Root-Exponential Accuracy for Coarse Quantization of Finite Frame Expansions

Felix Krahmer, Rayan Saab, and Rachel Ward

Abstract

In this note, we show that by quantizing the N-dimensional frame coefficients of signals in $\mathbb{R}^d$ using r-th-order Sigma-Delta quantization schemes, it is possible to achieve root-exponential accuracy in the oversampling rate $\lambda := N/d$. In particular, we construct a family of finite frames tailored specifically for coarse Sigma-Delta quantization that admit themselves as both canonical duals and Sobolev duals. Our construction allows for error guarantees that behave as $e^{-c\sqrt{\lambda}}$, where under a mild restriction on the oversampling rate, the constants are absolute. Moreover, we show that harmonic frames can be used to achieve the same guarantees, but with the constants now depending on d.

I. Introduction

Signal quantization is a fundamental problem in signal processing. Viewing a signal as a vector in $\mathbb{R}^d$, quantization involves replacing the vector with coefficients that are each chosen from a finite alphabet $\mathcal{A}$. In particular, one can represent a vector x in $\mathbb{R}^d$ by a vector q in $\mathcal{A}^N$, where $N > d$, in the following way. First, one computes a finite-frame expansion $y := Ex$, where E is an appropriately chosen full-rank matrix in $\mathbb{R}^{N\times d}$ (see Section II for a precise definition). Next, one applies a quantization scheme to replace y with q. This approach will be referred to as *frame quantization* in the sequel. More specifically, the quantization schemes we study in this paper are designed to allow for good *linear reconstruction* of x, i.e., we focus on approximation formulas of the form $\widetilde{x} := Fq$ where F is one of the infinitely many left-inverses of E.

Clearly, the goal of a good quantization scheme is to allow for an accurate reconstruction of x from q. Thus, for reasonable frame quantization schemes, one expects that $q \in \mathcal{A}^N$ should allow for increasingly accurate and robust approximation of x as N increases. In the following paragraphs, we will introduce two frame quantization schemes, the second of which, $\Sigma\Delta$ quantization, will be the main focus of this paper.

Felix Krahmer is with the Hausdorff Center for Mathematics, Universität Bonn, Bonn, Germany, email: felix.krahmer@hcm.uni-bonn.de

Rayan Saab is with the Department of Mathematics, Duke University, Durham, NC, email: rayans@math.duke.edu

Rachel Ward is with the Department of Mathematics, University of Texas at Austin, Austin, TX, USA, email: rward@ma.utexas.edu

This paper was presented in part at the 9th International Conference on Sampling Theory and Applications (SAMPTA2011)

A. Memoryless scalar quantization

In the context of quantization using finite-frame representations, the most intuitive approach is *memoryless scalar quantization* (MSQ), which requires replacing each coefficient of $y = Ex$ with its nearest element from $\mathcal{A}$. That is, y is replaced by $\widetilde{q}$, where $\widetilde{q}_i = \arg\min_{v \in \mathcal{A}} |y_i - v|$. On the other hand, this naive approach treats each of the coefficients of y independently, and does not exploit the correlations between coefficients of y resulting from the lower-dimensional representation $y = Ex$. Goyal et al. [1] show that, even when using an optimal reconstruction scheme to approximate x from its MSQ quantized frame coefficients, the expected value of the error cannot be better than $\mathcal{O}(\lambda^{-1})$. Here, the expectation is with respect to some probability measure on x that is, for example, absolutely continuous. One can do much better with other quantization schemes. In particular, Sigma-Delta ($\Sigma\Delta$) quantization schemes are more complex, but can achieve better error rates than MSQ by exploiting the redundancy inherent in y.

B. $\Sigma\Delta$ quantization of oversampled bandlimited functions

$\Sigma\Delta$ schemes were introduced for the quantization of oversampled bandlimited functions [2], and have since been studied extensively. In the setting of bandlimited functions, the oversampling rate λ is the ratio of the actual sampling rate to the Nyquist rate and the signal is reconstructed from the samples via a low-pass filter. Since the time-shifted versions of the low-pass filter as they are used in the reconstruction formula form an infinite dimensional frame, this setup can be seen as analogous to the finite-frame case discussed in this paper. In particular, the oversampling rate in the framework of bandlimited functions corresponds to the oversampling rate for finite frame expansions as above.

Daubechies and Devore [3] showed that if the samples of a bandlimited function are quantized according to a stable r-th-order $\Sigma\Delta$ scheme, the L^∞ approximation error $\|f - \widetilde{f}\|_\infty$ is $\mathcal{O}(\lambda^{-r})$. Subsequently, Güntürk [4] showed that certain 1-bit $\Sigma\Delta$ schemes (that is, $\mathcal{A} = \{-1, 1\}$) can achieve exponential precision, i.e., an L^∞ error decay of order $e^{-C_1\lambda}$, by choosing the order r as a function of λ. Here $C_1 < 1$ is a small constant[1]. This work was improved on by Deift et al. [5], who showed that the above constant can be pushed to $C_1 \approx 0.102$. In order to achieve exponential precision, these works use stable families of r-th-order $\Sigma\Delta$ schemes with approximation errors bounded by $C_2(r)\lambda^{-r}$. For well behaved $C_2(r)$, the optimal choice $r^\#(\lambda)$ achieves exponential precision.

C. $\Sigma\Delta$ quantization of finite frame expansions

The use of $\Sigma\Delta$ quantization in the setting of finite frames was first explored by Benedetto et al. [6]. In contrast to the setting of bandlimited functions where the error is most naturally measured with respect to the L^∞-norm, in the finite-dimensional setting it is more amenable to measure error with respect to the Euclidean, i.e. $\ell_2(\mathbb{R}^d)$, metric. In [6], it was shown that with linear reconstruction, even first-order $\Sigma\Delta$ schemes outperform MSQ when

[1] Subsequently C_i will denote a constant, indexed by order of appearance.

the frames are sufficiently redundant and chosen from appropriate families. Subsequent work showed that it is possible to achieve error bounds with respect to the Euclidean metric that decay like $\mathcal{O}(\lambda^{-r})$. For example, in [7], Bodmann et al. proved that with tight frames of special design, r-th-order schemes achieve an error decay rate of $\mathcal{O}(\lambda^{-r})$, when the left-inverse of the matrix E used in linear reconstruction is the Moore-Penrose inverse. Using a different approach, Blum et al. [8] showed that such an error rate can be achieved by using alternative left-inverses, called *Sobolev duals*, for any frame that arises via uniform sampling from piecewise smooth frame-paths. Recently, Güntürk et al. [9] showed that for randomly-generated frames, error bounds of $\mathcal{O}(\lambda^{-(r-1/2)\alpha})$, for $\alpha \in (0,1)$, are attainable via the use of Sobolev duals. In particular, the parameter α controls the probability (on the draw of the frame) with which the result holds. This allowed [9] to apply $\Sigma\Delta$ quantization in the context of compressed sensing [10], [11].

In this note, we combine the techniques of Blum et al. [8] and Güntürk [4]/Deift et al. [5] to show that it is possible to achieve root-exponential accuracy in the finite frame setting. In particular, we show that for a family of tight frames of special design that admit themselves as Sobolev duals, and for harmonic frames, root-exponential error rates of $\mathcal{O}(e^{-C\sqrt{\lambda}})$ are achievable.

