying expansion effectively, Benedetto et al. [10] showed that $\Sigma\Delta$ quantization schemes provide a viable quantization scheme for finite frame expansions in $\mathbb{R}^d$. In particular, [10] considers $x \in \mathbb{R}^k$ with $\|x\|_2 \leq 1$ and shows that the reconstruction error associated with the first-order $\Sigma\Delta$ quantization (i.e., $r = 1$) decays like λ^{-1} as λ increases. Here the "oversampling ratio" λ is defined as $\lambda = m/k$ if the underlying frame for $\mathbb{R}^k$ consists of m vectors. This is analogous to the error bound in the case of bandlimited functions with a first-order $\Sigma\Delta$ quantizer. Following [10], there have been several results on $\Sigma\Delta$ quantization of finite frames expansions improving on the $O(\lambda^{-1})$ approximation error by using higher order schemes, specialized frames, and alternative reconstruction techniques,

e.g., [11, 12, 13, 14, 15]. Two of these papers are of special interest for the purposes of this paper: Blum et al. showed in [14] that frames with certain smoothness properties (including harmonic frames) allow for the $\Sigma\Delta$ reconstruction error to decay like λ^{-r}, provided alternative dual frames—called Sobolev duals—are used for reconstruction. Soon after, [15] showed that by using higher order $\Sigma\Delta$ schemes whose order is optimally chosen as a function of the oversampling rate λ, one obtains that the worst-case reconstruction error decays like $e^{-C\sqrt{\frac{m}{k}}}$, at least in the case of two distinct structured families of frames: harmonic frames and the so-called *Sobolev self-dual frames* which were constructed in [15]. For a more comprehensive review of $\Sigma\Delta$ schemes and finite frames, see [16]. One of the fundamental contributions of this paper is to extend these results to wide families of random frames.

1.2.3. *Gaussian frames and quantization.* Based on the above mentioned results, one may surmise that structure and smoothness are in some way critical properties of frames, needed for good error decay in $\Sigma\Delta$ quantization. Using the error analysis techniques of [14], it can be seen, though, that what is critical for good error decay in $\Sigma\Delta$ quantization is the "smoothness" and "decay" properties of the *dual frame* that is used in the reconstruction—see Section 2.3, cf. [16]. This observation is behind the seemingly surprising fact that Gaussian frames, i.e., frames whose entries are drawn independently according to $\mathcal{N}(0,1)$, allow for polynomial decay in λ of the reconstruction error [17, Theorem A]—specifically, one can show that the approximation error associated with an rth-order $\Sigma\Delta$ scheme is of order $O(\lambda^{-r+\frac{1}{2}})$. The proof relies on bounding the extreme singular values of the product of powers of a deterministic matrix—the difference matrix D defined in Section 2.3—and a Gaussian random matrix. This result holds uniformly with high probability on the draw of the Gaussian frame.

1.3. **Compressed sensing setting.** Compressed sensing is a novel paradigm in mathematical signal processing that was spearheaded by the seminal works of Candes, Romberg, Tao [18], and of Donoho [19]. Compressed sensing is based on the observation that various classes of signals such as audio and images admit approximately sparse representations with respect to a known basis or frame. Central results of the theory establish that such signals can be recovered with high accuracy from a small number of appropriate, non-adaptive linear measurements by means of computatinally tractable reconstruction algorithms.

More precisely, one considers N-dimensional, k-sparse signals, i.e., vectors in the set
$$\Sigma_k^N := \{z \in \mathbb{R}^N, |\mathrm{supp(z)}| \le \mathrm{k}\}.$$
The generalized measurements are acquired via the $m \times N$ matrix Φ where $k < m \ll N$. The goal is to recover $z \in \Sigma_k^N$ from $y = \Phi z$. In this paper, we focus on random matrices Φ with independent sub-Gaussian entries in

the sense of Definitions 3.2 and 3.5 below. It is well-known [20] that if $m > Ck \log(N/k)$, where C is an absolute constant, with high probability, such a choice of Φ allows for the recovery of all $z \in \Sigma_k^N$ as the solution $z^{\#}$ of the ℓ_1-minimization-problem

$$z^{\#} = \arg \min_x \|x\|_1 \quad \text{subject to } y = \Phi x.$$

1.3.1. *Quantized compressed sensing.* The recovery guarantees for compressed sensing are provably stable with respect to measurement errors. Consequently, this allows to incorporate quantization into the theory, albeit naively, as one can treat the quantization error as noise. The resulting error bounds for quantized compressed sensing, however, are not satisfactory, mainly because additional quantized measurements will, in general, not lead to higher reconstruction accuracy [17]. The fairly recent literature on quantization of compressed sensing measurements mainly investigates two families of quantization methods: the 1-bit or multibit memoryless scalar quantization (MSQ) [21, 22, 23, 24] and $\Sigma\Delta$ quantization of arbitrary order [17].

The results on the MSQ scenario focus on replacing the naive reconstruction approach outlined above by recovery algorithms that exploit the structure of the quantization error. For Gaussian random measurement matrices, it has been shown that approximate reconstruction is possible via linear programming even if the measurements are coarsely quantized by just a single bit [23]. For non-Gaussian measurements, counterexamples with extremely sparse signals exist which show that, in general, corresponding results do not hold [23]. These extreme cases can be controlled by introducing the ℓ_∞-norm of the signal as an additional parameter, establishing recovery guarantees for arbitrary sub-Gaussian random measurement matrices, provided that this norm is not too large [24]. All these results yield approximations where the error does not decay faster than λ^{-1}, where in this case the oversampling rate λ is defined as the ratio of m, the number of measurements, to k, the sparsity level of the underlying signal. Again, it follows from [3] that for independently quantized measurements, i.e., MSQ, no quantization scheme and no recovery method can yield an error decay that is faster than λ^{-1}.

This bottleneck can be overcome by considering $\Sigma\Delta$ quantizers, which take into account the representations of previous measurements in each quantization step [17]. The underlying observation is that the compressed sensing measurements are in fact frame coefficients of the sparse signal restricted to its support. Accordingly, the problem of quantizing compressed sensing measurements is a frame quantization problem, even though the "quantizer" does not know what the underlying frame is. This motivates a two-stage approach for signal recovery:

In a first approximation, the quantization error is just treated as noise, a standard reconstruction algorithm is applied, and the indices of the largest coefficients are retained as a support estimate. Once the support has been identified, the measurements carry redundant information about the signal,

which is exploited in the quantization procedure by applying frame quantization techniques.

This two-stage approach has been analyzed in [17] for the specific case of Gaussian random measurement matrices. In particular, it was shown that under mild size assumptions on the non-zero entries of the underlying sparse signals, this approach can be carried through. Consequently, in the case of Gaussian measurement matrices, one obtains that the approximation error associated with an rth-order $\Sigma\Delta$ quantizer is of order $O(\lambda^{\alpha(-r+\frac{1}{2})})$ where $\alpha \in (0,1)$—see [17, Theorem B]. These results hold uniformly with high probability on the draw of the Gaussian measurement matrix provided $m \geq Ck(\log N)^{1/(1-\alpha)}$.

1.4. **Contributions.** Our work in this paper builds on these results, and generalizes them. Our contributions are two-fold. On the one hand, we establish corresponding results in the *compressed sensing setting* which allow arbitrary, independent, fixed variance *sub-Gaussian* (in the sense of Definition 3.2 below) random variables as measurement matrix entries. In particular, this includes the important case of Bernoulli matrices, whose entries are renormalized independent random signs. More precisely, in Theorem 4.10 we prove a refined version of the following.

