NEAR-OPTIMAL ENCODING FOR SIGMA-DELTA QUANTIZATION OF FINITE FRAME EXPANSIONS

MARK IWEN AND RAYAN SAAB

Abstract. In this paper we investigate encoding the bit-stream resulting from coarse Sigma-Delta quantization of finite frame expansions (i.e., overdetermined representations) of vectors. We show that for a wide range of finite-frames, including random frames and piecewise smooth frames, there exists a simple encoding algorithm —acting only on the Sigma-Delta bit stream— and an associated decoding algorithm that together yield an approximation error which decays exponentially in the number of bits used. The encoding strategy consists of applying a discrete random operator to the Sigma-Delta bit stream and assigning a binary codeword to the result. The reconstruction procedure is essentially linear and equivalent to solving a least squares minimization problem.

Key words: Vector quantization, frame theory, rate-distortion theory, random matrices, overdetermined systems, pseudoinverses

AMS subject classifications: 42C15, 94A12, 94A34, 65F20, 15B52, 68Q17

1. Introduction

In the modern era, the first step in signal processing consists of obtaining a digital representation of the signal of interest, i.e., quantizing it. This enables one to store, transmit, process, and analyze the signal via digital devices. Sigma-Delta ($\Sigma\Delta$) quantization was proposed in the 1960's as a quantization scheme for digitizing band-limited signals (see, e.g., [18]). Since then, and especially with the advent of very large scale integration (VLSI) technology, $\Sigma\Delta$ schemes have seen extensive use in the engineering community for analog-to-digital conversion of, for example, audio signals (cf. [25]). In the mathematical community, $\Sigma\Delta$ quantization has seen increasing interest since the work of Daubechies and Devore [8]. In this paper, we are interested in efficiently encoding the bit-stream resulting from $\Sigma\Delta$ quantization of finite-frame expansions. Here one models the signal as an element in a finite dimensional space, and its samples as inner products with a spanning set of vectors. The goal is, using only the samples, to obtain a digital representation of the original signal that allows its high fidelity reconstruction.

1.1. Overview and prior work.

For concreteness, let the vectors $\{\mathbf{f}_i\}_{i=1}^N \subset \mathbb{R}^d$ form a finite frame for $\mathbb{R}^d$. In other words, suppose there exist constants $0 < A \leq B < \infty$ such that the *frame matrix* $F \in \mathbb{R}^{N\times d}$ (with the vectors $\mathbf{f}_i$ as its rows) satisfies

$$A\|\mathbf{x}\|_2^2 \leq \|F\mathbf{x}\|_2^2 \leq B\|\mathbf{x}\|_2^2$$

M.I. was supported in part by NSA grant H98230-13-1-0275. R.S. was supported in part by a Banting Postdoctoral Fellowship, administered by the Natural Sciences and Engineering Research Council of Canada (NSERC). The majority of the work reported on herein was completed while the authors were visiting assistant professors at Duke University.

for all $x \in \mathbb{R}^d$. Thus any full rank matrix is a frame matrix. In the context of data acquisition, finite frames are useful at modeling the sampling (i.e., measurement) process. In various applications, the measurement vector can be expressed as

$$\mathbf{y} := F\mathbf{x} \in \mathbb{R}^N. \tag{1}$$

For example, in imaging applications, "multiplex" systems (see, e.g., [6]) collect linear combinations of the pixels of interest, thus their measurement vectors can the be represented using (1). Such systems have been devised using coded apertures (see, e.g., [20]), as well as digital micro-mirror arrays (e.g., [10]). Indeed, by simply collecting more measurements than the ambient dimension of the image, as is often the case, and ensuring that F is full rank, we find ourselves in the finite frame setting. Similarly, systems that acquire finite dimensional signals using filter banks allow their measurement process to be modeled via (1). For a more in-depth treatment of finite frames and filter banks, see [11].

In order to allow digital storage and computer processing, including the recovery of $\mathbf{x}$, one must quantize the finite frame expansion (1). Quantizing the finite frame expansion of $\mathbf{x}$ consists of replacing the entries of the measurement vector $\mathbf{y} := F\mathbf{x} \in \mathbb{R}^N$ with elements from a finite set. Giving these elements binary labels then enables digital storage and transmission of the quantized measurements. To be precise, let $\mathcal{A} \subset \mathbb{R}$ be a finite set (the quantization alphabet), and let $\mathcal{X}$ be a compact set in $\mathbb{R}^d$. A quantization scheme is a map

$$\mathcal{Q} : F\mathcal{X} \mapsto \mathcal{A}^N$$

and a reconstruction scheme is an "inverse" map

$$\Delta : \mathcal{A}^N \mapsto \mathbb{R}^d.$$

Note that, depending on the practical application, one may require that the quantization schemes satisfy certain properties. For example, it is often preferred that the quantization scheme acts progressively on the measurements, i.e., as they arrive, to avoid storing too many analog quantities. Nevertheless, one seeks quantization and reconstruction schemes with approximation errors $\|\mathbf{x} - \Delta(\mathcal{Q}(F\mathbf{x}))\|_2$ that are as small as possible for all $\mathbf{x} \in \mathcal{X}$.

$\Sigma\Delta$ quantization schemes form an important class of such progressive quantizers, and there has been much work focusing on their application to the finite frame setting. In particular, the research on $\Sigma\Delta$ quantization of finite frame expansions has focused on the decay of the approximation error as a function of the number of measurements and has typically considered $\mathcal{X} = \mathcal{B}^d$, the Euclidean ball in $\mathbb{R}^d$. The work of Benedetto, Powell, and Yılmaz [4] first showed that the reconstruction error associated with 1st order $\Sigma\Delta$ quantization decays linearly in the number of measurements. Several results followed, improving on this linear error decay by using various combinations of specialized frames, higher order quantization schemes, and different reconstruction techniques. Blum et al. [5] showed that frames with certain smoothness properties allow for polynomial decay in the $\Sigma\Delta$ reconstruction error, provided appropriate alternative dual frames are used for reconstruction. Motivated by applications in compressed sensing, a similar result [16] was shown for random frames whose elements are Gaussian random variables. This was followed by the results of [21] and [22] showing that there exist (deterministic and random) frames for which higher order $\Sigma\Delta$ schemes yield approximation errors that behave like $e^{-c\sqrt{\frac{N}{d}}}$, where c is a constant. Specifically, in [21, 22] this root-exponential accuracy is achieved by carefully choosing the order

of the scheme as a function of the oversampling rate N/d. For a more comprehensive review of $\Sigma\Delta$ schemes applied to finite frames, see [26].

While the above results progressively improved on the *coding efficiency* of $\Sigma\Delta$ quantization, it remains true that even the root-exponential performance $e^{-c\sqrt{\frac{N}{d}}}$ of [21, 22] is generally sub-optimal from an information theoretic perspective, including in the case where $\mathcal{X} = \mathcal{B}^d$. To be more precise, any quantization scheme tasked with encoding all possible points in $\mathcal{B}^d$ to within ϵ-accuracy must produce outputs which, in the case of an optimal encoder, each correspond to a unique subset of the unit ball having radius at most ϵ. Covering each of these subsets with a ball of radius at most ϵ then produces an ϵ-cover of $\mathcal{B}^d$. A simple volume argument now shows that covering $\mathcal{B}^d$ with balls of radius ϵ requires one to use at least $\left(\frac{1}{\epsilon}\right)^d$ such ϵ-balls.[1] Thus, quantizing $\mathcal{B}^d$ via an optimal map requires at least $d \ln \frac{c}{\epsilon}$ bits, or, viewed slightly differently: optimally quantizing $\mathcal{B}^d$ with b-bits yields an approximation error of the form $e^{-c\frac{b}{d}}$, where $c \in \mathbb{R}^+$ is a universal constant. Observing that the number of bits that result from a $\Sigma\Delta$ scheme is proportional to N and that the best known error rates are root-exponential in N, we conclude that $\Sigma\Delta$ schemes are sub-optimal.