Remark 1. In [7], Bodmann et al. study r-th order $\Sigma\Delta$ schemes that employ scalar quantizers operating on $[-L, L]$. Their schemes require the input sequence, i.e., the frame expansion of x, to be bounded by $L - (2^r - 1)\delta/2$ where δ is the quantization step size. Consequently, there is an upper bound on admissible values of r for these schemes to work and this does not allow one to optimize the value of r freely as a function of λ. A similar issue arises in [9], where the frames are random. On the other hand, in the bandlimited setting, [4] and [5] proposed $\Sigma\Delta$ schemes that do not suffer from an r-dependent constraint on the input sequence. However, the involved constants grow in r. By freely optimizing r as a function of λ, [4] and [5] one can balance these effects obtaining exponential precision in λ (measured in the L^∞ norm). In this paper, we use the $\Sigma\Delta$ schemes of [4] and [5] for frame quantization and Sobolev duals as in [8] for linear reconstruction. Consequently, we can freely optimize r as a function of λ. This allows us to obtain root-exponential precision in the $\ell_2(\mathbb{R}^d)$ norm.

D. Organization of the paper

The remainder of the paper is organized as follows. In Section II, we introduce the relevant basic concepts from frame theory and we describe $\Sigma\Delta$ quantization. In Section III, we construct a family of frames that admit themselves as both canonical and Sobolev duals and we show that they allow root exponential approximation errors. We derive explicit bounds on the constants; in particular, we show that the error is bounded by $C_3 e^{-C_4\sqrt{\lambda}}$, except for very small oversampling rates $\lambda := \frac{N}{d} \lesssim (\log d)^2$, where the constants C_3 and C_4 do not depend on the dimension d. In Section IV, we study the performance of harmonic frames, showing that they too allow root-exponential bounds on the reconstruction error, albeit without the explicit analysis of the dimension dependence of the error. Finally, in Section V, we include the results of numerical experiments showing that the effective decay rate of the error as a function of λ, when using the proposed schemes, is indeed root-exponential. This highlights the fact that our mathematical analysis (for the proposed frames and reconstruction method) is not sub-optimal but matches the empirically observed error decay.

II. Preliminaries

A. Finite frames

We say that a finite collection of vectors $\{e_n\}_{n=1}^N$ is a frame for $\mathbb{R}^d$ with frame bounds $0 < A \leq B < \infty$ if

$$\forall x \in \mathbb{R}^d, \qquad A\|x\|_2^2 \leq \sum_{n=1}^{N} |\langle x, e_n\rangle|^2 \leq B\|x\|_2^2, \tag{1}$$

where $\|\cdot\|_2$ denotes the Euclidean norm, and A and B are the largest and smallest numbers such that (1) holds, respectively. If $A = B$ we say that the frame is tight. If $\|e_n\|_2 = 1$ for each $n \in \{1, ..., N\}$, then we say that the frame is unit-norm. Given the frame vectors $\{e_n\}_{n=1}^N$, for convenience, we define the frame matrix $E \in \mathbb{R}^{N\times d}$ with e_k as its k-th row. A matrix $E \in \mathbb{R}^{N\times d}$ is thus a frame matrix if and only if it has rank d. Let x be a vector in $\mathbb{R}^d$. Then we say that $y = Ex$ is the frame expansion of x with respect to E. Equivalently we say that $y_n, n \in \{1, ..., N\}$, are the frame coefficients of x.

Consider a frame $\{f_n\}_{n=1}^N$ and let F be the matrix whose k-th column is f_k. F is called a *dual (or synthesis) frame* associated with $\{e_n\}_{n=1}^N$ if the frame matrix F in $\mathbb{R}^{d\times N}$ satisfies $FE = I_d$, where $I_d \in \mathbb{R}^{d\times d}$ is the identity matrix. In other words, a dual frame matrix F is a left inverse of E. As $N > d$, there are infinitely-many such left-inverses. In particular, the *canonical dual frame* (the Moore-Penrose inverse) of E is given by

$$E^\dagger := (E^*E)^{-1}E^*.$$

B. Sigma-Delta Quantization of Finite Frame Expansions

A *midrise quantization alphabet* is a set of the type

$$\mathcal{A} = \mathcal{A}_K^\delta = \{\pm(m - 1/2)\delta : 1 \leq m \leq K, m \in \mathbb{Z}\}.$$

For such an alphabet, we define the associated *scalar quantizer*

$$Q : \mathbb{R} \to \mathcal{A},\ Q(h) := \arg\min_{q\in\mathcal{A}} |h - q|.$$

A *quantization scheme* is a procedure that employs such a quantizer to represent multi- or even infinite-dimensional signals by a sequence of symbols from the alphabet $\mathcal{A}$. In the context of redundant representations, MSQ is the most basic form of quantization; here, x in $\mathbb{R}^d$ is encoded by quantizing the entries of its frame expansion $y = Ex$ independently to obtain a vector q of quantized coefficients, i.e. $q_n = Q(y_n)$. Subsequently, decoding is achieved by using a dual frame F to obtain the approximation $\widetilde{x} = Fq$. However, as mentioned previously, MSQ is suboptimal since it makes no use of the fact that the frame E maps $\mathbb{R}^d$ to a d-dimensional subspace of $\mathbb{R}^N$, spanned by the columns of E. On the other hand, $\Sigma\Delta$ schemes, a class of recursive algorithms first applied to the setting of finite frame expansions in [6], explicitly make use of the dependencies in the vectors of the reconstruction frame F to achieve robust, high precision quantization (see, e.g., [8]). Adopting the notation generally more common in the framework of bandlimited functions ([4],[5]), a general r-th-order $\Sigma\Delta$ scheme with alphabet $\mathcal{A}$ runs the following iteration for $n = 1, 2, \ldots, N$,

$$\begin{aligned} q_n &= Q\left(\rho(u_{n-1}, u_{n-2}, \cdots, u_{n-r}, y_n)\right) \\ \left(\Delta^r u\right)_n &= y_n - q_n. \end{aligned} \tag{2}$$

Here the operator Δ^r results from r subsequent concatenations of the finite difference operator $(\Delta w)_n = w_n - w_{n-1}$, $\rho : \mathbb{R}^{r+1} \mapsto \mathbb{R}$ is a fixed function known as the quantization rule, and Q is the scalar quantizer associated with $\mathcal{A}$ as above. We refer to the sequence u_n as the *state sequence*. In vector form, (2) can be restated as

$$D^r u = y - q, \tag{3}$$

where D is the first-order $N \times N$ difference matrix defined by

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

In this formulation, the iterative nature of (2) is reflected in the invertibility of D. Suppose that $F \in \mathbb{R}^{d\times N}$ is the dual frame to E used for linear reconstruction, and suppose that $\widetilde{x} = Fq$ is the reconstructed approximation to x. Using that $FD^r u = F(y-q) = FEx - \widetilde{x} = x - \widetilde{x}$, it was shown in [12] that the linear reconstruction error of a stable r-th-order $\Sigma\Delta$ scheme with state variables u can be bounded by

$$\|x - \widetilde{x}\|_2 = \|FD^r u\|_2 \leq \|FD^r\|_{2\to 2}\|u\|_2 \leq N^{1/2}\|FD^r\|_{2\to 2}\|u\|_\infty. \tag{5}$$