Theorem 1.1. *Let Φ be an $m \times N$ matrix whose entries are appropriately normalized independent sub-Gaussian random variables and suppose that $\lambda := m/k \geq \left(C\log(eN/k)\right)^{\frac{1}{1-\alpha}}$ where $\alpha \in (0,1)$. With high probability the r-th order $\Sigma\Delta$ reconstruction $\hat{z}$ satisfies*

$$\|z - \hat{z}\|_2 \leq C\lambda^{-\alpha(r-1/2)}\delta,$$

for all $z \in \Sigma_k^N$ for which $\min\limits_{j \in \text{supp}(z)} |z_j| > C\delta$. Here δ is the resolution of the $\Sigma\Delta$ quantization alphabet and C is an appropriate constant that depends only on r.

Our second line of contributions is on frame quantization: We show that using appropriate $\Sigma\Delta$ quantization schemes, we obtain root-exponential decay of the reconstruction error with both Gaussian and sub-Gaussian frame entries. In particular, in Theorem 4.4 we prove a refined version of the following result.

Theorem 1.2. *Let E be an $m \times k$ matrix whose entries are appropriately normalized independent sub-Gaussian random variables. Suppose that $\lambda := m/k$ satisfies $\lambda \geq \lambda_0$, where λ_0 is a constant independent of k and m. Then with high probability on the draw of E, the corresponding reconstruction $\hat{x}$ from a $\Sigma\Delta$ scheme of appropriate order satisfies*

$$\|x - \hat{x}\|_2 \leq Ce^{-c\sqrt{\lambda}}.$$

for all x with $\|x\|_2 \leq 1$. Here c, C are appropriate constants.

Note that a key element of our proof, which may be of independent interest, pertains to the extreme singular values of the product of a deterministic matrix with quickly decaying singular values and a sub-Gaussian matrix, see Proposition 4.1.

Remark 1.3. *All of the constants in the above theorems can be made explicit. Moreover, the quantization schemes are explicit and tractable, as are the reconstruction algorithms; however, the quantization scheme and reconstruction algorithms are different between Theorems 1.1 and 1.2. Please see Theorems 4.4 and 4.10 for the full details.*

1.5. **Organization.** The remainder of the paper is organized as follows. In Section 2 we review $\Sigma\Delta$ quantization and basic error analysis techniques that will be useful in the rest of the paper. In Section 3 we introduce the concept of a sub-Gaussian random matrix and recall some of its key properties as well as some important probabilistic tools. In Section 4, we prove a probabilistic lower bound on the singular values of the product of the matrix D^{-r}, where r is a positive integer and D is a difference matrix, and a sub-Gaussian random matrix. Finally, we use this result in combination with some known results on the properties of various $\Sigma\Delta$ quantization schemes to prove the main theorems.

2. Sigma-Delta quantization

An rth order $\Sigma\Delta$ quantizer $Q^{(r)} : \mathbb{R}^m \mapsto \mathcal{A}^m$ maps a sequence of inputs $(y_i)_{i=1}^m$ to a sequence $(q_i)_{i=1}^m$ whose elements take on values from $\mathcal{A}$ via the iteration

$$(1) \qquad q_i = Q\left(\rho(u_{i-1}, \cdots, u_{i-r}, y_i, \cdots, y_{i-r+1})\right)$$
$$(\Delta^r u)_i = y_i - q_i.$$

Here ρ is a fixed function known as the *quantization rule* and $(u_i)_{i=1}^m$ is a sequence of state variables initialized to zero, i.e., $u_i = 0$ for all $i \leq 0$. It is worth noting that designing a a good quantization rule in the case $r > 1$ is generally non-trivial, as one seeks stable $\Sigma\Delta$ schemes, i.e., schemes that satisfy

$$(2) \qquad \|y\|_\infty \leq C_1 \implies \|u\|_\infty \leq C_2,$$

for constants C_1 and C_2 that do not depend on m (note that for the remainder of this paper, the constants are numbered in the order of appearance; this allows to refer to constants introduced in previous results and proofs). In particular, stability is difficult to ensure when one works with a coarse quantizer associated with a small alphabet, the extreme case of which is 1-bit quantization corresponding to $\mathcal{A} = \{\pm 1\}$.

In this work we consider two different sets of assumptions. Our results on compressed sensing reconstruction require sufficiently fine alphabets, whereas the results on frame quantization make no assumptions on

the size of the alphabet —in particular, allowing for very coarse alphabets. In both cases we will work with the $2L$ level *mid-rise* alphabet

$$
(3) \qquad \mathcal{A} = \Big\{ \pm (2j+1)\delta/2, \quad j \in \{0,...,L-1\} \Big\}.
$$

2.1. **Greedy sigma-delta schemes.** We will work with the *greedy* $\Sigma\Delta$ quantization schemes

$$
(4) \qquad q_i = Q\Big(\sum_{j=1}^{r} (-1)^{j-1} \binom{r}{j} u_{i-j} + y_i \Big)
$$

$$
u_i = \sum_{j=1}^{r} (-1)^{j-1} \binom{r}{j} u_{i-j} + y_i - q_i.
$$

It is easily seen by induction that for the $2L$ level mid-rise alphabet and $\|y\|_\infty \leq C$, a sufficient condition for stability is $L \geq 2\lceil \frac{C}{\delta} \rceil + 2^r + 1$ as this implies

$$
(5) \qquad \|u\|_\infty \leq \delta/2.
$$

Note that to satisfy this stability condition, the number of levels L must increase with r.

2.2. **Coarse sigma-delta schemes.** We are also interested in coarse $\Sigma\Delta$ quantization, i.e., schemes where the alphabet size is fixed. In this case, guaranteeing stability with a smaller alphabet typically entails a worse (i.e., larger) stability constant. The coarse $\Sigma\Delta$ schemes that we employ were first proposed by Güntürk [6] and refined by Deift et al. [7]. Originally designed to obtain exponential accuracy in the setting of bandlimited functions, they were used to obtain root-exponential accuracy in the finite frame setup in [15]. At their core is a change of variables of the form $u = g * v$, where u is as in (1) and $g \in \mathbb{R}^{d+1}$ for some $d \geq r$ (with entries indexed by the set $\{0,\ldots,d\}$) such that $g_0 = 1$. The quantization rule is then chosen in terms of the new variables as $\rho(v_i, v_{i-1}, \ldots, y_i) = (h*v)_i + y_i$, where $h = \delta^{(0)} - \Delta^r g$ with $\delta^{(0)}$ the Kronecker delta. Then (1) reads as

$$
(6) \qquad q_i = Q((h*v)_i + y_i)
$$
$$
v_i = (h*v)_i + y_i - q_i,
$$

where, again, Q is the scalar quantizer associated with the $2L$ level mid-rise alphabet (3). By induction, one concludes

$$
\|h\|_1 \frac{\delta}{2} + \|y\|_\infty \leq L\delta \implies \|v\|_\infty \leq \frac{\delta}{2} \implies \|u\|_\infty \leq \|g\|_1 \frac{\delta}{2},
$$

i.e., a sufficient condition to guarantee stability for all bounded inputs $\|y\|_\infty \leq \mu$ is

$$
(7) \qquad \|h\|_1 \leq 2L - \frac{2\mu}{\delta}.
$$

Thus, one is interested in choosing g with minimal $\|g\|_1$ subject to $h = \delta^{(0)} - \Delta^r g$ and (7). This problem was studied in [6, 7] leading to the following proposition (cf. [15]).