This fact has been recognized in the mathematical literature on $\Sigma\Delta$ quantization. In particular, in the case where $d = 1$ and the frame F is the $N \times 1$ repetition frame with $F_{i1} = 1$ for all $i \in [N]$, there has been research seeking upper bounds on the maximum number of possible $\Sigma\Delta$ bit-streams (cf. [17], [15], [2]). For example, [17] showed that asymptotically in N, the number of bit-streams is bounded by $O(N^2)$ for first order single-bit $\Sigma\Delta$ schemes with certain initial conditions. This indicates that by losslessly encoding the possible $\Sigma\Delta$ bitstreams into codewords of length $O(\log(N))$, one can achieve the desired exponential error rates. However, to do that one needs to identify the $O(N^2)$ achievable sequences from among the 2^N potential ones, which to our knowledge is an unsolved problem. Moreover, to our knowledge, not much is known about the number of codewords generated by $\Sigma\Delta$ quantization in more general settings.

To help remedy this situation, this paper introduces a potentially *lossy encoding* stage, consisting of the map

$$\mathcal{E} : \mathcal{A}^N \mapsto \mathcal{C},$$

where $\mathcal{C}$ is such that $|\mathcal{C}| \ll |\mathcal{A}^N|$. Consequently, $\log_2 |\mathcal{C}|$ bits are sufficient for digitally representing the output of this encoder. To accommodate this additional encoding, the reconstruction is modified to approximate $\mathbf{x}$ directly from $\mathcal{C}$. Thus, we propose a decoder

$$\Delta : \mathcal{C} \mapsto \mathbb{R}^d,$$

where both the proposed decoder, Δ, and the proposed encoding map, $\mathcal{E}$, are *linear*, hence computationally efficient.

1.2. **Contributions.** For stable $\Sigma\Delta$ quantization schemes, we show that there exists an encoding scheme $\mathcal{E}$ acting on the output $\mathcal{Q}(F\mathbf{x})$ of the quantization, and a decoding scheme Δ, such that

$$\left.\begin{array}{rcl} \epsilon_{\Sigma\Delta} & := & \max\limits_{\mathbf{x}\in\mathcal{B}^d} \left\| \mathbf{x} - \Delta\Big(\mathcal{E}\big(\mathcal{Q}(F\mathbf{x})\big)\Big) \right\|_2 \leq CN^{-\alpha} \\ b_{\Sigma\Delta} & := & \ln|\mathcal{C}| \leq C' d \ln N, \end{array}\right\} \implies \epsilon_{\Sigma\Delta} \leq \exp\left(-c\frac{b_{\Sigma\Delta}}{d}\right).$$

[1] Moreover, there exists a covering with no more than $\left(\frac{3}{\epsilon}\right)^d$ elements (see, e.g., [24]).

where α, C, C', and c are positive constants that depend on the $\Sigma\Delta$ scheme and d. More specifically:

1. We show that there exist frames (the Sobolev self-dual frames), for which encoding by random subsampling of the integrated $\Sigma\Delta$ bit-stream (and labeling the output) yields an *essentially optimal rate-distortion tradeoff up to logarithmic factors of d.*
2. We show that random *Bernoulli* matrices in $\mathbb{R}^{m\times d}$, with $m \approx d$, are *universal* encoders. Provided one has a good frame for $\Sigma\Delta$ quantization, such Bernoulli matrices yield an *optimal rate-distortion tradeoff, up to constants.*
3. We show that in both cases above, the decoding can be done linearly and we provide an explicit expression for the decoder.

These contributions are made explicit in Theorems 3 and 4. Additionally, we note that $\Sigma\Delta$ schemes (see Section 2) act progressively on the samples $\mathbf{y} = F\mathbf{x}$, and do not require explicit knowledge of the frame that produced $\mathbf{y}$. Similarly, the Bernoulli encoding of the $\Sigma\Delta$ bit-stream does not require knowledge of the underlying frame. Nevertheless, and somewhat surprisingly, this encoding method allows the compression of the $\Sigma\Delta$-bitstream in a near optimal manner. It also allows the decoding to be done linearly via an operator $\mathbb{R}^m \mapsto \mathbb{R}^d$, hence in time proportional to md, as opposed to time proportional to Nd needed (in general) for reconstructing a signal from its unencoded $\Sigma\Delta$ bitstream. One of the favorable properties of coarse $\Sigma\Delta$ quantization schemes is their robustness to certain errors that can arise in practice due to (for example) circuit imperfections (cf. [8]). Such imperfections can affect the elements that implement scalar quantization (i.e., assigning discrete values to continuous ones by toggling at a threshold), or multiplication. We remark that our methods for compressing the $\Sigma\Delta$ bit-stream inherit whatever robustness properties the original $\Sigma\Delta$ quantizer possesses. In other words, by compressing the bit-stream, we do not lose any of the desirable properties of $\Sigma\Delta$ quantization.

1.3. **Organization.** In Section 2, we introduce notation and provide a mathematical overview of $\Sigma\Delta$ quantization. We also state certain results on random matrices, in particular Johnson-Lindenstrauss embeddings, which will be useful in the remainder of the paper. In Section 3 we show that random subsampling of the discretely integrated $\Sigma\Delta$ bit-stream allows a linear decoder to achieve exponentially decaying reconstruction error, uniformly for all $\mathbf{x} \in \mathcal{B}^d$. This result pertains to a particular choice of frames, the Sobolev self-dual frames [21], and is contingent on using 1st order $\Sigma\Delta$ schemes. In Section 4 we instead use a Bernoulli matrix for reducing the dimensionality of the integrated $\Sigma\Delta$-bit stream. Here, our result is more general and applies to stable $\Sigma\Delta$ schemes of arbitrary order, as well as to a large family of smooth and random frames. Finally, in Section 5 we illustrate our results with numerical experiments.

2. Preliminaries

Below we will denote the set $\{1, 2, \ldots, n-1, n\} \subset \mathbb{N}$ by $[n]$. For any matrix $M \in \mathbb{R}^{m\times N}$ we will denote the j^{th} column of M by $\mathbf{M}_j \in \mathbb{R}^m$. Furthermore, for a given subset $\mathcal{S} = \{s_1, \ldots, s_n\} \subset [N]$ with $s_1 < s_2 < \cdots < s_n$, we will let $M_\mathcal{S} \in \mathbb{R}^{m\times n}$ denote the submatrix of M given by

$$M_\mathcal{S} := (\mathbf{M}_{s_1} \ldots \mathbf{M}_{s_n}) .$$

The transpose of a matrix, $M \in \mathbb{R}^{m\times N}$, will be denoted by $M^{\mathrm{T}} \in \mathbb{R}^{N\times m}$, and the singular values of any matrix $M \in \mathbb{R}^{m\times N}$ will always be ordered as $\sigma_1(M) \geq \sigma_2(M) \geq \cdots \geq \sigma_{\min(m,N)}(M) \geq 0$. We will denote the standard indicator function by

$$\delta_{i,j} := \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases},$$

for $i, j \in \mathbb{N}$. Finally, given a frame matrix F, we define its Moore-Penrose pseudo-inverse to be $F^\dagger := (F^T F)^{-1} F^T$.

2.1. **Sigma-Delta Quantization.** Let $\mathcal{B}^d$ be the Euclidean unit ball in $\mathbb{R}^d$. Given $\mathbf{x} \in \mathcal{B}^d$, and a frame matrix $F \in \mathbb{R}^{N\times d}$, the simplest $\Sigma\Delta$ quantization scheme considered herein, in Section 3, is the *single bit first order greedy scheme*. Given $\mathbf{y} = F\mathbf{x}$, this scheme computes a vector $\mathbf{q} \in \{-1, 1\}^N$ via the following recursion with initial condition $u_0 = 0$:

$$q_i = \operatorname{sign}(y_i + u_{i-1}), \tag{2}$$

$$u_i = y_i + u_{i-1} - q_i \tag{3}$$

for all $i \in [N]$. To analyze this scheme as well as higher order schemes, it will be convenient to introduce the *difference matrix*, $D \in \mathbb{R}^{N\times N}$, given by

$$D_{i,j} := \begin{cases} 1 & \text{if } i = j \\ -1 & \text{if } i = j + 1 \\ 0 & \text{otherwise} \end{cases}. \tag{4}$$

We may restate the relationships between $\mathbf{x}$, $\mathbf{u}$, and $\mathbf{q}$ resulting from the above scheme as

$$D\mathbf{u} \;=\; F\mathbf{x} - \mathbf{q}. \tag{5}$$

Furthermore, a short induction argument shows that $|u_i| \leq 1$ for all $i \in [N]$ provided that $|y_i| \leq 1$ for all $i \in [N]$.