Here $\|\cdot\|_{2\to 2}$ denotes the matrix norm $\|M\|_{2\to 2} := \max\limits_{\|v\|_2=1} \|Mv\|_2$. In absence of further information about the vector u (which is typically the case), a reasonable quantization procedure should yield good bounds for both norm estimates on the right hand side of (5). To control the factor $\|u\|_\infty$, we concentrate on schemes that are *stable*, that is, there exist constants $C_5 > 0$ and $C_6 > 0$ such that for any $N > 0$ and $y \in \mathbb{R}^N$ one has

$$\|y\|_\infty \leq C_5 \implies \|u\|_\infty \leq C_6. \tag{6}$$

The constants C_5 and C_6 may depend on the order r, the quantization rule ρ, and the alphabet $\mathcal{A}$, but they should not depend on N (and hence not on the oversampling rate λ either). Stability is a crucial concept in the theory of $\Sigma\Delta$ quantization both for bandlimited signals (compare [13]) and for frames (see for example [6]). The construction of stable schemes that allow for good bounds on C_6 will be discussed in the next section. Now we can bound $\|y\|_\infty = \max\limits_{n\in\{1,\ldots,N\}} |\langle e_n, x\rangle| \leq \max\limits_{n\in\{1,\ldots,N\}} \|e_n\|_2\|x\|_2$. Thus in order to ensure that $\|y\|_\infty \leq C_5$ uniformly for all x with $\|x\|_2 \leq C_7$, we need that $\max_{n\in\{1,\ldots,N\}} \|e_n\|_2 \leq \frac{C_5}{C_7}$.

For such a frame E we then seek to find a dual frame F such that $\|FD^r\|_{2\to 2}$ is minimized. This is achieved by the *Sobolev dual* introduced in [8]. The r-th-order Sobolev dual frame of a given frame E is given by

$$F_r := (D^{-r}E)^{\dagger}D^{-r}.$$

As desired, F_r is the left-inverse of E that minimizes the norm $\|FD^r\|_{2\to 2}$ over all left inverses F, $FE = I$ (see [8]). Now two approaches are conceivable: On the one hand, one can attempt to design E to yield particularly good bounds for this minimum. We will follow this approach in Section III introducing a class of frames where the canonical dual and the Sobolev dual coincide. On the other hand, one can work with a given frame. We will follow this approach in Section IV, analyzing the bounds for the harmonic frame, as it has been discussed for example in [6].

C. Superpolynomial Sigma-Delta Quantization

Note that the constants C_5 and C_6 in (6) depend on r, so a sharper analysis will require taking these dependencies into account. The first deduction of superpolynomial decay from explicitly r-dependent bounds for the solution of system (2) was provided in [3] in the context of $\Sigma\Delta$ quantization for bandlimited functions. In [3], the core idea is to choose the order r of the $\Sigma\Delta$ modulator adaptively as a function of the oversampling rate and to choose the underlying quantization rule to be a non-linear function that involves a concatenation of sign functions.

In [4], the author derives a framework that allows for stronger error decay rates (exponential in the context of bandlimited functions). The approach is based on an auxiliary sequence v_n that is defined recursively in terms of r of its non-subsequent previous values and an associated linear quantization rule. The optimal error decay in this framework is provided in [5].

More specifically, one formally substitutes $u = g * v$ for a given $g \in \mathbb{R}^{\{0,\dots,m\}}$ for some $m \geq r$ with $g_0 = 1$ and chooses the quantization rule in terms of the new variables to be $\rho(v_n, v_{n-1}, \dots, y_n, y_{n-1}, \dots) = (h * v)_n + y_n$, where $h = \delta^{(0)} - \Delta^r g$ with $\delta^{(0)}$ the Kronecker delta. Then (2) reads as follows.

$$q_n = Q((h * v)_n + y_n) \tag{7}$$

$$v_n = (h * v)_n + y_n - q_n, \tag{8}$$

Note that as $(\Delta^r g)_0 = g_0 = 1$ and hence $h_0 = 0$, this formula describes again how v_n is computed recursively from v_j, $j < n$. Now by definition of the midrise quantization alphabet $\mathcal{A}_K^{\delta}$ and its scalar quantizer Q, one has

$$|v_n| \leq \max\left(\frac{\delta}{2}, \|h\|_1\|(v_j)_{j=1}^{n-1}\|_\infty + \|y\|_\infty - \left(K - \frac{1}{2}\right)\delta\right),$$

which inductively shows that $\|v_n\|_\infty \leq \frac{\delta}{2}$, i.e., stability, for all input sequences y with $\|y\|_\infty \leq \mu$ provided that $\|h\|_1\frac{\delta}{2} + \mu \leq K\delta$. Here $\|\cdot\|_1$ denotes the ℓ_1 norm given by $\|v\|_1 = \sum |h_j|$.

Stability of this auxiliary scheme automatically implies that the scheme in the original variables $u = g * v$ is also stable as long as the quantized bits are computed using the v_n's. One estimates

$$\|u\|_\infty \leq \|g\|_1\|v\|_\infty \leq \frac{\delta}{2}\|g\|_1. \tag{9}$$

These estimates motivate the study of the following optimization problem first posed in [4].

$$\text{Minimize } \|g\|_1 \text{ over all } g \in \ell^1 \text{ subject to } \|h\|_1 = \|\Delta^r g\|_1 - 1 \leq 2K - \frac{2\mu}{\delta}. \tag{10}$$

To make this problem more tractable, the author restricts the problem to minimally sparse h, i.e., with only r non-zero entries (albeit distributed over a longer interval). This idea allows for the construction of admissible pairs (g, h) that yield the bound

$$\|g\|_1 \leq C_8 C_9^r r^r \tag{11}$$

for some constants C_8, C_9 that depend on μ. With the currently best-known constants resulting from the optimized constructions derived in [5], we can summarize these considerations as follows.

Proposition 2. *There exists a universal constant $C_8 > 0$ such that for any midrise quantization alphabet $\mathcal{A} = \mathcal{A}_K^\delta$, for any order $r \in \mathbb{N}$, and for all $\mu < \delta\left(K - \frac{1}{2}\right)$, there exists $g \in \mathbb{R}^m$ for some $m > r$ such that the $\Sigma\Delta$ scheme given in* (7) *is stable for all input signals y with $\|y\|_\infty \leq \mu$ and*

$$\|u\|_\infty \leq C_8 C_9^r r^r \frac{\delta}{2}, \tag{12}$$

where $u = g * v$ as above and $C_9 = \left(\left\lceil \frac{\pi^2}{(\cosh^{-1}\gamma)^2} \right\rceil \frac{e}{\pi}\right)$ with $\gamma := 2K - \frac{2\mu}{\delta}$.

III. Sobolev self-dual frames

In this section we construct a family of frames $\mathcal{F}_{d,N}(r)$ for $\mathbb{R}^d$, parametrized explicitly by an order $r \in \mathbb{Z}, r \geq 1$. In particular, for any d, N, and r, we construct frames that admit themselves as both canonical and Sobolev duals of order r. We show that the optimal choice of frames from this family allows for a root-exponential error decay rate (by linear reconstruction) when used for the redundant $\Sigma\Delta$ quantization of signals in $\mathbb{R}^d$. Constructing such frames for $r = 1$ and $r > 1$ will be the focus of the next two subsections, respectively. To that end we now focus on some useful properties of D, defined in (4).