Proposition 2.1. *There exists a universal constant $C_3 > 0$ such that for any midrise quantization alphabet $\mathcal{A} = \mathcal{A}_L^\delta$, for any order $r \in \mathbb{N}$, and for all $\mu < \delta\left(L - \frac{1}{2}\right)$, there exists $g \in \mathbb{R}^{d+1}$ for some $d \geq r$ such that the $\Sigma\Delta$ scheme given in (6) is stable for all input signals y with $\|y\|_\infty \leq \mu$ and*

$$(8) \qquad \|u\|_\infty \leq cC_3^r r^r \frac{\delta}{2},$$

where $u = g * v$ as above, $C_3 = \left(\left\lceil \frac{\pi^2}{(\cosh^{-1}\gamma)^2} \right\rceil \frac{e}{\pi}\right)$ and $\gamma := 2L - \frac{2\mu}{\delta}$.

2.3. **Sigma-Delta error analysis.** As above, assume that $x \in \mathbb{R}^k$ and the frame matrix $E \in \mathbb{R}^{m \times k}$. If the vector of frame coefficients $y = Ex \in \mathbb{R}^m$ is $\Sigma\Delta$ quantized to yield the vector $q \in \mathcal{A}^m$, then linear reconstruction of x from q using some dual frame F of E (i.e., $FE = I$) produces the estimate $\hat{x} := Fq$. We would like to control the reconstruction error $\eta := x - \hat{x}$. Writing the state variable equations (1) in vector form, we have

$$(9) \qquad D^r u = y - q,$$

where D is the $m \times m$ difference matrix with entries given by

$$(10) \qquad D_{ij} = \begin{cases} 1 & i = j \\ -1 & i = j + 1 \\ 0 & \text{otherwise} \end{cases}.$$

Thus,

$$(11) \qquad \eta = x - Fq = F(y - q) = FD^r u.$$

Working with with stable $\Sigma\Delta$ schemes, one can control $\|u\|_2$ via $\|u\|_\infty$. Thus, it remains to bound the operator norm $\|FD^r\| := \|FD^r\|_{\ell_2^m \mapsto \ell_2^k}$ and a natural choice for F is

$$(12) \qquad F := \arg\min_{G:GE=I} \|GD^r\| = (D^{-r}E)^\dagger D^{-r}.$$

This so-called Sobolev dual frame was first proposed in [14]. Here $A^\dagger := (A^*A)^{-1}A^*$ is the $k \times m$ Moore-Penrose (left) inverse of the $m \times k$ matrix A. Since (12) implies that $FD^r = (D^{-r}E)^\dagger$, the singular values of $D^{-r}E$ will play a key role in this paper.

We begin by presenting some important properties of the matrix D. The following proposition is a quantitative version of Proposition 3.1 of [17].

Proposition 2.2. *The singular values of the matrix D^{-r} satisfy*

$$\frac{1}{(3\pi r)^r}\left(\frac{m}{j}\right)^r \leq \sigma_j(D^{-r}) \leq (6r)^r\left(\frac{m}{j}\right)^r.$$

Proof. Note that (see, e.g., [17])

$$\frac{1}{\pi}\left(\frac{m+1/2}{j-1/2}\right) \leq \sigma_j(D^{-1}) \leq \frac{1}{2}\left(\frac{m+1/2}{j-1/2}\right).$$

Moreover, by Weyl's inequalities [25] on the singular values of Hermitian matrices, it holds that (see [17] for the full argument)

$$\left(\sigma_{\min(j+2r,m)}(D^{-1})\right)^r \leq \sigma_j(D^{-r}) \leq \left(\sigma_{\max(j-2r,1)}(D^{-1})\right)^r.$$

Combining the above inequalities, we obtain

$$\frac{1}{\pi^r}\frac{(m+1/2)^r}{\min(j+2r-1/2,m-1/2)^r} \leq \sigma_j(D^{-r}) \leq \frac{1}{2^r}\frac{(m+1/2)^r}{\max(j-2r-1/2,1/2)^r}.$$

Observing that

$$\min(j+2r-1/2, m-1/2) \leq j+2r \leq 3rj$$

for $r, j \in \mathbb{Z}^+$ establishes the lower bound.

For the upper bound, note that $j - 2r - 1/2 \geq (2r+1)^{-1}j/2$, for $j \geq 2r+1$ and $1/2 \geq (2r+1)^{-1}j/2$ for $j \in \{1, ..., 2r\}$. Thus,

$$\frac{(m+1/2)^r}{2^r \max(j-2r-1/2, 1/2)^r} \leq \frac{2^r m^r}{2^r(4r+2)^{-r}j^r} \leq (6r)^r\left(\frac{m}{j}\right)^r.$$

$\square$

3. Sub-Gaussian random matrices

Here and throughout, $x \sim \mathcal{D}$ denotes that the random variable x is drawn according to a distribution $\mathcal{D}$. Furthermore, $\mathcal{N}(0, \sigma^2)$ denotes the zero-mean Gaussian distribution with variance σ^2. The following definition provides a means to compare the tail decay of two distributions.

Definition 3.1. *If two random variables $\eta \sim \mathcal{D}_1$ and $\xi \sim \mathcal{D}_2$ satisfy $P(|\eta| > t) \leq KP(|\xi| > t)$ for some constant K and all $t \geq 0$, then we say that η is K-dominated by ξ (or, alternatively, by $\mathcal{D}_2$).*

Definition 3.2. *A random variable is sub-Gaussian with parameter $c > 0$ if it is e-dominated by $\mathcal{N}(0, c^2)$.*

Remark 3.3. *One can also define sub-Gaussian random variables via their moments or, in case of zero mean, their moment generating functions. See [20] for a proof that all these definitions are equivalent.*

Remark 3.4. *Examples of sub-Gaussian random variables include Gaussian random variables, all bounded random variables (such as Bernoulli), and their linear combinations.*

Definition 3.5. *We say that a matrix E is sub-Gaussian with parameter c, mean μ and variance σ^2 if its entries are independent sub-Gaussian random variables with mean μ, variance σ^2, and parameter c.*

The contraction principle (see, for example, Lemma 4.6 of [26]) will allow us to derive estimates for sub-Gaussian random variables via the corresponding results for Gaussians.

Lemma 3.6 (Contraction Principle). *Let $G : \mathbb{R}^+ \mapsto \mathbb{R}^+$ be a non-decreasing convex function and let (η_i) and (ξ_i) be two finite symmetric sequences of random variables such that there exists a constant $K \geq 1$ such that for each i, η_i is K-dominated by ξ_i. Then, for any finite sequence (x_i) in a Banach space equipped with a norm $\| \cdot \|$ we have*

$$\mathbb{E}G(\|\sum_i \eta_i x_i\|) \leq \mathbb{E}G(K\|\sum_i \xi_i x_i\|).$$

While the contraction principle as well as the following chaos estimate are formulated for random vectors, we mainly work with random matrices. Thus, it is convenient to "vectorize" the matrices: for a matrix A, we denote by $\vec{A}$ the vector formed by stacking its columns into a single column vector.

To state the more refined chaos estimate, we need the concept of the Talagrand γ_2-functional (see, e.g., [26] for more details).

Definition 3.7. *For a metric space (T, d), an admissible sequence of T is a collection of subsets of T, $\{T_r : r \geq 0\}$, such that for every $s \geq 1$, $|T_r| \leq 2^{2^r}$ and $|T_0| = 1$. The γ_2 functional is defined by*

$$\gamma_2(T, d) = \inf \sup_{t \in T} \sum_{r=0}^{\infty} 2^{r/2} d(t, T_r),$$

where the infimum is taken with respect to all admissible sequences of T.