More generally, for a given alphabet $\mathcal{A}$ and $r \in \mathbb{Z}^+$ we may employ an r^{th}*-order* $\Sigma\Delta$ *quantization scheme* with quantization rule $\rho : \mathbb{R}^{r+1} \mapsto \mathbb{R}$ and scalar quantizer $Q : \mathbb{R} \mapsto \mathcal{A}$. Such a scheme, with initial conditions $u_0 = u_{-1} = \cdots = u_{1-r} = 0$, computes $\mathbf{q} \in \mathcal{A}^N$ via the recursion

$$q_i = Q\left(\rho(y_i, u_{i-1}, u_{i-2}, \ldots, u_{i-r})\right), \tag{6}$$

$$u_i = y_i - q_i - \sum_{j=1}^{r} \binom{r}{j} (-1)^j u_{i-j} \tag{7}$$

for all $i \in [N]$. Here, the scalar quantizer Q is defined via its action

$$Q(v) = \operatorname*{arg\,min}_{q\in\mathcal{A}} |q - v|.$$

In this paper, we focus on *midrise* alphabets of the form

$$\mathcal{A}_K^\delta = \{\pm(2n-1)\delta/2 : n \in [K]\}, \tag{8}$$

where δ denotes the quantization step size. For example, when $K = 1$, we have the 1-bit alphabet $\mathcal{A}_1^\delta = \{\pm\frac{\delta}{2}\}$. As before, we may restate the relationships between $\mathbf{x}$, $\mathbf{u}$, and $\mathbf{q}$ as

$$D^r\mathbf{u} \;=\; F\mathbf{x} - \mathbf{q}. \tag{9}$$

As in the case of the first order scheme, we will ultimately need a bound on $\|\mathbf{u}\|_\infty := \max_{i\in[N]} |u_i|$ below. Hence, we restrict our attention to *stable r^{th}-order schemes*. That is, r^{th}-order schemes for which (6) and (7) are guaranteed to always produce vectors $\mathbf{u} \in \mathbb{R}^N$ having $\|\mathbf{u}\|_\infty \leq C_{\rho,Q}(r)$ for all $N \in \mathbb{N}$, and $\mathbf{y} \in \mathbb{R}^N$ with $\|\mathbf{y}\|_\infty \leq 1$. Moreover, for our definition of stability we require that $C_{\rho,Q} : \mathbb{N} \mapsto \mathbb{R}^+$ be entirely independent of both N and $\mathbf{y}$. Finally, it is important to note that stable r^{th}-order $\Sigma\Delta$ schemes with $C_{\rho,Q}(r) = O(r^r)$ do indeed exist (see, e.g., [14, 9]), even when $\mathcal{A}$ is a 1-bit alphabet. In particular, we cite the following proposition [9] (cf. [21]).

Proposition 1. *There exists a universal constant $c > 0$ such that for any midrise quantization alphabet $\mathcal{A} = \mathcal{A}_L^\delta$, for any order $r \in \mathbb{N}$, and for all $\eta < \delta\left(L - \frac{1}{2}\right)$, there exists an rth order $\Sigma\Delta$ scheme which is stable for all input signals y with $\|y\|_\infty \leq \eta$. It has*

$$\|u\|_\infty \leq cC^r r^r \frac{\delta}{2}, \tag{10}$$

where $C = \left(\left\lceil \frac{\pi^2}{(\cosh^{-1}\gamma)^2}\right\rceil \frac{e}{\pi}\right)$ and $\gamma := 2L - \frac{2\eta}{\delta}$.

In what follows we will need the singular value decomposition of D essentially computed by von Neumann in [27] (see also [21]). It is $D = U\Sigma V^T$, where

$$U_{i,j} = \sqrt{\frac{2}{N+1/2}} \cos\left(\frac{2(i-1/2)(N-j+1/2)\pi}{2N+1}\right), \tag{11}$$

$$\Sigma_{i,j} = \delta_{i,j}\sigma_j(D) = 2\delta_{i,j}\cos\left(\frac{j\pi}{2N+1}\right), \tag{12}$$

and

$$V_{i,j} = (-1)^{j+1}\sqrt{\frac{2}{N+1/2}} \sin\left(\frac{2ij}{2N+1}\pi\right). \tag{13}$$

Note that the difference matrix, D, is full rank (e.g., see (12)). Thus, we may rearrange (5) to obtain

$$\mathbf{u} = D^{-1}F\mathbf{x} - D^{-1}\mathbf{q}. \tag{14}$$

More generally, rearranging (9) tells us that

$$\mathbf{u} = D^{-r}F\mathbf{x} - D^{-r}\mathbf{q} \tag{15}$$

for any r^{th}-order scheme.

2.2. **Johnson-Lindenstrauss Embeddings and Bounded Orthonormal Systems.** We will utilize *linear Johnson-Lindenstrauss embeddings* [19, 13, 1, 7, 3, 23] of a given finite set $\mathcal{S} \subset \mathbb{R}^N$ into $\mathbb{R}^m$.

Definition 1. *Let $\epsilon, p \in (0,1)$, and $\mathcal{S} \subset \mathbb{R}^N$ be finite. An $m \times N$ matrix M is a linear Johnson-Lindenstrauss embedding of $\mathcal{S}$ into $\mathbb{R}^m$ if the following holds with probability at least $1-p$:*

$$(1-\epsilon)\| \mathbf{u}-\mathbf{v} \|_2^2 \leq \| M\mathbf{u} - M\mathbf{v} \|_2^2 \leq (1+\epsilon)\| \mathbf{u}-\mathbf{v} \|_2^2$$

for all $\mathbf{u}, \mathbf{v} \in \mathcal{S}$. In this case we will say that M is a JL(N,m,ϵ,p)-embedding of $\mathcal{S}$ into $\mathbb{R}^m$.

We will say that a matrix $B \in \{-1,1\}^{m\times N}$ is a *Bernoulli random matrix* iff each of its entries is independently and identically distributed so that

$$\mathbb{P}\left[B_{i,j}=1\right] \;=\; \mathbb{P}\left[B_{i,j}=-1\right] \;=\; \frac{1}{2}$$

for all $i \in [m]$ and $j \in [N]$. The following theorem is proven in [1].

Theorem 1. *Let $m, N \in \mathbb{N}$, $\mathcal{S} \subset \mathbb{R}^N$ finite, and $\epsilon, p \in (0,1)$. Let $B \in \{-1,1\}^{m\times N}$ be a Bernoulli random matrix, and set $\widetilde{B} = \frac{1}{\sqrt{m}}B$. Then, $\widetilde{B}$ will be a JL(N,m,ϵ,p)-embedding of $\mathcal{S}$ into $\mathbb{R}^m$ provided that $m \geq \frac{4+2\log_{|\mathcal{S}|}(1/p)}{\epsilon^2/2-\epsilon^3/3}\ln|\mathcal{S}|$.*

Let $\mathcal{D} \subset \mathbb{R}^n$ be endowed with a probability measure μ. Further, let $\Psi = \{\psi_1, \ldots, \psi_N\}$ be an orthonormal set of real-valued functions on $\mathcal{D}$ so that

$$\int_{\mathcal{D}} \psi_i\left(\,\mathbf{t}\,\right)\overline{\psi_j\left(\,\mathbf{t}\,\right)}d\mu\left(\mathbf{t}\right) \;=\; \delta_{i,j}.$$

We will refer to any such Ψ as an *orthonormal system*. More specifically, we utilize a particular type of orthonormal system:

Definition 2. *We call $\Psi = \{\psi_1, \ldots, \psi_N\}$ a bounded orthonormal system with constant $K \in \mathbb{R}^+$ if*

$$\left\|\psi_k\right\|_\infty := \sup_{\mathbf{t}\,\in\,\mathcal{D}} \left|\psi\left(\,\mathbf{t}\,\right)\right| \;\leq\; K \text{ for all } k \in [N].$$

For any orthonormal system, Ψ, on $\mathcal{D} \subset \mathbb{R}^n$ with probability measure μ, we may create an associated *random sampling matrix*, $R' \in \mathbb{R}^{m\times N}$, as follows: First, select m points $\mathbf{t}_1, \ldots, \mathbf{t}_m \in \mathcal{D}$ independently at random according to μ.[2] Then, form the matrix R' by setting $R'_{i,j} := \psi_j\left(\,\mathbf{t}_i\,\right)$ for each $i \in [m]$ and $j \in [N]$. The following theorem concerning random sampling matrices created from bounded orthonormal systems is proven in [12].[3]

Theorem 2. *Let $R' \in \mathbb{R}^{m\times N}$ be a random sampling matrix created from a bounded orthonormal system with constant K. Let $\mathcal{S} \subset [N]$ have have cardinality $|\mathcal{S}| = d$, and set $\widetilde{R}' = \frac{1}{\sqrt{m}}R'$. Then, for $\epsilon \in (0,1)$, we will have*

$$\sqrt{1-\epsilon} \;\leq\; \sigma_d\left(\widetilde{R}'_{\mathcal{S}}\right) \;\leq\; \sigma_1\left(\widetilde{R}'_{\mathcal{S}}\right) \;\leq\; \sqrt{1+\epsilon}$$

with probability at least $1-p$ provided that $m \geq (8/3)K^2\epsilon^{-2}d\ln(2d/p)$.