Recall that for any matrix M in $\mathbb{R}^{m\times n}$ of rank k, there exists a singular value decomposition (SVD) of the form $M = U_M S_M V_M^*$, where $U_M \in \mathbb{R}^{m\times k}$ is a matrix with orthonormal columns, $S_M \in \mathbb{R}^{k\times k}$ is a diagonal matrix with strictly non-negative entries, and $V_M \in \mathbb{R}^{n\times k}$ is a matrix with orthonormal columns. We will use an equivalent form of the above factorization, with $M = \widetilde{U}_M \widetilde{S}_M V_M^*$. Here, $\widetilde{U}_M \in \mathbb{R}^{m\times m}$ is orthonormal, $\widetilde{S}_M \in \mathbb{R}^{m\times k}$ is “diagonal” (that is, it contains a $k \times k$ diagonal submatrix, with the remaining entries being zero), and $V_M \in \mathbb{R}^{n\times k}$ is a matrix with orthonormal columns as before.

In particular, the difference matrix D admits a singular value decomposition $D = U_D S_D {V_D}^*$ where U_D and V_D are orthonormal matrices and S_D is a diagonal matrix given respectively (see [14], [15]) by

$$U_D(k,l) = \sqrt{\frac{2}{N+1/2}} \cos\left(\frac{2(k-1/2)(N-l+1/2)\pi}{2N+1}\right), \tag{13}$$

$$V_D(k,l) = (-1)^{k+1}\sqrt{\frac{2}{N+1/2}} \sin\left(\frac{2kl\pi}{2N+1}\right), \tag{14}$$

$$S_D(k,l) = 2\delta^{(k,l)} \cos\left(\frac{l\pi}{2N+1}\right). \tag{15}$$

Above, $k, l \in \{1, \ldots, N\}$, $\delta^{(k,l)}$ is the Kronecker delta, and $M(k,l)$ indicates the entry on the k-th row and l-th column of M.

We now briefly summarize how Sobolev self-dual frames arise. Let E and F be dual frames, i.e., $FE = I$ and note that in this section we will design both E and F. Recall that in the context of $\Sigma\Delta$ quantization of redundant frame expansions, we aim to control the error associated with linear reconstruction. Since the above error is given by $\|x - FQ^{\Sigma\Delta}(Ex)\|_2 \le \|FD^r\|_{2\to 2}\|u\|_2$ (where $Q^{\Sigma\Delta}$ denotes r-th-order $\Sigma\Delta$ quantization), we seek E and F such that $\|FD^r\|_{2\to 2}$ is nicely bounded. In particular, it is natural to consider only the Sobolev duals, which minimize $\|FD^r\|_{2\to 2}$ over all duals of E. With this choice of F, $\|FD^r\|_{2\to 2} = \frac{1}{\sigma_{min}(D^{-r}E)}$. On the other hand, for stability considerations we seek E so that $\|Ex\|_2$ is bounded, and thus it is reasonable to restrict our attention to tight frames with frame bound 1. With this choice, the expression $\frac{1}{\sigma_{min}(D^{-r}E)}$ is minimized when E consists of the right singular vectors of D^{-r} corresponding to the largest singular values. As a result, the Sobolev dual and the canonical dual of E agree, the frame is Sobolev self-dual. This argument is made precise in Lemma 3 and Theorem 8.

A. First-order Sobolev self-dual frames

We begin with the construction for the case $r = 1$ and some of its useful properties.

Lemma 3. *Suppose that $E \in \mathbb{R}^{N\times d}$ is a frame matrix for $\mathbb{R}^d$, with frame vectors $\{e_n\}_{n=1}^N$ given by*

$$e_n(l) = \sqrt{\frac{2}{N+1/2}} \cos\left(\frac{(n-1/2)(d-l+1/2)\pi}{N+1/2}\right), \qquad l \in \{1, \ldots, d\}. \tag{16}$$

Let F and $E^\dagger$ be the first order Sobolev dual and canonical dual of E, respectively. Then

(i) E is a tight frame with frame bound 1,

(ii) $F = E^\dagger = E^*$,

(iii) $\|FD\|_{2\to 2} = 2\cos\left(\frac{(N-d+1)\pi}{2N+1}\right)$.

Proof: By definition, $E = [u_{N-d+1}|\cdots|u_{N-1}|u_N]$, where u_i are the columns of U_D as above. As U_D is unitary, the columns are orthonormal, which implies (i). Furthermore, $R := U_D^* U_E \in \mathbb{R}^{N\times d}$ is of the form

$$R(i,j) = \delta^{(N-d-i,j)}. \tag{17}$$

In other words, the entries of R are zero except on the diagonal of its lowermost square $d \times d$ submatrix, where they are 1. Thus,

$$E^\dagger = R^* U_D^* = E^*.$$

To finish the proof of (ii), recall that $FD = \left(D^{-1}E\right)^{\dagger}$, which directly gives

$$\begin{aligned} F &= \left(V_D(S_D^{-1}R)\right)^{\dagger} V_D S_D^{-1} U_D^* \\ &= R^* U_D^* = E^*. \end{aligned}$$

To prove (iii) we write FD using the SVDs of F and D to get

$$\begin{aligned} FD &= (R^* U_D^*)(U_D S_D V_D^*) \\ &= (R^* S_D) V_D^*, \end{aligned}$$

which is itself an SVD of FD. Therefore,

$$\|FD\|_{2\to 2} = 2\cos\left(\frac{(N-d+1)\pi}{2N+1}\right).$$

■

B. Higher order self-dual frames

To deal with the case $r > 1$, we examine the properties of D^r. To that end, let $D^r = U_{D^r} S_{D^r} V_{D^r}^*$ with $r \geq 1$, be the singular value decomposition of D^r, and note that D^r is a Toeplitz matrix. In what follows, we will assume that U_{D^r} can been computed (numerically), but we do not provide an explicit expression for its elements. Our technique in generalizing the results of the previous section to the case $r \geq 1$ will be very similar to the approach used in the proof of Lemma 3. The main difference is that rather than compute S_{D^r}, we will approximate it by $(S_D)^r$ using Weyl's inequalities (see, e.g., [16, Thm 4.3.6]) as in [9].

Theorem 4 (Weyl)**.** *Let Σ and Δ be $N \times N$ Hermitian matrices with eigenvalues*
$\lambda_1(\Sigma) \geq \lambda_2(\Sigma) \geq \ldots \geq \lambda_N(\Sigma)$ and $\lambda_1(\Delta) \geq \lambda_2(\Delta) \geq \ldots \geq \lambda_N(\Delta)$.
Let $\lambda_1(X) \geq \lambda_2(X) \geq \ldots \geq \lambda_N(X)$ be the eigenvalues of $X = \Sigma + \Delta$. Then,

1) $\lambda_i(X) \geq \lambda_{i+j}(\Sigma) + \lambda_{N-j}(\Delta) \quad \forall j \in \{0, 1, 2, ..., N-i\}$
2) $\lambda_i(X) \leq \lambda_{i-j}(\Sigma) + \lambda_{j+1}(\Delta) \quad \forall j \in \{0, 1, 2, ..., i-1\}$.

We will apply Weyl's theorem to $\Sigma = (DD^*)^r$, $\Delta = -(DD^*)^r + D^r D^{*r}$ and $X = D^r D^{*r}$. This will yield estimates of the eigenvalues of $D^r D^{*r}$ and hence estimates of the singular values of S_{D^r} in terms of $(S_D)^r$ (the r-th powers of the singular values of D). To that end, we require estimates of the singular values of $D^r D^{*r} - (DD^*)^r$.