Furthermore, for $\mathcal{A}$ a set of matrices, we denote by $d_{Fr}(\mathcal{A}) := \sup_{A \in \mathcal{A}} \|A\|_{Fr}$ and $d_{\ell_2 \to \ell_2}(\mathcal{A}) := \sup_{A \in \mathcal{A}} \|A\|_{\ell_2 \to \ell_2}$ the diameter in the Frobenius norm $\| \cdot \|_{Fr}$ and the spectral norm $\| \cdot \|_{\ell_2 \to \ell_2}$, respectively. Here the Frobenius norm is given by $\|A\|_{Fr} = \|\vec{A}\|_2$ and the spectral norm is given by $\|A\|_{\ell_2 \to \ell_2} = \sup_{\|x\|_2 = 1} \|Ax\|_2$. The following theorem is a slightly less general version of [27, Thm. 3.1].

Theorem 3.8. *Let $\mathcal{A}$ be a symmetric set of matrices, that is, $\mathcal{A} = -\mathcal{A}$, and let ξ be a random vector whose entries ξ_j are independent, sub-Gaussian random variables of mean zero, variance one, and parameter c. Set*

$$\mu = \gamma_2(\mathcal{A}, \| \cdot \|_{\ell_2 \to \ell_2}) \left(\gamma_2(\mathcal{A}, \| \cdot \|_{\ell_2 \to \ell_2}) + d_{Fr}(\mathcal{A})\right),$$

$$\nu_1 = d_{\ell_2 \to \ell_2}(\mathcal{A})(\gamma_2(\mathcal{A}, \| \cdot \|_{\ell_2 \to \ell_2}) + d_{Fr}(\mathcal{A})), \quad \text{and} \quad \nu_2 = d_{\ell_2 \to \ell_2}^2(\mathcal{A}).$$

Then, for $t > 0$,

$$\mathbb{P}\left(\sup_{A \in \mathcal{A}} \left|\|A\xi\|_2^2 - \mathbb{E}\|A\xi\|_2^2\right| \geq C_4 \mu + t\right) \leq 2\exp\left(-C_5 \min\left\{\frac{t^2}{\nu_1^2}, \frac{t}{\nu_2}\right\}\right).$$

The constants C_4, C_5 depend only on c.

4. Main results

4.1. Estimates of singular values and operator norms.

As argued above, a key quantity to control the $\Sigma\Delta$ reconstruction error both in the context of compressed sensing and frame quantization is the norm $\|FD^r u\|_2$, where F is the Sobolev dual of a sub-Gaussian frame E and u is the associated state vector. This quantity can be controlled by the product of the operator norm $\|FD^r\| = \frac{1}{\sigma_{\min}(D^{-r}E)}$ and the vector norm $\|u\|_2$. We will estimate these quantities separately in the following two propositions. An estimate for the first quantity can be deduced from the following proposition together with two observations: First, recall that singular values are invariant under unitary transformations, so $\sigma_{\min}(D^{-r}E) = \sigma_{\min}(SV^*E)$, where USV^* is the singular value decomposition of D^{-r}. Second, when estimated using Proposition 2.2, the singular values of D^{-r} are bounded exactly as in the following assumptions.

Proposition 4.1. *Let E be an $m \times k$ sub-Gaussian matrix with mean zero, unit variance, and parameter c, let $S = \operatorname{diag}(s)$ be a diagonal matrix, and let V be an orthonormal matrix, both of size $m \times m$. Further, let $r \in \mathbb{Z}^+$ and suppose that $s_j \geq C_6^r \left(\frac{m}{j}\right)^r$, where C_6 is a positive constant that may depend on r. Then there exist constants $C_7, C_8 > 0$ (depending on c and C_6) such that for $0 < \alpha < 1$ and $\lambda := \frac{m}{k} \geq C_7^{\frac{1}{1-\alpha}}$*

$$\mathbb{P}\left(\sigma_{\min}(\tfrac{1}{\sqrt{m}}SV^*E) \leq \lambda^{\alpha(r-1/2)}\right) \leq 2\exp(-C_8 m^{1-\alpha}k^\alpha).$$

In particular, C_8 depends only on c, while C_7 can be expressed as $f(c)C_6^{-\frac{2r}{2r-1}}$ provided $C_6 \leq 1/2$.

Proof. The matrix SV^*E has dimensions m and k, so by the Courant min-max principle applied to the transpose one has

$$(13) \qquad \sigma_{\min}(SV^*E) = \min_{\substack{W \subset \mathbb{R}^m \\ \dim W = m-k+1}} \sup_{z \in W: \|z\|_2 = 1} \|E^*VSz\|_2$$

Noting that, for $m \geq \widetilde{k} := C_9 m^{1-\alpha}k^\alpha > k$, where the constant C_9 will be determined later, each $m - k + 1$-dimensional subspace intersects the span $V_{\widetilde{k}}$ of the first $\widetilde{k}$ standard basis vectors in at least a $\widetilde{k} - k + 1$-dimensional space, this expression is bounded from below by

$$(14)$$
$$\min_{\substack{W \subset V_{\widetilde{k}} \\ \dim W = \widetilde{k}-k+1}} \sup_{z \in W: \|z\|_2 = 1} \|E^*VSz\|_2 \geq \min_{\substack{W \subset V_{\widetilde{k}} \\ \dim W = \widetilde{k}-k+1}} \sup_{z \in W: \|z\|_2 = s_{\widetilde{k}}} \|E^*Vz\|_2$$

$$(15) \qquad\qquad = \min_{\substack{W \subset \mathbb{R}^{\widetilde{k}} \\ \dim W = \widetilde{k}-k+1}} \sup_{z \in W: \|z\|_2 = 1} s_{\widetilde{k}} \|E^*VP_{\widetilde{k}}^*z\|_2.$$

The inequality follows from the observation that $V_{\widetilde{k}}$ is invariant under S and the smallest singular value of $S|_{V_{\widetilde{k}}}$ is $s_{\widetilde{k}}$. In the last step, $P_{\widetilde{k}} \in \mathbb{R}^{\widetilde{k} \times m}$ denotes

the projection of an m-dimensional vector onto its first $\widetilde{k}$ components. We note that (15), again by the Courant min-max principle, is equal to

$$(16) \qquad s_{\widetilde{k}} \sigma_k(E^* V P_{\widetilde{k}}^*) = s_{\widetilde{k}} \sigma_{\min}(P_{\widetilde{k}} V^* E) = s_{\widetilde{k}} \inf_{y \in S^{k-1}} \|P_{\widetilde{k}} V^* E y\|_2$$

Now, as $\mathbb{E}\|P_{\widetilde{k}} V^* E y\|_2^2 = \widetilde{k}$,

$$(17) \qquad \inf_{y \in S^{k-1}} \|P_{\widetilde{k}} V^* E y\|_2^2 \geq \left(\widetilde{k} - \sup_{y \in S^{k-1}} \left| \|P_{\widetilde{k}} V^* E y\|_2^2 - \mathbb{E}\|P_{\widetilde{k}} V^* E y\|_2^2 \right| \right).$$

Thus, noting that $\frac{\lambda^{\alpha(r-1/2)}}{s_{\widetilde{k}}} < m^{\alpha(r-1/2)} k^{-\alpha(r-1/2)} C_6^{-r} m^{-r} \widetilde{k}^r = C_6^{-r} C_9^{r-\frac{1}{2}} \frac{\sqrt{\widetilde{k}}}{\sqrt{m}}$ and that by choosing $C_9 = \min(\frac{1}{2} C_6^{\frac{2r}{2r-1}}, \frac{1}{2})$ we ensure that $1 - C_6^{-2r} C_9^{2r-1} \geq \frac{1}{2}$,

(18)

$$\mathbb{P}(\sigma_{\min}(\tfrac{1}{\sqrt{m}} S V^* E) \leq \lambda^{\alpha(r-1/2)})$$