Note that Theorem 2 also applies to the special case where our orthonormal system, Ψ, consists of the N columns of a rescaled unitary matrix $U \in \mathbb{R}^{N\times N}$ (i.e., $\psi_j = \sqrt{N}\mathbf{U}_j$ for all $j \in [N]$). Here, $\mathcal{D} = [N] \subset \mathbb{R}$, $\psi_j(i) = \sqrt{N}U_{i,j}$ for all $i, j \in [N]$, and μ is the discrete uniform measure on $[N]$. In this case we will consider the random sampling matrix, R', for Ψ to be the product $\sqrt{N}RU$, where $R \in \{0,1\}^{m\times N}$ is a random matrix with exactly one nonzero entry per row (which is selected uniformly at random). We will refer to any such random matrix, $R \in \{0,1\}^{m\times N}$, as a *random selector matrix*.

[2] So that $\mathbb{P}\left[\mathbf{t}_j \in \mathcal{S}\right] = \mu\left(\mathcal{S}\right)$ for all measurable $\mathcal{S} \subseteq \mathcal{D}$ and $j \in [m]$.

[3] The specific form of the lower bound used for m below is taken from Theorem 12.12 of [12].

3. Exponential Accuracy for First Order Sigma-Delta via Random Sampling

In this section we will deal only with first order Sigma-Delta. Hence, given $\mathbf{x} \in \mathcal{B}^d$, the vectors $\mathbf{q}, \mathbf{u} \in \mathbb{R}^N$ will always be those resulting from (2) and (3) above. Our objective in this section is to demonstrate that a small set of sums of the bit stream produced by the first order scheme considered herein suffices to accurately encode the vector being quantized. Furthermore, and somewhat surprisingly, the number of sums we must keep in order to successfully approximate the quantized vector is *entirely independent of* N (though the reconstruction error depends on N). Proving this will require the following lemma.

Lemma 1. *For every $\mathbf{x} \in \mathcal{B}^d$ we will have $D^{-1}\mathbf{q} \in \{-N, \ldots, N\}^N \subset \mathbb{Z}^N$.*

Proof: Note that $q_i \in \{-1, 1\}$ for all $i \in [N]$ (see (2)). Furthermore, it is not difficult to check that

$$\left(D^{-1}\right)_{i,j} = \begin{cases} 1 & \text{if } j \leq i \\ 0 & \text{otherwise} \end{cases}.$$

Thus, we have that $\left(D^{-1}\mathbf{q}\right)_i \in \{-i, \ldots, i\}$ for all $i \in [N]$. $\square$

We are new equipped to prove the main theorem of this section.

Theorem 3. *Let $\epsilon, p \in (0, 1)$, and $R \in \{0, 1\}^{m \times N}$ be a random selector matrix. Then, there exists a frame $F \in \mathbb{R}^{N \times d}$ such that*

$$\left\| \mathbf{x} - \left(RD^{-1}F\right)^\dagger RD^{-1}\mathbf{q} \right\|_2 \leq \frac{\sqrt{2}\pi}{\sqrt{1-\epsilon}} \left(\frac{d^{\frac{3}{2}}}{N}\right)$$

for all $\mathbf{x} \in \mathcal{B}^d \subset \mathbb{R}^d$ with probability at least $1 - p$, provided that $m \geq (16/3)\epsilon^{-2} d \ln(2d/p)$. Here, $\mathbf{q}$ is the output of the first order $\Sigma\Delta$ quantization scheme (2) and (3), applied to $F\mathbf{x}$. Furthermore, $RD^{-1}\mathbf{q}$ can always be encoded using $b \leq m(\log_2 N + 1)$ bits.

Proof: Let $U, \Sigma, V \in \mathbb{R}^{N \times N}$ be defined as in (11), (12), and (13), respectively. Define $F \in \mathbb{R}^{N \times d}$ to be the (renormalized) last d columns of U,

$$F := \sqrt{\frac{N}{2d}} \left(\mathbf{U}_{N-d+1} \ldots \mathbf{U}_N\right). \tag{16}$$

We refer to the frame corresponding to F as the 1st order Sobolev self-dual frame. Denoting the i^{th} row of F by $\mathbf{f}_i \in \mathbb{R}^d$, we note that (11) implies that

$$\|\mathbf{y}\|_\infty \;=\; \|F\mathbf{x}\|_\infty \;\leq\; \max_{i \in [N]} \|\mathbf{f}_i\|_2 \|\mathbf{x}\|_2 \;\leq\; \sqrt{\frac{N}{2d}} \sqrt{\frac{2d}{N + 1/2}} \cdot \|\mathbf{x}\|_2 \;\leq\; 1 \tag{17}$$

for all $\mathbf{x} \in \mathcal{B}^d$. Now, apply the random selector matrix, R, to (14) to obtain

$$R\mathbf{u} \;=\; RD^{-1}F\mathbf{x} - RD^{-1}\mathbf{q}.$$

Since our goal is to obtain an upper bound on

$$\|(RD^{-1}F)^\dagger R\mathbf{u}\|_2 \;=\; \|\mathbf{x} - (RD^{-1}F)^\dagger RD^{-1}\mathbf{q}\|_2 \tag{18}$$

and since $\|Ru\|_2$ is easily controlled (see the discussion after (5)), it behooves us to study $RD^{-1}F \in \mathbb{R}^{m\times d}$. Observe that

$$D^{-1}F \;=\; \sqrt{\frac{N}{2d}}V\Sigma^{-1}U^{\mathrm{T}}\left(\mathbf{U}_{N-d+1}\dots\mathbf{U}_N\right) \;=\; \sqrt{\frac{N}{2d}}\left((\mathbf{V})_{N-d+1}\dots(\mathbf{V})_N\right)\;\widetilde{\Sigma}, \tag{19}$$

where $\widetilde{\Sigma}\in\mathbb{R}^{d\times d}$ has

$$\widetilde{\Sigma}_{i,j} \;=\; \frac{\delta_{i,j}}{\sigma_{N-d+j}(D)}.$$

Let $\mathcal{S}=\{N-d+1,\dots,N\}\subset[N]$. Then,

$$RD^{-1}F \;=\; \sqrt{\frac{N}{2d}}RV_{\mathcal{S}}\widetilde{\Sigma} \;=\; \sqrt{\frac{m}{2d}}\left(\frac{\sqrt{N}RV}{\sqrt{m}}\right)_{\mathcal{S}}\widetilde{\Sigma}.$$

Applying Theorem 2 now tells us that

$$\sqrt{\frac{m}{2d}}\cdot\frac{\sqrt{1-\epsilon}}{\sigma_{N-d+1}(D)} \;\le\; \sigma_d\left(RD^{-1}F\right) \;\le\; \sigma_1\left(RD^{-1}F\right) \;\le\; \sqrt{\frac{m}{2d}}\cdot\frac{\sqrt{1+\epsilon}}{\sigma_N(D)} \tag{20}$$

with probability at least $1-p$, provided that $m\ge(16/3)\epsilon^{-2}d\ln(2d/p)$.