Proposition 5. *Let $\Delta \in \mathbb{R}^{N\times N}$ be as above and let $\mathcal{I} = \{(i,j) : (i,j) \in \{1, ..., r\} \times \{1, ..., r\} \cup \{N-r+1, ..., N\} \times \{N-r+1, ..., N\}$. Then $\Delta_{i,j} = 0$ except possibly when $(i,j) \in \mathcal{I}$. We make no claims over the value of $\Delta_{i,j}$ when $(i,j) \in \mathcal{I}$.*

The proof of this proposition follows trivially from explicitly evaluating $\Sigma_{i,j}$ and $X_{i,j}$ on $\mathcal{I}^c = \{(i,j) : i \in \{1, ..., N\}, j \in \{1, ..., N\}, (i,j) \notin \mathcal{I}\}$ and noting that they are equal. The details are omitted. In fact, the middle

$N-2r$ rows of Σ and X form identical matrices. Specifically, the entries in the $(r+1)$-th row comprise the coefficients of the polynomial $(-1)^r(1-z)^{2r}$. The $(r+1+j)$-th rows, $j \in \{0,...,N-2r-1\}$, are formed by shifting the coefficients in the $(r+1)$-th row j times to the right. For a full proof, see [9].

Thus, Δ has at most $2r$ non-zero eigenvalues. We make no assumptions about their signs (the ordering of eigenvalues matters in applying Weyl's inequalities). On the other hand, we are certain that the $N-2r$ middle eigenvalues are zero. Denoting by $\lambda_j(M)$ the j-th largest eigenvalue of a Hermitian matrix $M \in \mathbb{R}^{N\times N}$, we are now ready to prove the following proposition.

Proposition 6. *For $D \in \mathbb{R}^{N\times N}$ as before and with $N > 4r$,*

$$\lambda_{\min(j+2r,N)}(DD^*)^r \leq \lambda_j(D^r D^{*r}) \leq \lambda_{\max(j-2r,1)}(DD^*)^r, \qquad j=1,\dots,N.$$

Proof: Noting that $(DD^*)^r$ and $D^r D^{*r}$ are Hermitian, using Weyl's inequalities we will first bound the middle eigenvalues of $D^r D^{*r}$. Specifically,

$$\lambda_{j+2r}((DD^*)^r) \leq \lambda_j(D^r D^{*r}) \leq \lambda_{j-2r}((DD^*)^r) \quad \forall j \in \{2r+1,...,N-2r\}.$$

This leaves the largest $2r$ and smallest $2r$ eigenvalues. We start with the largest ones noting that $\lambda_{2r}(D^r D^{*r}) \leq ... \leq \lambda_1(D^r D^{*r}) = \|D^r\|^2_{2\to 2}$ by definition. But $\|D^r\|^2_{2\to 2} \leq \|D\|^{2r}_{2\to 2} = \lambda_1((DD^*)^r)$, so we have a bound from above for the largest $2r$ eigenvalues. Now to bound them from below just apply the relevant Weyl inequalities. This yields

$$\lambda_{j+2r}((DD^*)^r) \leq \lambda_j(D^r D^{*r}) \leq \lambda_1((DD^*)^r), \quad \forall j \in \{1,...,2r\}.$$

We now turn to the smallest eigenvalues. To that end, recall that for any invertible matrix $M \in \mathbb{R}^{N\times N}$, $\lambda_j(MM^*) = (\sigma_j(M))^2$, where $\sigma_j(M)$ denotes the j-th largest singular value of M. Moreover, $\sigma_j(M) = (\sigma_{N-j+1}(M^{-1}))^{-1}$.

Now note that $(\sigma_1(D^{-r}))^2 = \|D^{-r}\|^2_{2\to 2} \leq \|D^{-1}\|^{2r}_{2\to 2} = (\sigma_1(D^{-1}))^{2r}$. We can thus conclude that

$$\lambda_{N-2r+1}(D^r D^{*r}) \geq ... \geq \lambda_N(D^r D^{*r}) \geq \lambda_N((DD^*)^r).$$

We have thus bounded all the smallest eigenvalues from below. To obtain upper bounds, we again use Weyl's inequalities. This yields

$$\lambda_N((DD^*)^r) \leq \lambda_j(D^r D^{*r}) \leq \lambda_{j-2r}((DD^*)^r) \quad \forall j \in \{N-2r+1,...,N\}.$$

■

This trivially yields the following bounds on the singular values of D^r, i.e., the diagonal entries of S_{D^r} which we will refer to by $\sigma_j(D^r), j \in \{1,...,N\}$, where $\sigma_1(D^r) \geq \sigma_2(D^r) \geq \cdots \geq \sigma_N(D^r)$.

Proposition 7. *For $D \in \mathbb{R}^{N\times N}$ as before and with $N > 4r$, one has*

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

We can now present a main result of this section.

Theorem 8. *Let $U_{D^r} = [u_1|u_2|\cdots|u_N]$ be the matrix containing the left singular vectors of D^r, corresponding to the decreasing arrangement of the singular values of D^r. Let $E = [u_{N-d+1}|\cdots|u_{N-1}|u_N]$ and denote by F and $E^\dagger$ the r-th-order Sobolev dual and canonical dual of E, respectively. Then*

(i) *E is a tight frame with frame bound 1,*

(ii) *$F = E^\dagger = E^*$,*

(iii) $\|FD^r\|_{2\to 2} \le \left(2\cos\left(\frac{(N-d-2r+1)\pi}{2N+1}\right)\right)^r$.

Proof: The proof of this theorem traces exactly the same steps as the proof of Lemma 3. The only exception is that to obtain (iii) we need to apply the conclusions of Proposition 7 with $j = N-d+1$. The details are omitted. ■

C. Root-exponential accuracy

The main result of this section is the following.

Theorem 9. *For $0 < K \in \mathbb{Z}$ and $0 < \delta \in \mathbb{R}$, let $x \in \mathbb{R}^d$ be such that $\|x\|_2 \le \mu < \delta\left(K - \frac{1}{2}\right)$, and suppose that we wish to quantize a redundant representation of x with oversampling rate $\lambda = N/d$ using the alphabet $\mathcal{A} = \mathcal{A}_K^\delta$. If $\lambda \ge C_{12}(\log d)^2$, then there exists a Sobolev self-dual frame E and an associated Sigma-Delta quantization scheme $Q^{\Sigma\Delta}$, both of order $r^\# = r(\lambda)$, such that*

$$\|x - E^* Q^{\Sigma\Delta}(Ex)\|_2 \le C_3 e^{-C_4\sqrt{\lambda}}.$$

Here, C_3, C_4 and C_{12} are constants independent of d and x.

Proof: Let E and F be as in Theorem 8. A quick calculation shows that

$$\begin{aligned}\|FD^r\|_{2\to 2} &\le \left(2\cos\left(\frac{(N-d-2r+1)\pi}{2N+1}\right)\right)^r \\ &\le \pi^r\left(\frac{d+2r}{N}\right)^r.\end{aligned}$$

Let $\|\cdot\|_{2\to\infty}$ denote the matrix norm defined by $\|M\|_{2\to\infty} := \max\limits_{\|v\|_2=1}\|Mv\|_\infty$. We will now bound $\|E\|_{2\to\infty}$, recalling that E is a restriction of the matrix U of left singular vectors, which is orthonormal. We obtain

$$\|E\|_{2\to\infty} = \max_{n\in\{1,..,N\}}\|e_n\|_2 \le 1,$$

and consequently, for any x with $\|x\|_2 \le \mu$, $\|Ex\|_\infty \le \mu$.