(19)

$$\leq \mathbb{P}\left(\sup_{y \in S^{k-1}} \left| \|\tfrac{1}{\sqrt{m}} P_{\widetilde{k}} V^* E y\|_2^2 - \mathbb{E}\|\tfrac{1}{\sqrt{m}} P_{\widetilde{k}} V^* E y\|_2^2 \right| \geq (1 - C_6^{-2r} C_9^{2r-1}) \frac{\widetilde{k}}{m} \right)$$

(20)

$$\leq \mathbb{P}\left(\sup_{y \in S^{k-1}} \left| \|\tfrac{1}{\sqrt{m}} P_{\widetilde{k}} V^* E y\|_2^2 - \mathbb{E}\|\tfrac{1}{\sqrt{m}} P_{\widetilde{k}} V^* E y\|_2^2 \right| \geq \frac{\widetilde{k}}{2m} \right).$$

Note that this choice of C_9 also ensures $\widetilde{k} \leq m$, which is required above. We will estimate (20) using Theorem 3.8, similarly to the proof of [27, Thm. A.1]. Indeed, we can write

$$(21) \qquad \tfrac{1}{\sqrt{m}} P_{\widetilde{k}} V^* E y = W_y \xi,$$

where $\xi = \overrightarrow{E^*}$ is a vector of length km with independent sub-Gaussian entries of mean zero and unit variance, and

$$(22) \qquad W_y = \frac{1}{\sqrt{m}} P_{\widetilde{k}} V^* \begin{pmatrix} y^T & 0 & \cdots & 0 \\ 0 & y^T & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & y^T \end{pmatrix}.$$

In order to apply Theorem 3.8, we need to estimate, for $\mathcal{A} = \{W_y : y \in S^{k-1}\}$, $d_{Fr}(\mathcal{A})$, $d_{\ell_2 \to \ell_2}(\mathcal{A})$, and $\gamma_2(\mathcal{A}, \|\cdot\|_{\ell_2 \to \ell_2})$. We obtain for $A = W_y \in \mathcal{A}$:

$$(23) \qquad \|A\|_{Fr}^2 = \frac{1}{m} \sum_{j=1}^{k} \sum_{p_1, p_2 = 1}^{\widetilde{k}, m} y_j^2 V_{p_1, p_2}^2 = \frac{\widetilde{k}}{m}, \quad \text{so} \quad d_{Fr}(\mathcal{A}) = \sqrt{\frac{\widetilde{k}}{m}}.$$

Furthermore, we have, for $z \in \mathbb{R}^k$,

(24)
$$\|W_z\|_{\ell_2 \to \ell_2} = \left\| \frac{1}{\sqrt{m}} P_{\widetilde{k}} V^* \begin{pmatrix} z^T & 0 & \cdots & 0 \\ 0 & z^T & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & z^T \end{pmatrix} \right\|_{\ell_2 \to \ell_2} \leq \left\| \frac{1}{\sqrt{m}} \begin{pmatrix} z^T & 0 & \cdots & 0 \\ 0 & z^T & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & z^T \end{pmatrix} \right\|_{\ell_2 \to \ell_2},$$

so the quantities $d_{\ell_2 \to \ell_2}(\mathcal{A})$ and $\gamma_2(\mathcal{A}, \|\cdot\|_{\ell_2 \to \ell_2})$ can be estimated in exact analogy to [27, Thm. A.1]. This yields $d_{\ell_2 \to \ell_2}(\mathcal{A}) = \frac{1}{\sqrt{m}}$ and $\gamma_2(\mathcal{A}, \|\cdot\|_{\ell_2 \to \ell_2}) \leq C_{10} \sqrt{\frac{k}{m}}$ for some constant C_{10} depending only on c. With these estimates, we obtain for the quantities μ, ν_1, ν_2 in Theorem 3.8

(25)
$$\mu \leq (2C_{10} + 2) \frac{\sqrt{k\widetilde{k}}}{m}$$

(26)
$$\nu_1 \leq (C_{10} + 1) \frac{\sqrt{\widetilde{k}}}{m}$$

(27)
$$\nu_2 \leq \frac{1}{m},$$

so the resulting tail bound reads

(28)
$$\mathbb{P}\left(\sup_{y \in S^{k-1}} \left| \|\frac{1}{\sqrt{m}} P_{\widetilde{k}} V^* E y\|_2^2 - \mathbb{E} \|\frac{1}{\sqrt{m}} P_{\widetilde{k}} V^* E y\|_2^2 \right| \geq C_4 (2C_{10} + 2) \frac{\sqrt{k\widetilde{k}}}{m} + t \right) \leq e^{-C_5 \min\left(\frac{t^2 m^2}{(C_{10}+1)\widetilde{k}}, mt \right)}.$$

where C_4 and C_5 are the constants depending only on c as they appear in Theorem 3.8. Note that $k = \widetilde{k} \frac{\lambda^{-(1-\alpha)}}{C_9}$, so for oversampling rates $\lambda > \left((4C_4(2C_{10}+2))^2 / C_9 \right)^{\frac{1}{1-\alpha}} =: C_7^{\frac{1}{1-\alpha}}$, we obtain $C_4 E \leq \frac{\widetilde{k}}{4m}$ and hence, choosing $t = \frac{\widetilde{k}}{4m}$, we obtain the result

(29)
$$\mathbb{P}\left(\sup_{y \in S^{k-1}} \left| \|\frac{1}{\sqrt{m}} P_{\widetilde{k}} V^* E y\|_2^2 - \mathbb{E} \|\frac{1}{\sqrt{m}} P_{\widetilde{k}} V^* E y\|_2^2 \right| \geq \frac{\widetilde{k}}{2m} \right) \leq e^{-C_8 \widetilde{k}}$$

where, as desired, the constant $C_8 := \frac{C_5}{16(C_{10}+1)}$ depends only on the sub-Gaussian parameter c. $\qquad\square$

In contrast to the term $\|FD^r\|$ analyzed in the previous proposition, $\|u\|_2$ crucially depends on the quantization procedure that is employed. The procedure employed will be fundamentally different in the frameworks of compressed sensing and frame quantization. While the quantization level in the compressed sensing scheme is chosen sufficiently fine to allow for accurate support recovery via standard compressed sensing techniques, there is no need for this in the context of frame quantization and the quantization scheme employed can be coarse.

In both cases, we will employ $\Sigma\Delta$ schemes which are stable in the sense of (2). As explained in Section 2, $\|u\|_2$ can be controlled for such schemes via the input Ex. More precisely, to bound $\|u\|_2$, we require a bound on $\|Ex\|_\infty \leq \|E\|_{\ell_2 \to \ell_\infty} \|x\|_2$. Since the matrices $E = E(\omega)$ are random we derive bounds on the operator norms $\|E\|_{\ell_2 \to \ell_\infty}$ that hold with high probability on the draw of E.

Proposition 4.2. *Let $\widetilde{E}$ be an $m \times k$ sub-Gaussian matrix with mean zero, unit variance, and parameter c, let $E = \frac{1}{\sqrt{m}} \widetilde{E}$ and fix $\alpha \in (0,1)$. Denote the associated oversampling rate by $\lambda := m/k$. Then, with probability at least $1 - e^{-\frac{1}{4}\lambda^{1-\alpha} k}$, we have for all $\lambda > C_{11}^{\frac{1}{1-\alpha}}$*

$$(30) \qquad \|E\|_{\ell_2 \to \ell_\infty} \leq e^{1/2} \lambda^{-\frac{\alpha}{2}}.$$

Here C_{11} is a constant that may depend on c, but that is independent of k and α.