Whenever (20) holds we may approximate $\mathbf{x}\in\mathcal{B}^d$ by

$$\hat{\mathbf{x}} := \left(RD^{-1}F\right)^{\dagger}RD^{-1}\mathbf{q},$$

and then use (18) to estimate the approximation error as

$$\|\mathbf{x}-\hat{\mathbf{x}}\|_2 \;=\; \left\|\left(RD^{-1}F\right)^{\dagger}R\mathbf{u}\right\|_2 \;\le\; \sqrt{\frac{2d}{m}}\cdot\frac{\sigma_{N-d+1}(D)}{\sqrt{1-\epsilon}}\,\|R\mathbf{u}\|_2\,.$$

Using (12) and recalling that $|u_i|\le 1$ for all $i\in[N]$ since $|y_i|\le 1$ for all $i\in[N]$ (see (17)), we obtain

$$\|\mathbf{x}-\hat{\mathbf{x}}\|_2 \;\le\; \frac{2\sqrt{2d}}{\sqrt{1-\epsilon}}\cdot\cos\left(\frac{(N-d+1)\pi}{2N+1}\right) \;\le\; \frac{2\sqrt{2d}}{\sqrt{1-\epsilon}}\cdot\left(\frac{\pi}{2}-\frac{(N-d+1)\pi}{2N+1}\right) \;\le\; \frac{\sqrt{2}\pi}{\sqrt{1-\epsilon}}\left(\frac{d^{\frac{3}{2}}}{N}\right).$$

Finally, Lemma 1 tells us that $RD^{-1}\mathbf{q}$ can always be encoded using $b\le m(\log_2 N+1)$ bits. □

Remark 1. *Theorem 3 provides the desired exponentially decaying rate-distortion bounds. In particular, by choosing the smallest integer $m\ge(16/3)\epsilon^{-2}d\ln(2d/p)$, the rate is*

$$\mathcal{R} = m(\log_2 N+1),$$

and the distortion is

$$\mathcal{D} = \frac{\sqrt{2}\pi d^{3/2}}{N\sqrt{1-\epsilon}}.$$

Expressing the distortion in terms of the rate, we obtain

$$\mathcal{D}(\mathcal{R}) = \frac{2\sqrt{2}\pi d^{3/2}}{\sqrt{1-\epsilon}}\cdot 2^{-\mathcal{R}/m} = C_1(\epsilon)\cdot d^{3/2}\exp\left(-\frac{\mathcal{R}}{C_2(\epsilon)d\ln(2d/p)}\right). \tag{21}$$

Above, $C_1(\epsilon)=2\pi\cdot\sqrt{\frac{2}{1-\epsilon}}$ and $\frac{16}{3\ln 2\cdot\epsilon^2}+\frac{1}{d\ln 2\cdot\ln(2d/p)}\ge C_2(\epsilon)\ge\frac{16}{3\ln 2\cdot\epsilon^2}$.

Theorem 3 demonstrates that only $O(\log N)$-bits must be saved and/or transmitted in order to achieve $O(1/N)$-accuracy (neglecting other dependencies). Moreover, from a practical point of view the use of random sampling matrices is appealing as they allow for the $O(\log N)$ bits to be computed "on the fly" and with minimal cost. However, the first order scheme we consider herein has deficiencies that merit additional consideration. Primarily, the first order scheme (2) and (3) must still be executed before its output bitstream can be compressed. Hence, utilizing the compressed sigma delta encoding described in this section has an $O(N)$ "cost" associated with it (as the N-dimensional vector $F\mathbf{x}$ must be acquired and quantized). This cost may be prohibitive for some applications. Compared to the first order scheme we considered here, higher order schemes can allow the same reconstruction accuracy with a smaller number of measurements. Hence, we will consider results for higher order schemes (and more general frames) in the next section.

4. Exponential Accuracy For General Frames and Orders via Bernoulli Random Matrices

In this section we will deal with a more general class of stable r^{th}-order Sigma-Delta schemes. Hence, given $\mathbf{x} \in \mathcal{B}^d$, the vectors $\mathbf{q}, \mathbf{u} \in \mathbb{R}^N$ will always be those resulting from (6) and (7) above. The main result of this section will require the following lemma, which is essentially proven in [3].

Lemma 2. *Let $\epsilon, p \in (0,1)$, $\{\mathbf{v}_1, \ldots, \mathbf{v}_d\} \subset \mathbb{R}^N$, and $B \in \{-1,1\}^{m\times N}$ be a Bernoulli random matrix. Set $\widetilde{B} = \frac{1}{\sqrt{m}}B$. Then,*

$$(1-\epsilon)\| \mathbf{x} \|_2 \leq \| \widetilde{B}\mathbf{x} \|_2 \leq (1+\epsilon)\| \mathbf{x} \|_2$$

for all $\mathbf{x} \in span\{\mathbf{v}_1, \ldots, \mathbf{v}_d\}$ with probability at least $1-p$, provided that

$$m \geq \frac{4d\ln(12/\epsilon) + 2\ln(1/p)}{\epsilon^2/8 - \epsilon^3/24}.$$

Proof: Combine Theorem 1 with the proof of Lemma 5.1 and the subsequent discussion in [3]. □

In addition to considering more general r^{th}-order quantization schemes, we will also consider a more general class of frames, $F \in \mathbb{R}^{N\times d}$. More specifically, we will allow any frame matrix which adheres to the following definition.

Definition 3. *We will call a frame matrix $F \in \mathbb{R}^{N\times d}$ an (r, C, α)-frame if*

(1) $\|F\mathbf{x}\|_\infty \leq 1$ *for all* $\mathbf{x} \in \mathcal{B}^d$, *and*
(2) $\sigma_d\left(D^{-r}F\right) \geq C \cdot N^\alpha$.

Roughly speaking, the first condition of Definition 3 ensures that the frame F is uniformly bounded, while the second condition can be interpreted as a type of smoothness requirement. We are now properly equipped to prove the main theorem of this section.

Theorem 4. *Let $\epsilon, p \in (0,1)$, $B \in \{-1,1\}^{m\times N}$ be a Bernoulli random matrix, and $F \in \mathbb{R}^{N\times d}$ be an (r, C, α)-frame with $r \in \mathbb{N}$, $\alpha \in (1,\infty)$, and $C \in \mathbb{R}^+$. Consider $\mathbf{q}$, the quantization of*

$F\mathbf{x}$ via a stable r^{th}-order scheme with alphabet $\mathcal{A}_A^{2\mu}$ and stability constant $C_{\rho,Q}(r) \in \mathbb{R}^+$ (see (6), (7), (8) and the subsequent discussion). Then, the following are true.

(i) *The reconstruction error (i.e., the distortion) satisfies*

$$\left\| \mathbf{x} - \left(BD^{-r}F\right)^{\dagger} BD^{-r}\mathbf{q} \right\|_2 \;\leq\; \frac{C_{\rho,Q}(r) \cdot N^{1-\alpha}}{C \cdot (1-\epsilon)}$$

for all $\mathbf{x} \in \mathcal{B}^d \subset \mathbb{R}^d$ with probability at least $1-p$, provided that $m \geq \frac{4d\ln(12/\epsilon)+2\ln(1/p)}{\epsilon^2/8-\epsilon^3/24}$.

(ii) *$BD^{-r}\mathbf{q}$ can always be encoded using $b \leq m[(r+1)\log_2 N + \log_2 A + 1]$ bits.*

Proof: Apply a Bernoulli random matrix, $B \in \{-1,1\}^{m\times N}$, to (15) and then renormalize by $m^{-1/2}$ to obtain

$$\widetilde{B}\mathbf{u} \;=\; \widetilde{B}D^{-r}F\mathbf{x} - \widetilde{B}D^{-r}\mathbf{q}, \tag{22}$$

where $\widetilde{B} = \frac{1}{\sqrt{m}}B$. Considering $\widetilde{B}D^{-r}F \in \mathbb{R}^{m\times d}$, we note that Lemma 2 guarantees that $\widetilde{B}$ is an near-isometry on $\text{span}\{D^{-r}\mathbf{F}_1, \ldots, D^{-r}\mathbf{F}_d\}$. Thus,

$$(1-\epsilon)\cdot C \cdot N^{\alpha} \;\leq\; \sigma_d\left(\widetilde{B}D^{-r}F\right) \tag{23}$$

with probability at least $1-p$, provided that $m \geq \frac{4d\ln(12/\epsilon)+2\ln(1/p)}{\epsilon^2/8-\epsilon^3/24}$.