Let us now use the $\Sigma\Delta$ quantization schemes of Proposition 2 (see also (7)). These schemes yield the bound

$\|u\|_\infty \leq \left(C_8 \frac{\delta}{2}\right) C_9^r r^r$ for $\|Ex\|_\infty \leq \mu$ (see (9) and (11)). Letting $C_{10} = 3\pi C_9$ we obtain

$$\begin{aligned} \|FD^r u\|_2 &\leq \|FD^r\|_{2\to 2}\|u\|_2 \leq \|FD^r\|_{2\to 2}\|u\|_\infty \sqrt{N} \\ &\leq C_8 \left(\pi C_9\right)^r \left(\frac{d+2r}{N}\right)^r r^r N^{1/2} \frac{\delta}{2} \\ &\leq C_8 C_{10}^r \max\left(\left(\frac{1}{N}\right)^{r-1/2} r^{2r}, \left(\frac{d}{N}\right)^r r^r N^{1/2}\right) \frac{\delta}{2} \\ &\leq C_8 C_{10}^r \left(\frac{d}{N}\right)^{r-1/2} \max\left(r^{2r}, r^r d^{1/2}\right) \frac{\delta}{2}. \end{aligned} \tag{18}$$

Thus, the family of Sobolev self-dual frames (of arbitrary order) satisfies the above error bounds. Assuming that the first term in the maximum dominates, we optimize over the order r for a given oversampling rate $\lambda = N/d$. Thus, we set

$$r^{\#} = \left\lfloor \arg\min_r \left(C_{10}\right)^r \lambda^{-r} r^{2r} \right\rfloor = \left\lfloor e^{-1} \sqrt{\frac{\lambda}{C_{10}}} \right\rfloor.$$

Substituting $r^{\#}$ in (18), i.e., choosing $E = E_{r^{\#}}$ (the Sobolev self-dual frame of order $r^{\#}$) yields the error bound (see, e.g., [4])

$$\|x - EQ^{\Sigma\Delta}(Ex)\|_2 = \|F_{r^{\#}} D^{r^{\#}} u\|_2 \leq \left(C_8 e^2 \frac{\delta}{2}\right) \sqrt{\lambda} e^{-C_{11}\sqrt{\lambda}} \leq C_3 \exp(-C_4 \sqrt{\lambda}) \tag{19}$$

where $C_{11} = \frac{2}{\sqrt{C_{10}}e}$, $C_3 = \frac{\delta}{2}\frac{C_8}{C_{11}} e^2$, and $C_4 = C_{11}/2$.

The above bound holds provided that $(r^{\#})^{2r^{\#}} \geq (r^{\#})^{r^{\#}} d^{1/2}$, i.e., provided that $(\frac{C_{11}}{2}\sqrt{\lambda})^{\frac{C_{11}}{2}\sqrt{\lambda}} > d^{1/2}$. Equivalently, we require $\frac{C_{11}}{2}\sqrt{\lambda} \log(\frac{C_{11}}{2}\sqrt{\lambda}) \geq \frac{1}{2}\log d$. This is satisfied if $\frac{C_{11}}{2}\sqrt{\lambda} \geq 2\log d$ and $\log(\frac{C_{11}}{2}\sqrt{\lambda}) \geq \frac{1}{4}$. Since $d \geq 2$, we have $2\log d > e^{1/4}$ and the latter condition is redundant. Thus, for (19) to hold it suffices to have $\lambda \geq \left(\frac{4\log(d)}{C_{11}}\right)^2 =: C_{12}(\log d)^2$. ■

Remark 10. The above estimates provide an error bound even when the minimum requirement for the oversampling rate is not met. In fact, when the term $r^r d^{1/2}$ dominates in the maximization of (18), we obtain a bound of $C_{13} d^{\frac{1}{2}} e^{-C_{14}\lambda}$. The explicit d-dependence of the constant for comparatively small oversampling rates is to be expected, because for $\lambda = 1$, the errors arising at each sample are independent. Thus the total error will behave like $d^{\frac{1}{2}}$ if the quantization accuracy for the individual samples stays fixed.

Remark 11. Using Proposition 2, the constant C_4 can be bounded explicitly by $C_4 \leq \left(3e^2 \left\lceil \frac{\pi^2}{(\cosh^{-1}\gamma)^2} \right\rceil\right)^{-\frac{1}{2}}$, where $\gamma = 2K - \frac{2\mu}{\delta}$.

Remark 12. In Lemma 3 and Theorem 8 we may replace E with $\widetilde{E} := EW$, where $W \in \mathbb{R}^{d\times d}$ is any orthonormal matrix, without changing the conclusions. In fact, the proof is invariant under a right multiplication by W. In Theorem 9 one can then choose any such $\widetilde{E}$ in place of E.

IV. Bounds for harmonic frames

In this section, we show that harmonic frames allow for root-exponential error decay, in the oversampling rate $\lambda = N/d$, when used for the $\Sigma\Delta$ quantization of redundant frame expansions. In particular, here too, we will use the $\Sigma\Delta$ scheme (7). We start by defining the harmonic frames for $\mathbb{R}^d$. Let

$$\begin{aligned} E_0(t) &= \frac{1}{\sqrt{2}} \\ E_{2j-1}(t) &= \cos(2\pi jt), \quad j \geq 1 \\ E_{2j}(t) &= \sin(2\pi jt), \quad j \geq 1. \end{aligned}$$

The harmonic frame $E \in \mathbb{R}^{N \times d}$ is given by the coefficients

$$e_{kn} = \sqrt{\frac{2}{d}} E_k \left(\frac{n}{N} \right),$$

where n ranges from 1 to N and k ranges from 0 to $2m$ (when $d = 2m+1$ is odd) or from 1 to $2m$ (when $d = 2m$ is even). Note that in both cases, the harmonic frame is a unit-norm tight frame, thus $\|Ex\|_2^2 = \frac{N}{d} \|x\|_2^2$.

As in the previous section, we seek to bound $\|FD^r\|_{2\to 2}$, where F is the r-th-order Sobolev dual of the harmonic frame E. To that end, we will provide a lower bound for the smallest singular value of the matrix $D^{-r}E$. This allows us to bound from above the largest singular value (i.e., the norm) of the canonical dual of $D^{-r}E$, which is FD^r. The Riemann sum argument of [8] plays a crucial role in our proof. The underlying idea of this argument is to interpret the iterated sum corresponding to the application of the operator D^{-r} as a Riemann sum and then to approximate it by the corresponding integral. We hence need to estimate the vector valued functions $E^{(r)} \in \mathbb{R}^d$ whose coordinates are defined recursively via

$$\begin{aligned} E_k^{(0)}(t) &= E_k(t) \\ E_k^{(r)}(t) &= \int_0^t E_k^{(r-1)}(s) ds. \end{aligned}$$

We will proceed by providing a lower bound for the coordinates of $E^{(r)}(t)$ in Proposition 15 using a Taylor expansion. Then we obtain a lower bound for $\sigma_{min}\left(D^{-r}E\right)$ by controlling $\inf\limits_{v \in \mathbb{R}^d : \|v\|_2 = 1} \left| \left\langle v, E^{(r)}(t) \right\rangle \right|$ explicitly, via Proposition 16 below. The main idea here is that the resulting bound can be expressed using a Vandermonde matrix; then the estimate follows from the invertibility of Vandermonde matrices. We will then use this result and the Riemann sum argument to obtain our upper bound on $\|FD^r\|_{2\to 2}$ in Lemma 17. Equipped with this bound, we will then be able to show our desired result on $\Sigma\Delta$ quantization for harmonic frame expansions, Theorem 18.