Proof. Since

$$(31) \qquad \|E\|_{\ell_2 \to \ell_\infty} = \max_{i \in \{1,\dots,m\}} \left(\sum_{j=1}^{k} e_{ij}^2 \right)^{1/2} = \max_{\eta \text{ is a row of } E} \|\eta\|_2,$$

we will focus on bounding the norm $\|\eta\|_2$ for random vectors $\eta \in \mathbb{R}^k$ consisting of independent sub-Gaussian entries η_i with parameter $\frac{1}{\sqrt{m}}$. Using Markov's inequality as well as the contraction principle applied to the increasing convex function $G : x \mapsto e^{x^2 t/2}$ and the sequence $x_i = e_i$ of standard basis vectors, we reduce to the case of a k-dimensional random vector $\xi \sim \mathcal{N}(0, \frac{1}{m}I)$:

$$(32) \qquad \mathbb{P}\left(\|\eta\|_2^2 \geq \Theta/\lambda \right) \leq \inf_{t>0} e^{-t\Theta/(2\lambda)} \mathbb{E} e^{\|\eta\|_2^2 t/2}$$

$$(33) \qquad \leq \inf_{t>0} e^{-t\Theta/(2\lambda)} \mathbb{E} e^{e\|\xi\|_2^2 t/2}$$

$$(34) \qquad = \inf_{t>0} e^{-t\Theta/(2\lambda)} \left(1 - et/m \right)^{-k/2}$$

$$(35) \qquad \leq \left(\frac{\Theta}{e} \right)^{k/2} \exp\left(-(\Theta/e - 1)k/2 \right)$$

$$(36) \qquad = \Theta^{k/2} \exp\left(-\frac{\Theta}{e} \frac{k}{2} \right),$$

where we set $t = m(\frac{1}{e} - \frac{1}{\Theta})$ to obtain the third inequality. Applying a union bound over the m rows and specifying $\Theta = e\lambda^{1-\alpha}$ we obtain for λ sufficiently

large:

(37)

$$\mathbb{P}(\|E\|_{\ell_2 \to \ell_\infty} \geq e^{1/2}\lambda^{-\alpha/2}) \leq me^{k/2}\lambda^{(1-\alpha)k/2} \exp\left(-\lambda^{1-\alpha}k/2\right)$$

(38)
$$= k\lambda \left(\lambda^{1-\alpha} \exp\left(-\frac{1}{2}\lambda^{1-\alpha}\right)\right)^{k/2} e^{-\frac{1}{4}m^{1-\alpha}k^\alpha}$$

(39)
$$=: k\lambda f(\lambda^{1-\alpha})^{-k/2} e^{-\frac{1}{4}m^{1-\alpha}k^\alpha}$$

(40)
$$\leq e^{-\frac{1}{4}m^{1-\alpha}k^\alpha},$$

where we used that f is independent of k and grows superlinearly, so above some threshold $C_{11}^{\frac{1}{1-\alpha}}$, $f(\lambda^{1-\alpha})^{-k/2}$ can absorb both λ and k.

$\square$

Remark 4.3. *Clearly, when the entries of $\widetilde{E}$ are bounded random variables, there exists a finite, deterministic upper bound on the operator norm $\|E\|_{\ell_2 \to \ell_\infty}$, obtained via the bounds on the matrix entries. In fact, for Bernoulli matrices, the resulting bounds are sharp.*

4.2. **Root-exponential accuracy for sub-Gaussian frames.**

Theorem 4.4. *Let $\widetilde{E}$ be an $m \times k$ sub-Gaussian matrix with mean zero, unit variance, and parameter c, let $E = \frac{1}{\sqrt{m}}\widetilde{E}$ and suppose that*

$$\lambda := \frac{m}{k} \geq C_{12},$$

where C_{12} is an appropriate constant that only depends on c. For a vector $x \in \mathbb{R}^k$, denote by $Q^{(r^*)}(Ex)$ the $2L$-level $\Sigma\Delta$ quantization of Ex using a scheme as given in (6) which satisfies (8) with order $r^* := \lfloor \lambda^{1/2}/(2C_{13}) \rfloor$ and $\delta > \frac{4e^{1/2}}{\lambda^{1/4}L}$. Denote by F the r^*-th order Sobolev dual of E. Then, with probability exceeding $1 - 3e^{-C_{14}\sqrt{mk}}$ on the draw of E, the reconstruction error satisfies

$$\|x - FQ^{(r)}(Ex)\|_2 \leq C_{15}\sqrt{k}e^{-C_{13}\lambda^{1/2}}\delta$$

uniformly for all x with $\|x\|_2 \leq 1$, where C_{13}, C_{14}, and C_{15} are appropriate constants depending only on c. If the entries of $\widetilde{E}$ are bounded by an absolute constant K almost surely, the condition on δ can be relaxed to $\delta > \frac{4K}{L}\lambda^{-1/2}$.

Proof. Observe the following facts:

(I) Since $FD^r = (D^{-r}E)^\dagger$, then $\|FD^r\| = \frac{1}{\sigma_{min}(D^{-r}E)} = \frac{1}{\sigma_{min}(SV^*E)}$, where $D^{-r} = USV^*$ is the singular value decomposition of D^{-r}. Thus, by Proposition 4.1 with $C_6 = 1/(3\pi r)$ and $\alpha = 1/2$, it holds that with probability greater than $1 - 2e^{-C_8\sqrt{mk}}$,

$$\|FD^r\| \leq \lambda^{-\frac{1}{2}(r-1/2)}$$

provided $\lambda \geq C_{16}^2 (3\pi r)^{\frac{4r}{(2r-1)}}$. Since $(3\pi r)^{\frac{1}{2r-1}}$ is decreasing with $r \geq 1$, the condition is satisfied if

$$(41) \qquad \lambda^{1/2} \geq C_{16}(3\pi)^2 r =: C_{17} r.$$

Without loss of generality, we may assume $C_{17} \geq 1$.

(II) By Proposition 4.2 with $\alpha = 1/2$ and for $\lambda > C_{11}^2$, we have

$$\|E\|_{\ell_2 \to \ell_\infty} \leq e^{1/2} \lambda^{-1/4}$$

with probability greater than $1 - e^{-\frac{1}{4}\sqrt{mk}}$. If the entries of E are bounded by $K/\sqrt{m}$ a.s. then

$$\|E\|_{\ell_2 \to \ell_\infty} = \max_{\eta \text{ is a row of E}} \|\eta\|_2 \leq K\lambda^{-1/2} \qquad \text{almost surely.}$$

(III) As this implies that the quantizer input satisfies $\|Ex\|_\infty \leq e^{1/2}\lambda^{-1/4}\|x\|_2$ or $\|Ex\|_\infty \leq K\lambda^{-1/2}\|x\|_2$, respectively, our assumption (8) implies that for all signals $x \in \mathbb{R}^k$ with $\|x\|_2 \leq 1$, the resulting state vector satisfies

$$\|u\|_\infty \leq C_{18} C_{19}^r r^r \delta/2,$$

where $C_{19} = \lceil \frac{\pi^2}{(\cosh^{-1}\gamma)^2} \rceil \frac{e}{\pi}$ for $\gamma = 2L - 2e^{1/2}\lambda^{-1/4}/\delta$ or $\gamma = 2L - 2K\lambda^{-1/2}/\delta$. By Cauchy-Schwarz, this entails

$$\|u\|_2 \leq C_{18} C_{19}^r r^r \delta/2\sqrt{m}.$$

The respective constraints on δ ensure that in both cases, $\frac{3}{2}L < \gamma < 2L$, so C_{19} is well-defined and bounded above by 10. Moreover, γ and C_{19} effectively do not depend on λ.