Given that (23) holds, we may approximate $\mathbf{x} \in \mathcal{B}^d$ using $BD^{-r}\mathbf{q} \in \mu\mathbb{Z}^m$ by

$$\hat{\mathbf{x}} := \frac{1}{\sqrt{m}}\left(\widetilde{B}D^{-r}F\right)^{\dagger} BD^{-r}\mathbf{q} \;=\; \left(\widetilde{B}D^{-r}F\right)^{\dagger} \widetilde{B}D^{-r}\mathbf{q},$$

and then use (22) to estimate the approximation error as

$$\|\mathbf{x}-\hat{\mathbf{x}}\|_2 \;=\; \left\|\left(\widetilde{B}D^{-r}F\right)^{\dagger}\widetilde{B}\mathbf{u}\right\|_2 \;=\; \left\|\left(BD^{-r}F\right)^{\dagger}B\mathbf{u}\right\|_2 \;\leq\; \frac{N^{-\alpha}}{C\cdot(1-\epsilon)}\left\|\widetilde{B}\mathbf{u}\right\|_2.$$

Noting that $\|\mathbf{u}\|_\infty \leq C_{\rho,Q}(r)$ since $\|F\mathbf{x}\|_\infty \leq 1$ (by definition of (r,C,α)-frames), we obtain

$$\|\mathbf{x}-\hat{\mathbf{x}}\|_2 \;\leq\; \frac{N^{-\alpha}}{C\cdot(1-\epsilon)}\left\|\widetilde{B}\mathbf{u}\right\|_2 \;\leq\; \frac{C_{\rho,Q}(r)\cdot N^{1-\alpha}}{C\cdot(1-\epsilon)}.$$

Finally, a short argument along the lines of Lemma 1 tells us that $BD^{-r}\mathbf{q} \in \mu\mathbb{Z}^m$ will always have $\|BD^{-r}\mathbf{q}\|_\infty \leq 2\mu A \cdot N^{r+1}$. Thus, $BD^{-r}\mathbf{q}$ can be encoded using $b \leq m[(r+1)\log_2 N + \log_2 A + 1]$ bits. Note that μ does not influence the number of required bits. □

Remark 2. *Theorem 4 also provides the desired exponentially decaying rate-distortion bounds. In particular, by choosing the smallest integer $m \geq \frac{4d\ln(12/\epsilon)+2\ln(1/p)}{\epsilon^2/8-\epsilon^3/24}$, the rate is*

$$\mathcal{R} = m[(r+1)\log_2 N + \log_2 A + 1],$$

and the distortion is

$$\mathcal{D} = \frac{C_{\rho,Q}(r)\cdot N^{1-\alpha}}{C\cdot(1-\epsilon)}.$$

Expressing the distortion in terms of the rate, we obtain

(24)

$$\mathcal{D}(\mathcal{R}) = \frac{C_{\rho,Q}(r)\cdot(2A)^{(\alpha-1)/(r+1)}}{C\cdot(1-\epsilon)}\cdot 2^{-(\mathcal{R}(\alpha-1)/m(r+1))} \;\leq\; \bar{C}_{\rho,Q}(A,\epsilon,\alpha,r)\cdot\exp\left(-\frac{\mathcal{R}}{d\cdot C_3(\epsilon,p)}\right).$$

Above, $\bar{C}_{\rho,Q}(A,\epsilon,\alpha,r) = \frac{C_{\rho,Q}(r)\cdot(2A)^{(\alpha-1)/(r+1)}}{C\cdot(1-\epsilon)}$ *and* $\frac{4(r+1)\ln(12/\epsilon p)}{\ln 2\cdot(\alpha-1)(\epsilon^2/8-\epsilon^3/24)} + \frac{1}{d} \geq C_3(\epsilon,p) > 0.$

Remark 3. *The choice of Bernoulli matrices in Theorem 4 is motivated by two properties. First, one can encode their action on the integrated* $\Sigma\Delta$ *bit-stream in a lossless manner. Second, Bernoulli matrices (of appropriate size) act as near isometries on* $\text{span}\{D^{-r}\mathbf{F}_1,\ldots,D^{-r}\mathbf{F}_d\}$*. In fact, any encoding matrix drawn from a distribution satisfying the above two properties would work for compressing the* $\Sigma\Delta$ *bit-stream. For example,* [1] *also studied other discrete random matrices (whose entries are* ± 1 *with probability* $1/6$ *each, and* 0 *with probability* $2/3$*) and showed that they serve as Johnson-Lindenstrauss embeddings.*

It is informative to compare the rate-distortion bounds resulting from Theorems 3 and 4 in the case of the first order Sigma-Delta scheme (and frame) considered by Theorem 3. If we define $F \in \mathbb{R}^{N\times d}$ as per (16) we can see, by considering (17) and (19), that it will be a $\big(1, d^{-3/2}(\sqrt{2}\pi)^{-1}, 3/2\big)$-frame. Furthermore, the first order scheme considered by Theorem 3 has $A = 1$ and $C_{\rho,Q}(1) = 1$. Hence, we see that (24) becomes

$$\mathcal{D}(\mathcal{R}) = \frac{2^{3/4}\pi d^{3/2}}{1-\epsilon}\cdot 2^{-\mathcal{R}/4m} \;\leq\; \frac{2^{3/4}\pi d^{3/2}}{1-\epsilon}\cdot 2^{-\mathcal{R}\left(\frac{\epsilon^2-\epsilon^3/3}{128\cdot d\ln(12/\epsilon)+64\ln(1/p)}\right)}$$

in this case. Comparing this expression to (21) we can see that the dependence on d has been improved (i.e., by a log factor) in the denominator of the exponent. However, we have sacrificed some computational simplicity for this improvement since $BD^{-1}\mathbf{q}$ will generally require more effort to compute then $RD^{-1}\mathbf{q}$ in practice.

Importantly, though, Theorem 4 also enables one to obtain near-optimal rate-distortion bounds. Moreover, fixing the desired distortion, fewer samples N may now be used than in Theorem 3 (hence less computation for quantization and encoding) via the use of higher order quantization schemes. As a result, we will be able to use (r, C, α)-frames having only $O(N^{\frac{1}{\alpha-1}})$ rows below while still achieving $O(1/N)$ accuracy (ignoring dependences on other parameters such as d, etc.). This represents a clear improvement over the first order scheme we have considered so far for all $\alpha > 2$, provided such (r, C, α) frames exists. In the next section we will briefly survey some examples of currently known (r, C, α)-frames, for general $r \in \mathbb{N}$, which are suitable for use with the type of stable r^{th}-order sigma delta schemes considered herein.

4.1. **Examples of** (r, C, α)**-frames.** In this section we briefly survey some (r, C, α)-frames that can by utilized in concert with Theorem 4 above.

Example 1. Sobolev self-dual frames [21]

Our first example of a family of (r, C, α)-frames represents a generalization of the frame utilized by Theorem 3 to higher orders. Let $U_{D^r} = (\mathbf{U}_1 ... \mathbf{U}_N)$ be the matrix of left singular vectors of D^r, corresponding to a decreasing arrangement of the singular values. Then, we refer to $F_{(r)} = (\mathbf{U}_{N-d+1} ... \mathbf{U}_N)$ as the (r^{th}-order) Sobolev self-dual frame. $F_{(r)}$ is an (r, C, α)-frame with $C = \pi^{-r}(d + 2r)^{-r}$ and $\alpha = r$ (see [21], Theorem 8).

For these frames, with fixed r, using the $\Sigma\Delta$ schemes of Proposition 1 and a Bernoulli encoding matrix, the exponent in the rate-distortion expression $\mathcal{D}(\mathcal{R})$ behaves like $-\frac{r-1}{r+1}\frac{\mathcal{R}}{d}$. Specifically, considering Example 1 with $r = 1$ we see that (16), when unnormalized, is a $(1, \pi^{-1}(d+2)^{-1}, 1)$-frame instead of a $(1, d^{-3/2}(\sqrt{2}\pi)^{-1}, 3/2)$-frame. Hence, the bound provided for the unscaled matrix by Example 1 with $r = 1$ is weaker than the result for the rescaled matrix, whenever d is significantly smaller than N. What prevents us from rescaling F when $r \geq 2$ is that we have insufficient information regarding the left singular vectors of D^r.