Remark 13. It is interesting to note that D^r is a banded Toeplitz matrix, hence “close” to being a circulant matrix C_r. Circulant matrices are diagonal in the discrete Fourier transform basis. In particular, here, the columns of the Harmonic frame correspond to the singular vectors of C_r associated with the smallest singular values. In other words, had D^r been circulant, the Harmonic frame and its canonical dual could be used to obtain root-exponential precision in the $\Sigma\Delta$ frame quantization context. However, since D^r is only “close” to circulant, it is not true that the Harmonic frame diagonalizes it; hence, more work is necessary to obtain root-exponential precision and the use of Sobolev duals is warranted. Furthermore, we note that it is not possible to modify the $\Sigma\Delta$ scheme to induce a circulant matrix in the analysis, as that would correspond to a non-causal system where updating the current state variable requires knowledge of future values of the sequence.

Remark 14. We will assume from now on that r is above a sufficiently large d-dependent threshold, in particular large enough to satisfy equation (21) below. This assumption is justified, as root-exponential decay will eventually be achieved by choosing the order r for each N such that $r(N) \to \infty$ when $N \to \infty$. The values of r below this threshold hence correspond to finitely many values of N and can be treated by possibly introducing an additional (d-dependent) constant, thus adjusting the values of C_{15}, C_{16}, etc., in the following results.

Proposition 15. *Let $0 \leq t \leq 1$, $j > 0$, r sufficiently large, and let $E_k^{(r)}(t)$ be as above. Then*

$$|E_0^{(r)}(t)| = \frac{t^r}{\sqrt{2}r!} \tag{20}$$

$$|E_{2j}^{(r)}(t)| \geq \frac{t^r}{r!}\sum_{\ell=0}^{m-1}(-1)^\ell\left(\frac{2\pi jt}{r+2m}\right)^{2\ell+1} - 2\left(\frac{t}{r}\right)^{r+2m}$$

$$|E_{2j-1}^{(r)}(t)| \geq \frac{t^r}{r!}\sum_{\ell=0}^{m-1}(-1)^\ell\left(\frac{2\pi jt}{r+2m}\right)^{2\ell} - 2\left(\frac{t}{r}\right)^{r+2m}.$$

Proof: The identity (20) follows directly by induction in r. For $j > 0$, we have by repeated integration of the series expansions of sine and cosine, for r odd

$$E_{2j-1}^{(r)}(t) = \frac{(-1)^{\frac{r-1}{2}}}{(2\pi j)^r}\sum_{\ell=0}^{\infty}(-1)^\ell\frac{(2\pi jt)^{r+2\ell}}{(r+2\ell)!}$$

$$E_{2j}^{(r)}(t) = \frac{(-1)^{\frac{r+1}{2}}}{(2\pi j)^r}\sum_{\ell=0}^{\infty}(-1)^\ell\frac{(2\pi jt)^{r+2\ell+1}}{(r+2\ell+1)!}$$

and for r even

$$E_{2j-1}^{(r)}(t) = \frac{(-1)^{\frac{r}{2}}}{(2\pi j)^r}\sum_{\ell=0}^{\infty}(-1)^\ell\frac{(2\pi jt)^{r+2\ell}}{(r+2\ell)!}$$

$$E_{2j}^{(r)}(t) = \frac{(-1)^{\frac{r}{2}}}{(2\pi j)^r}\sum_{\ell=0}^{\infty}(-1)^\ell\frac{(2\pi jt)^{r+2\ell+1}}{(r+2\ell+1)!}.$$

Using that for r large enough (cf. Remark 14), each term dominates the subsequent one, we write

$$\left|\sum_{\ell=0}^{\infty}(-1)^\ell\frac{(2\pi jt)^{r+2\ell}}{(r+2\ell)!}\right| = \sum_{\ell'=0}^{\infty}\frac{(2\pi jt)^{r+4\ell'}}{(r+4\ell')!} - \frac{(2\pi jt)^{r+4\ell'+2}}{(r+4\ell'+2)!}. \tag{21}$$

To bound this expression, we estimate for integers $0 < \gamma < 2m$:

$$\frac{1}{(r+\gamma)!} - \frac{1}{r!(r+2m)^\gamma} = \frac{r!(r+2m)^\gamma - (r+\gamma)!}{(r+\gamma)!r!(r+2m)^\gamma} = \frac{r^{\gamma-1}\left(2\gamma m - \sum_{k=1}^{\gamma}k\right) + O(r^{\gamma-2})}{(r+\gamma)!(r+2m)^\gamma}$$

$$= \frac{r^{\gamma+3}\left(\sum_{k=1}^{\gamma}(2m-k)\right) + O(r^{\gamma+2})}{(r+\gamma+2)!(r+2m)^{\gamma+2}} > \frac{\frac{r^2}{3}\left(r^{\gamma+1}\left(\sum_{k=1}^{\gamma+2}(2m-k)\right) + O(r^\gamma)\right)}{(r+\gamma+2)!(r+2m)^{\gamma+2}}$$

$$> \frac{(2\pi jt)^2}{(r+\gamma+2)!} - \frac{(2\pi jt)^2}{r!(r+2m)^{\gamma+2}}.$$

Denoting by $\mathbb{I}_A$ the indicator (characteristic) function of the event A, and combining the above estimate for each $\gamma = 4\ell'$ with (21), we obtain

$$
\begin{aligned}
&\left|\sum_{\ell=0}^{\infty}(-1)^{\ell}\frac{(2\pi jt)^{r+2\ell}}{(r+2\ell)!}\right| \\
\geq\ & \sum_{\ell'=0}^{\left\lfloor\frac{m}{2}\right\rfloor-1}\frac{(2\pi jt)^{r+4\ell'}}{(r+4\ell')!}-\frac{(2\pi jt)^{r+4\ell'+2}}{(r+4\ell'+2)!}+\mathbb{I}_{m\text{ is odd}}\frac{(2\pi jt)^{r+2m-2}}{(r+2m-2)!}-\sum_{\alpha=m}^{\infty}\frac{(2\pi jt)^{r+2\alpha}}{(r+2\alpha)!} \\
\geq & \frac{(2\pi jt)^{r}}{r!}\left(\sum_{\ell'=0}^{\left\lfloor\frac{m}{2}\right\rfloor-1}\frac{(2\pi jt)^{4\ell'}}{(r+2m)^{4\ell'}}-\frac{(2\pi jt)^{4\ell'+2}}{(r+2m)^{4\ell'+2}}+\mathbb{I}_{m\text{ is odd}}\frac{(2\pi jt)^{r+2m-2}}{(r+2m)^{r+2m-2}}-\sum_{\alpha=m}^{\infty}\frac{(2\pi jt)^{2\alpha}}{r^{2\alpha}}\right) \\
\geq & \frac{(2\pi jt)^{r}}{r!}\left(\sum_{\ell=0}^{m-1}(-1)^{\ell}\frac{(2\pi jt)^{2\ell}}{(r+2m)^{2\ell}}-\left(\frac{2\pi jt}{r}\right)^{2m}\frac{1}{1-\left(\frac{2\pi jt}{r}\right)^{2}}\right) \\
\geq & \frac{(2\pi jt)^{r}}{r!}\left(\sum_{\ell=0}^{m-1}(-1)^{\ell}\frac{(2\pi jt)^{2\ell}}{(r+2m)^{2\ell}}-2\left(\frac{2\pi jt}{r}\right)^{2m}\right)
\end{aligned}
$$

and thus

$$
|E_{2j-1}^{(r)}| \geq \frac{t^r}{r!}\left(\sum_{\ell=0}^{m-1}(-1)^{\ell}\frac{(2\pi jt)^{2\ell}}{(r+2m)^{2\ell}}-2\left(\frac{2\pi jt}{r}\right)^{2m}\right).
$$