Thus, the reconstruction error $x - FQ^{(r)}(Ex) = FD^r u$ satisfies

$$\|FD^r u\|_2 \leq C_{18} C_{19}^r r^r \lambda^{-\frac{1}{2}(r-1/2)}\frac{\delta}{2}\sqrt{m} =: f(r).$$

Motivated by the observation

$$\arg\min_r f(r) = e^{-1}\frac{\lambda^{1/2}}{C_{19}}$$

we define $r^* = \left\lfloor \frac{e^{-1}\lambda^{1/2}}{C_{17}C_5} \right\rfloor$, which is adjusted to satisfy (41). As the above argument only works for $r \geq 1$, we need the additional requirement that $\lambda \geq C_{20} := 100e^2 C_{17}^2$. We estimate for such λ

$$(42)$$

$$f(r^*) \leq C_4 e \cdot m^{1/2}\lambda^{1/4}(C_{17}e)^{-\frac{1}{e}\frac{\lambda^{1/2}}{C_{17}C_5}}\delta/2 \leq C_4 e\sqrt{k}\lambda^{3/4}\exp\left(-\frac{1}{e}\frac{\lambda^{1/2}}{C_{17}C_5}\right)\delta/2$$

$$(43)$$

$$\leq C_{21}\sqrt{k}\lambda^{\frac{3}{4}}e^{-2C_{13}\lambda^{1/2}}\delta,$$

$$(44)$$

$$\leq C_{22}\sqrt{k}e^{-C_{13}\lambda^{1/2}}\delta,$$

where C_{13}, C_{21}, and C_{22} are appropriate constants. Here the last inequality stems from the observation that the polynomial factor is dominated by the exponential term. The theorem follows by choosing $C_{12} = \max\{C_{11}^2, C_{20}\}$.
□

By combining slight modifications of (I), (II), (III) in the proof of Theorem 4.4 above, one can deduce the following proposition which will be useful when dealing with quantization in the compressed sensing context. Below F is r-th order Sobolev dual of E and $Q^{(r)}(Ex)$ is the r-th order, $2L$-level greedy $\Sigma\Delta$ scheme, with step-size δ as in Section 2.1.

Proposition 4.5. *Let $\widetilde{E}$ be an $m \times k$ sub-Gaussian matrix with mean zero, unit variance, and parameter c, let $E = \frac{1}{\sqrt{m}}\widetilde{E}$, and fix an integer $r \geq 1$. Suppose that*

$$\lambda := \frac{m}{k} \geq C'_{12},$$

where C'_{12} is an appropriate constant that only depends on c and r. Let $0 < \alpha < 1$. For an integer $L \geq \lceil \frac{e^{1/2}\lambda^{-\alpha/2}}{\delta} \rceil + 2^r + 1$, with probability exceeding $1 - 3e^{-C'_{14}m^{1-\alpha}k^{\alpha}}$ on the draw of E, the reconstruction error satisfies

$$\|x - FQ^{(r)}(Ex)\|_2 \leq \lambda^{-\frac{\alpha}{2}(r-1/2)}\frac{\delta}{2}\sqrt{m}$$

uniformly for all x with $\|x\|_2 \leq 1$, where C'_{14}, is an appropriate constant depending only on c. If the entries of $\widetilde{E}$ are bounded by an absolute constant K almost surely, the condition on δ can be relaxed to $L \geq \lceil \frac{K\lambda^{-1/2}}{\delta} \rceil + 2^r + 1$.

Proof. The proof proceeds along the same lines as the proof of Thm 17. In fact, parts (I) and (II) are identical, albeit with a general α rather than $\alpha = \frac{1}{2}$. In part (III) we obtain that the state vector satisfies

$$\|u\|_\infty \leq \delta/2,$$

which entails

$$\|u\|_2 \leq \frac{\delta}{2}\sqrt{m}.$$

Combining (I) and (III) we see that

$$\|FD^r u\|_2 = \|x - FQ^{(r)}(Ex)\|_2 \leq \lambda^{-\frac{\alpha}{2}(r-1/2)}\frac{\delta}{2}\sqrt{m}.$$

Since the involved constants only depend on c, the statement follows.   □

Remark 4.6. *Note that in Theorem 4.4, the normalization of the entries depends on the number of measurements taken. In particular, this entails that the number of measurements needs to be assigned a-priori. In practice, however, one may wish to acquire as many measurements as possible and quantize the measurements on the fly. This can be accomplished when the entries of $\widetilde{E}$ are bounded by some constant K. In this case, E consists of entries bounded by $K/\sqrt{m}$, so Theorem 4.4 predicts root-exponential accuracy for all quantization levels $\delta > 2K\lambda^{-1/2}$. Hence the quantization level*

can be chosen independently of m even if the whole system is multiplied by $\sqrt{\lambda}$. This multiplication yields normalized rows rather than columns, so the normalization does not depend on m, as desired.

4.3. **Polynomial accuracy for compressed sensing quantization.** Analogously to [17], the above bounds on the singular values can be used to establish recovery guarantees from $\Sigma\Delta$ quantized compressed sensing measurements. To obatin these guarantees, as in [17], the decoding is performed via a two-stage algorithm: First, a *robust* compressed sensing decoder is used to recover a coarse estimate and particularly the support of the underlying sparse signal from the quantized measurements. Next, the estimate is refined using a Sobolev-dual based reconstruction method.

Definition 4.7. *The restricted isometry constant (see, e.g., [18]) γ_k of a matrix $\Phi \in \mathbb{R}^{m \times N}$ is the smallest constant for which*

$$(1 - \gamma_k)\|x\|_2^2 \leq \|\Phi x\|_2^2 \leq (1 + \gamma_k)\|x\|_2^2$$

for all $x \in \Sigma_k^N$.

Definition 4.8. *Let $\varepsilon > 0$, let m, N be positive integers such that $m < N$ and suppose that $\Phi \in \mathbb{R}^{m \times N}$. We say that $\Delta : \mathbb{R}^{m \times N} \times \mathbb{R}^m \to \mathbb{R}^N$ is a robust compressed sensing decoder with parameters (k, a, γ), $k < m$, and constant C if*

$$(45) \qquad\qquad \|x - \Delta(\Phi, \Phi x + e)\| \leq C\varepsilon,$$

for all $x \in \Sigma_k^N$, $\|e\|_2 \leq \varepsilon$, and all matrices Φ with a restricted isometry constant $\gamma_{ak} < \gamma$.

Remark 4.9. *Examples for robust compressed sensing decoders include ℓ_1-minimization [18], greedy algorithms such as orthogonal matching pursuit [28], and greedy-type algorithms such as CoSaMP [29].*

One then obtains the following result.

Theorem 4.10. *Let $r \in \mathbb{Z}^+$, fix $a \in \mathbb{N}$, $\gamma < 1$, and $c, C > 0$. Then there exist constants $C_{23}, C_{24}, C_{25}, C_{26}$ depending only on these parameters such that the following holds.*

Fix $0 < \alpha < 1$, let Φ be an $m \times N$ sub-Gaussian matrix with mean zero, unit variance, and parameter c, and suppose that $k \in \mathbb{N}$ satisfies

$$\lambda := \frac{m}{k} \geq \left(C_{23} \log(eN/k) \right)^{\frac{1}{1-\alpha}}.$$

Furthermore, suppose that $\Delta : \mathbb{R}^{m \times N} \times \mathbb{R}^m \to \mathbb{R}^N$ is a robust compressed sensing decoder with parameters (k, a, γ) and constant C.