Example 2. Harmonic frames

A *harmonic frame*, $F \in \mathbb{R}^{N\times d}$, is defined via the following related functions:

$$F_0(t) \;=\; \frac{1}{\sqrt{2}}, \tag{25}$$

$$F_{2j-1}(t) \;=\; \cos(2\pi jt),\ j \geq 1, \text{ and} \tag{26}$$

$$F_{2j}(t) \;=\; \sin(2\pi jt),\ j \geq 1. \tag{27}$$

We then define $F_{j,k} = \sqrt{\frac{2}{d}} \cdot F_{j'}(k/N)$, where $j' = j - d \bmod 2$ for all $k \in [N]$ and $j \in [d + d \bmod 2]$. In addition to Sobolev self-dual frames, we note that harmonic frames also yield general (r, C, α)-frames. For sufficiently large N, a harmonic frame is an (r, C, α)-frame with $C = C_1 e^{r/2} r^{-(r+C_2)}$ and $\alpha = r + 1/2$ (see [21], Lemma 17). Here, C_1 and C_2 are constants that possibly depend on d. For this example, with fixed r, using the $\Sigma\Delta$ schemes of Proposition 1 and a Bernoulli encoding matrix, the exponent in the rate-distortion expression $\mathcal{D}(\mathcal{R})$ behaves like $-\frac{r-1/2}{r+1}\frac{\mathcal{R}}{d}$.

Example 3. Frames generated from piecewise-C^1 uniformly sampled frame paths

Note that the example above is a special case of a *smooth frame* [5]. As one might expect, more general classes of smooth frames also yield (r, C, α)-frames. One such class of frames consists of those generated from piecewise-C^1 uniformly sampled frame paths, as defined in [5]. For convenience, we reproduce the definition below.

Definition 4. *A vector valued function $E : [0, 1] \mapsto \mathbb{R}^d$ given by $E(t) = [\mathbf{E}_1(t), \mathbf{E}_2(t), ..., \mathbf{E}_d(t)]$ is a piecewise-C^1 uniformly sampled frame path if*

1. *for all $n \in [d]$, $\mathbf{E}_n : [0, 1] \mapsto \mathbb{R}$ is piecewise-C^1,*
2. *the functions $\mathbf{E}_n$ are linearly independent, and*
3. *there exists an N_0 such that for all $N \geq N_0$, the matrix F with entries $F_{ij} = \mathbf{E}_j(i/N)$ is a frame matrix.*

In this case, we say that the frame F is generated from a piecewise-C^1 uniformly sampled frame path.

For any piecewise-C^1 uniformly sampled frame path, there is an $N_0 \in \mathbb{N}$ such that for all $N > N_0$, any frame generated from the frame path is an (r, C, α)-frame for some C (possibly depending on r and d) and $\alpha = r + 1/2$ (see [5], Theorem 5.4 and its proof). Here, again, with fixed r, using the $\Sigma\Delta$ schemes of Proposition 1 and a Bernoulli encoding matrix, the exponent in the rate-distortion expression $\mathcal{D}(\mathcal{R})$ behaves like $-\frac{r-1/2}{r+1}\frac{\mathcal{R}}{d}$. Example 3 deals

with smooth frames of a fairly general type, albeit at the cost of less precision in specifying C. Perhaps more surprisingly, decidedly non-smooth frames also yield (r, C, α)-frames in general. In particular, we may utilize Bernoulli random matrices as both our bit stream compression operator, *and* our (r, C, α)-frame.

Example 4. Bernoulli and Sub-Gaussian frames

Let $\gamma \in [0, 1]$. Then, there exists constant c_1 and c_2, such that with probability exceeding $1 - 2e^{-c_1 N^{1-\gamma} d^\gamma}$, a frame F whose entries are $\pm\frac{1}{\sqrt{d}}$ Bernoulli random variables is an (r, C, α)-frame, provided $N \geq (c_2 r)^{\frac{1}{1-\gamma}} d$. Here $C = d^{-\gamma(r-1/2)-1/2}$ and $\alpha = 1/2 + \gamma(r - 1/2)$. See [22, Proposition 4.1] for a proof.

In fact, Bernoulli frames are a special case of a more general class of frames whose entries are sub-Gaussian random variables. These more general types of random matrices also serve as (r, C, α)-frames.

Definition 5. *If two random variables η and ξ 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.*

Definition 6. *We say that a matrix is sub-Gaussian with parameter c, mean μ, and variance σ^2 if its entries are independent and e-dominated by a Gaussian random variable with parameter c, mean μ, variance σ^2.*

Let $\gamma \in [0, 1]$. Then, there exists a constant $c_1 > 0$ such that, with probability exceeding $1 - 3e^{-c_1 N^{1-\gamma} d^\gamma}$, a random sub-Gaussian frame matrix F with mean zero, variance $1/N$, and parameter c will be a (r, C, α)-frame whenever $\frac{N}{d} \geq (c_2 r)^{\frac{1}{1-\gamma}}$ where c_2 depends only on c. Here $C = d^{-\gamma(r-1/2)}$ and $\alpha = \gamma(r - 1/2)$. See [22, Propositions 4.1 and 4.2] for a proof. Consequently, using the $\Sigma\Delta$ schemes of Proposition 1 together with a Bernoulli encoding matrix and a Sub-Gaussian frame results in the exponent of the rate-distortion expression, $\mathcal{D}(\mathcal{R})$, behaving like $-\frac{\gamma \cdot r - \frac{1}{2}(\gamma+2)}{r+1} \frac{\mathcal{R}}{d}$.

5. Numerical experiments

In this section we present numerical experiments to illustrate our results. To illustrate the results of Section 3, we first generate 5000 points uniformly from $\mathcal{B}^d$, with $d = 2, 6,$ and 10. We then compute, for various N, the 1-bit 1st order greedy $\Sigma\Delta$-quantization of $F\mathbf{x}$, where F is an $N \times d$ Sobolev self-dual frame. $RD^{-1}\mathbf{q}$, where R is an $m \times N$ random selector matrix with $m = 10d$ is then employed to recover an estimate $\hat{\mathbf{x}} = (RD^{-1}F)^\dagger RD^{-1}\mathbf{q}$ of $\mathbf{x}$. In Figure 1 we plot (in log scale) the maximum and mean of $\|\hat{\mathbf{x}} - \mathbf{x}\|_2$ over the 5000 realizations of x versus the induced bit-rate.

Our second experiment is similar, albeit we now use a third order 1-bit $\Sigma\Delta$ quantizer according to the schemes of [9] to quantize the harmonic frame expansion of vectors in $\mathcal{B}^d$, with $d = 2, 6,$ and 10. Here, we use a $d \times m$ Bernoulli matrix, with $m = 5d$ to encode $BD^{-1}\mathbf{q}$ and subsequently obtain $\hat{\mathbf{x}} = (BD^{-1}F)^\dagger BD^{-1}\mathbf{q}$. As before, we plot the maximum and mean of $\|\hat{\mathbf{x}} - \mathbf{x}\|_2$ over the 5000 realizations of $\mathbf{x}$ versus the induced bit-rate.

For our third experiment, we fix $d = 20$ and use the $\Sigma\Delta$ schemes of [9] with $r = 1, 2,$ and 3 to quantize the Bernoulli frame coefficients, and we use Bernoulli matrices with $m = 5d$ to encode. In Figure 3 we show the maximum error versus the bit-rate. Note the different slopes corresponding to $r = 1, 2,$ and 3. This observation is in agreement with the prediction

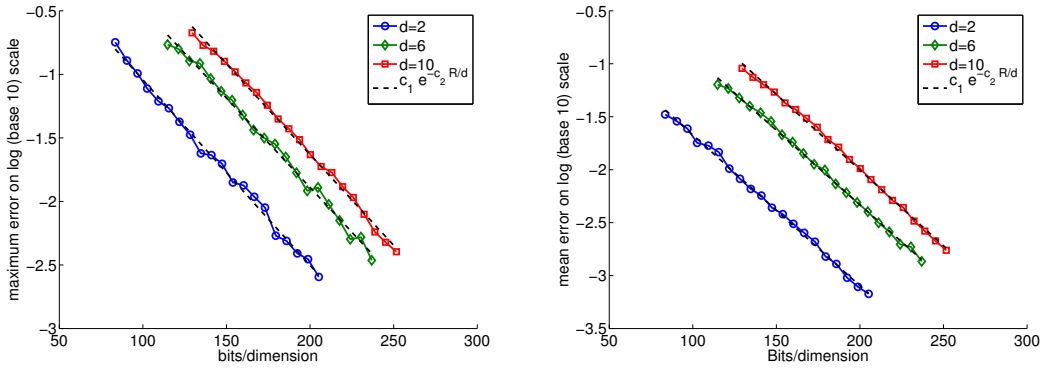

FIGURE 1. (left) The maximum and (right) mean ℓ_2-norm error (in $\log_{10}$ scale) plotted against the number of bits per dimension (b/d). Here we use a 1st order greedy $\Sigma\Delta$ scheme to quantize and a random selector matrix to encode.

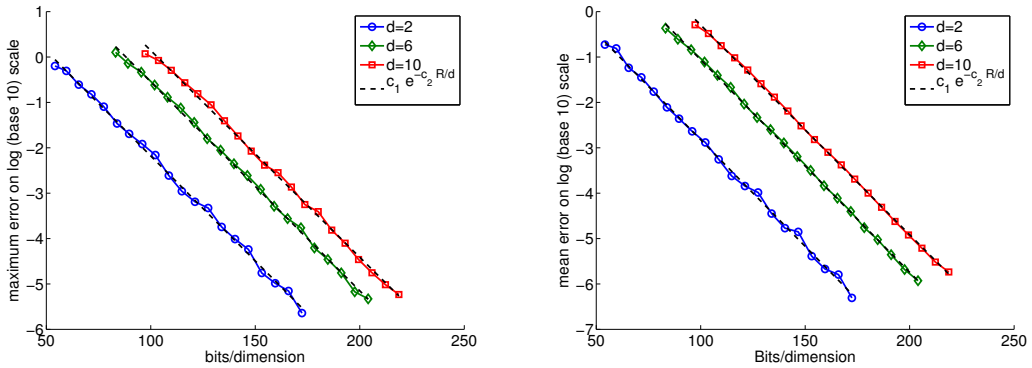

FIGURE 2. (left) The maximum and (right) mean ℓ_2-norm error (in $\log_{10}$ scale) plotted against the number of bits per dimension (b/d). Here we use a third order $\Sigma\Delta$ scheme to quantize and a Bernoulli matrix to encode.

(see the discussion around Example 4) that the exponent in the rate-distortion expression $\mathcal{D}(\mathcal{R})$ is a function of r.

REFERENCES

[1] D. Achlioptas. Database-friendly random projections. In *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 274–281. ACM, 2001.

[2] U. Ayaz. Sigma-delta quantization and sturmian words. Master's thesis, University of British Columbia, 2009.

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

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

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

[6] D. J. Brady. Multiplex sensors and the constant radiance theorem. *Optics Letters*, 27(1):16–18, 2002.

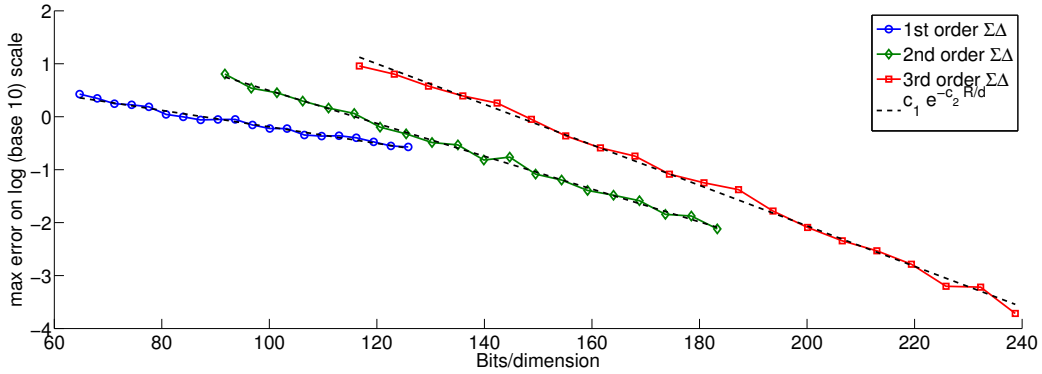

FIGURE 3. The maximum ℓ_2-norm error (in $\log_{10}$ scale) plotted against the number of bits per dimension (b/d). Here $d = 20$ and $\Sigma\Delta$ schemes with $r = 1, 2$ and 3 are used to quantize the frame coefficients. A Bernoulli matrix is used for encoding.

[7] S. Dasgupta and A. Gupta. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65, 2003.

[8] 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.*, 158(2):679–710, 2003.

[9] P. Deift, F. Krahmer, and C. Güntürk. An optimal family of exponentially accurate one-bit sigma-delta quantization schemes. *Communications on Pure and Applied Mathematics*, 64(7):883–919, 2011.

[10] M. F. Duarte, M. A. Davenport, D. Takhar, J. N. Laska, T. Sun, K. F. Kelly, and R. G. Baraniuk. Single-pixel imaging via compressive sampling. *Signal Processing Magazine, IEEE*, 25(2):83–91, 2008.

[11] M. Fickus, M. L. Massar, and D. G. Mixon. Finite frames and filter banks. In *Finite Frames*, pages 337–379. Springer, 2013.

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

[13] P. Frankl and H. Maehara. The Johnson-Lindenstrauss lemma and the sphericity of some graphs. *Journal of Combinatorial Theory, Series B*, 44(3):355–362, 1988.

[14] C. Güntürk. One-bit sigma-delta quantization with exponential accuracy. *Communications on Pure and Applied Mathematics*, 56(11):1608–1630, 2003.

[15] C. Güntürk, J. Lagarias, and V. Vaishampayan. On the robustness of single-loop sigma-delta modulation. *Information Theory, IEEE Transactions on*, 47(5):1735–1744, 2001.

[16] C. Güntürk, M. Lammers, A. Powell, R. Saab, and Ö. Yılmaz. Sobolev duals for random frames and sigma-delta quantization of compressed sensing measurements. *Found. Comput. Math.*, 13:1–36, 2013.

[17] S. Hein, K. Ibraham, and A. Zakhor. New properties of sigma-delta modulators with dc inputs. *Communications, IEEE Transactions on*, 40(8):1375–1387, 1992.

[18] H. Inose and Y. Yasuda. A unity bit coding method by negative feedback. *Proceedings of the IEEE*, 51(11):1524–1535, 1963.

[19] W. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary mathematics*, 26:189–206, 1984.

[20] T. P. Kohman. Coded-aperture x-or γ-ray telescope with least-squares image reconstruction. i. design considerations. *Review of scientific instruments*, 60(11):3396–3409, 1989.

[21] F. Krahmer, R. Saab, and R. Ward. Root-exponential accuracy for coarse quantization of finite frame expansions. *Information Theory, IEEE Transactions on*, 58(2):1069–1079, 2012.

[22] F. Krahmer, R. Saab, and Ö. Yılmaz. Sigma-delta quantization of sub-Gaussian frame expansions and its application to compressed sensing. *Preprint, arXiv:1306.4549*, 2013.

[23] F. Krahmer and R. Ward. New and improved Johnson-Lindenstrauss embeddings via the restricted isometry property. *SIAM Journal on Mathematical Analysis*, 43(3):1269–1281, 2011.

[24] G. Lorentz, M. von Golitschek, and Y. Makovoz. *Constructive approximation: advanced problems*. Grundlehren der mathematischen Wissenschaften. Springer, 1996.

[25] S. Norsworthy, R. Schreier, G. Temes, et al. *Delta-sigma data converters: theory, design, and simulation*, volume 97. IEEE press New York, 1997.

[26] A. Powell, R. Saab, and Ö. Yılmaz. Quantization and finite frames. In P. Casazza and G. Kutinyok, editors, *Finite Frames: Theory and Aplications*, pages 305–328. Birkhauser, 2012.

[27] J. Von Neumann. Distribution of the ratio of the mean square successive difference to the variance. *The Annals of Mathematical Statistics*, 12(4):367–395, 1941.

Mark Iwen
Department of Mathematics, Michigan State University
Department of Electrical and Computer Engineering, Michigan State University
E-mail address: iwenmark@msu.edu

Rayan Saab
Department of Mathematics, University of California, San Diego
E-mail address: rsaab@ucsd.edu