Consider the $2L$-level rth order greedy $\Sigma\Delta$ schemes with step-size δ, $L \geq \lceil \frac{K\lambda^{-1/2}}{\delta} \rceil + 2^r + 1$. Denote by q the quantization output resulting from Φz where $z \in \mathbb{R}^N$. Then with probability exceeding $1 - 4e^{-C_{24} m^{1-\alpha} k^\alpha}$ for all $z \in \Sigma_k^N$ having $\min\limits_{j \in \text{supp}(z)} |z_j| > C_{25}\delta$:

(i) *The support of z, T, coincides with the support of the best k-term approximation of $\Delta(q)$.*

(ii) *Denoting by E and F the sub-matrix of Φ corresponding to the support of z and its rth order Sobolev dual respectively, and by $x \in \mathbb{R}^k$ the restriction of z to its support, we have*

$$\|x - Fq\|_2 \leq C_{26}\lambda^{-\alpha(r-1/2)}\delta.$$

The proof essentially traces the same steps as in [17]. In particular, as Φ is a sub-Gaussian matrix, and by choosing C_{23} large enough, we can ensure that $m \geq 2C_{27}\gamma^{-2}ak\ln\frac{eN}{ak}$ and obtain $\gamma_{ak} \leq \gamma$ with probability exceeding $1 - 2e^{-\gamma^2 m/(2C_{27})}$ (see, e.g., [30, Theorem 9.2]). Here C_{27} depends only c. Since Δ is a robust compressed sensing decoder, (i) follows from [17, Proposition 4.1].

Once the support of x, say T, is recovered, Proposition 4.5 applies to the $m \times k$ sub-Gaussian matrix Φ_T, the submatrix of Φ consisting of columns of Φ indexed by T. Finally, (ii) follows from a union bound over all submatrices consisting of k columns of Φ. This union bound, coupled with the probability bound used to obtain a restricted isometry constant $\gamma_{ak} < \gamma$ yields the probability in the statement of the theorem as well as the condition on λ. As all the proof ingredients established above follow closely the corresponding results in [17], we omit the details.

Remark 4.11. *Note that root-exponential error decay, using quantization with a fixed number of levels, as obtained for the frame quantization case in Theorem 4.4 is hindered in the compressed sensing scenario by the support recovery step of the two stage algorithm in [17].*

Funding

Acknowledgment

References

[1] P. G. Casazza and G. Kutyniok, Eds., *Finite frames: Theory and applications.* Springer, 2013.

[2] Y. C. Eldar and G. Kutyniok, Eds., *Compressed Sensing: Theory and Applications.* Cambridge University Press Cambridge, 2012, vol. 95.

[3] V. Goyal, M. Vetterli, and N. Thao, "Quantized overcomplete expansions in $\mathbb{R}^N$: analysis, synthesis, and algorithms," *IEEE Trans. Inf. Theory*, vol. 44, no. 1, pp. 16–31, Jan 1998.

[4] H. Inose and Y. Yasuda, "A unity bit coding method by negative feedback," *Proc. IEEE*, vol. 51, no. 11, pp. 1524–1535, 1963.

[5] 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. Math.*, vol. 158, no. 2, pp. 679–710, 2003.

[6] C. Güntürk, "One-bit sigma-delta quantization with exponential accuracy," *Comm. Pure Appl. Math.*, vol. 56, no. 11, pp. 1608–1630, 2003.

[7] P. Deift, C. S. Güntürk, and F. Krahmer, "An optimal family of exponentially accurate one-bit sigma-delta quantization schemes," *Comm. Pure Appl. Math.*, vol. 64, no. 7, pp. 883–919, 2011.

[8] A. R. Calderbank and I. Daubechies, "The pros and cons of democracy," *IEEE Trans. Inf. Theory*, vol. 48, no. 6, pp. 1721–1725, 2002, special issue on Shannon theory: perspective, trends, and applications.

[9] F. Krahmer and R. Ward, "Lower bounds for the error decay incurred by coarse quantization schemes," *Appl. Comput. Harmonic Anal.*, vol. 32, no. 1, pp. 131–138, 2012.

[10] J. Benedetto, A. Powell, and O. Yılmaz, "Sigma-delta ($\Sigma\Delta$) quantization and finite frames," *IEEE Trans. Inf. Theory*, vol. 52, no. 5, pp. 1990–2005, 2006.

[11] B. G. Bodmann, V. I. Paulsen, and S. A. Abdulbaki, "Smooth frame-path termination for higher order sigma-delta quantization," *J. Fourier Anal. and Appl.*, vol. 13, no. 3, pp. 285–307, 2007.

[12] J. J. Benedetto, A. M. Powell, and Ö. Yılmaz, "Second-order sigma-delta ($\Sigma\Delta$) quantization of finite frame expansions," *Applied and Computational Harmonic Analysis*, vol. 20, no. 1, pp. 126 – 148, 2006.

[13] M. Lammers, A. Powell, and Ö. Yılmaz, "Alternative dual frames for digital-to-analog conversion in sigma–delta quantization," *Adv. Comput. Math.*, pp. 1–30, 2008.

[14] J. Blum, M. Lammers, A. Powell, and O. Yılmaz, "Sobolev duals in frame theory and sigma-delta quantization," *J. Fourier Anal. and Appl.*, vol. 16, no. 3, pp. 365–381, 2010.

[15] F. Krahmer, R. Saab, and R. Ward, "Root-exponential accuracy for coarse quantization of finite frame expansions," *IEEE Trans. Inf. Theory*, vol. 58, no. 2, pp. 1069 –1079, 2012.

[16] A. Powell, R. Saab, and Ö. Yılmaz, "Quantization and finite frames," in *Finite Frames: Theory and Aplications*, P. Casazza and G. Kutinyok, Eds. Springer, 2013, pp. 305–328.

[17] C. Güntürk, M. Lammers, A. Powell, R. Saab, and Ö. Yılmaz, "Sobolev duals for random frames and quantization of compressed sensing measurements," *Foundations of Computational Mathematics*, vol. 13, no. 1, pp. 1–36, 2013.

[18] E. J. Candès, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Comm. Pure Appl. Math.*, vol. 59, pp. 1207–1223, 2006.

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

[20] R. Vershynin, "Introduction to the non-asymptotic analysis of random matrices," in *Compressed Sensing: Theory and Applications*, Y. Eldar and G. Kutyniok, Eds. Cambridge: Cambridge Univ Press, 2012, pp. xii+544.

[21] L. Jacques, D. Hammond, and J. Fadili, "Dequantizing compressed sensing: When oversampling and non-gaussian constraints combine," *IEEE Trans. Inf. Theory*, vol. 57, no. 1, pp. 559–571, 2011.

[22] L. Jacques, J. Laska, P. Boufounos, and R. Baraniuk, "Robust 1-bit compressive sensing via binary stable embeddings of sparse vectors," *IEEE Trans. Inf. Theory*, vol. 59, no. 4, pp. 2082–2102, 2013.

[23] Y. Plan and R. Vershynin, "One-bit compressed sensing by linear programming," *arXiv preprint arXiv:1109.4299*, 2011.

[24] A. Ai, A. Lapanowski, Y. Plan, and R. Vershynin, "One-bit compressed sensing with non-gaussian measurements," *Lin. Alg. Appl.*, 2013, to appear.

[25] R. Horn and C. Johnson, *Matrix analysis.* Cambridge: Cambridge University Press, 1990, corrected reprint of the 1985 original.

[26] M. Ledoux and M. Talagrand, *Probability in Banach Spaces: Isoperimetry and Processes.* New York: Springer, 1991.

[27] F. Krahmer, S. Mendelson, and H. Rauhut, "Suprema of chaos processes and the restricted isometry property," *Comm. Pure Appl. Math.*, 2013, to appear.

[28] T. Zhang, "Sparse recovery with orthogonal matching pursuit under RIP," *IEEE Trans. Inform. Theory*, vol. 57, no. 9, pp. 6215 – 6221, 2011.

[29] D. Needell and J. Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples," *Appl. Comput. Harmon. Anal.*, vol. 26, no. 3, pp. 301–321, 2008.

[30] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing.* Springer, 2013, to appear.