Similarly, one obtains

$$
|E_{2j}^{(r)}| \geq \frac{t^r}{r!}\left(\sum_{\ell=0}^{m-1}(-1)^{\ell}\frac{(2\pi jt)^{2\ell+1}}{(r+2m)^{2\ell+1}}-2\left(\frac{2\pi jt}{r}\right)^{2m}\right).
$$

■

Proposition 16. *Let $v \in \mathbb{R}^d$, be such that $\|v\|_2 = 1$. There exist constants C_{15}, C_{16} and C_{17} independent of v and r (but possibly depending on d) such that for all r large enough*

$$
\left|\left\langle v, E^{(r)}(t)\right\rangle\right| \geq \frac{t^r}{r!}C_{15}\left(\frac{t}{r+d}\right)^{d-1}
$$

and

$$
\int_0^1\left|\left\langle v, E^{(r)}(t)\right\rangle\right|^2 dt \geq \frac{C_{16}e^r}{r^{2r+C_{17}}}
$$

Proof: By Proposition 15, we have

$$
\begin{aligned}
&\left\langle v, E^{(r)}(t)\right\rangle \\
&\geq \frac{t^r}{r!}\left(\frac{v_0}{\sqrt{2}}\mathbb{I}_{d\text{ is odd}}+\sum_{j=1}^{m}v_{2j-1}\left(\sum_{\ell=0}^{m-1}(-1)^{\ell}\left(\frac{2\pi jt}{r+d}\right)^{2\ell}\right)+v_{2j}\left(\sum_{\ell=0}^{m-1}(-1)^{\ell}\left(\frac{2\pi jt}{r+d}\right)^{2\ell+1}\right)\right)+O\left(\left(\frac{t}{r}\right)^{r+2m}\right) \\
&= \frac{t^r}{r!}\left(\frac{v_0}{\sqrt{2}}\mathbb{I}_{d\text{ is odd}}+\sum_{\ell=0}^{m-1}(-1)^{\ell}\left(\frac{2\pi t}{r+d}\right)^{2\ell}\left[\sum_{j=1}^{m}v_{2j-1}j^{2\ell}\right]+(-1)^{\ell}\left(\frac{2\pi t}{r+d}\right)^{2\ell+1}\left[\sum_{j=0}^{m}v_{2j}j^{2\ell+1}\right]\right)+O\left(\left(\frac{t}{r}\right)^{r+2m}\right)
\end{aligned}
\tag{22}
$$

Noting that $V = \left((j^2)^{\ell}\right)_{j=1,\ell=0}^{m,m-1}$ is a Vandermonde matrix, hence invertible, three scenarios are possible:

- $v_0 \neq 0$
- $v' = (v_1, v_3, \ldots, v_{2m-1}) \neq 0$, then $\sum_{j=1}^{m} v_{2j-1} j^{2\ell} = (Vv')_\ell \neq 0$ for some j.
- $v'' = (v_2, v_4, \ldots, v_{2m}) \neq 0$, which implies that also $v''' = (1 \cdot v_2, 2 \cdot v_4, \ldots, m \cdot v_{2m}) \neq 0$. In this case $\sum_{j=1}^{m} v_{2j} j^{2\ell+1} = Vv''' \neq 0$, for some j.

In all three cases, the polynomial

$$P_v(s) := \frac{v_0}{\sqrt{2}} \mathbb{I}_{d \text{ is odd}} + \sum_{\ell=0}^{m-1} (-1)^\ell s^{2\ell} \left[\sum_{j=1}^{m} v_{2j-1} j^{2\ell} \right] + (-1)^\ell s^{2\ell+1} \left[\sum_{j=1}^{m} v_{2j} j^{2\ell+1} \right]$$

is not identically zero. As at most $2m-1$ derivatives of P_v can vanish at 0, we can bound $|P_v(t)|$ near zero from below by $C_{18} t^{2m-1}$. Let

$$C_v = \sup\{C \leq 1 : |P_v(t)| \geq 2Ct^{d-1} \text{ in some neighborhood of } 0\}$$

and

$$q_v = \sup\{q \leq 1 : |P_v(t)| \geq C_v t^{d-1} \text{ for all } |t| \leq q\}.$$

Note that by the factor 2 in the definition of C_v, all q_v are strictly greater than 0. Both C_v and q_v are continuous functions of v, so they assume their minimum on the compact set $\mathbb{S} = \{v : \|v\|_2 = 1\}$. Hence we find $C_{19} = \min\{C_v\}$ and $C_{20} = \min\{q_v\}$ such that for all $v \in \mathbb{S}$, one has, for all $0 \leq s \leq C_{20}$

$$P_v(s) \geq C_{19} s^{d-1}.$$

With (22), this implies that there exists a constant C_{15} independent of v such that for all r large enough

$$\left| \left\langle v, E^{(r)}(t) \right\rangle \right| \geq \frac{t^r}{r!} C_{15} \left(\frac{t}{r+d} \right)^{d-1}.$$

Then

$$\begin{aligned} \int_0^1 \left| \left\langle v, E^{(r)}(t) \right\rangle \right|^2 dt &\geq \int_0^1 \frac{C_{15}^2 t^{2r+2d-2}}{(r!)^2 (r+d)^{2d-2}} dt \\ &= \frac{C_{15}^2}{(r!)^2 (r+d)^{2d-2} (r+2d-1)} \\ &\geq \frac{C_{16} e^r}{r^{2r+C_{17}}}. \end{aligned}$$

■

The next lemma provides a bound for $\|FD^r\|_{2\to2}$.

Lemma 17. *Let F be the r-th-order Sobolev dual of the harmonic frame $E \in \mathbb{R}^{N\times d}$, then there exist (possibly d-dependent) constants C_{21} and C_{22} , such that*

$$\|FD^r\|_{2\to2} \leq C_{21} e^{-r/2} N^{-(r+1/2)} r^{r+C_{22}} \big(1 + O(N^{-1})\big).$$

Proof: By Proposition 16 and following the Riemann sum argument (Lemma $A.1$) in [8] we can now estimate the smallest singular value of $D^{-r}E$ as follows.

$$\begin{aligned}\big(\sigma_{min}(D^{-r}E)\big)^2 &= \inf_{v\in\mathbb{R}^N:\|v\|_2=1}\sum_{i_r=1}^{N}\left|\left\langle v, \sum_{i_{r-1}=1}^{i_r}\cdots\sum_{i_1=1}^{i_2}\sum_{i_0=1}^{i_1}E\Big(\frac{i_0}{N}\Big)\right\rangle\right|^2\\ &\geq \inf_{v\in\mathbb{R}^N:\|v\|_2=1} N^{2r+1}\int_0^1\left|\left\langle v, E^{(r)}(t)\right\rangle\right|^2 dt + O(N^{2r})\\ &\geq \frac{C_{16}e^r}{r^{2r+C_{17}}}N^{2r+1}+O(N^{2r}) =: \alpha.\end{aligned}$$

By the definition of the Sobolev dual, FD^r is the canonical dual frame of $D^{-r}E$, hence $\|FD^r\|_{2\to 2}=\sigma_{min}(D^{-r}E)$ one obtains

$$\|FD^r\|_{2\to2}\leq \alpha^{-1/2} = C_{21}e^{-r/2}N^{-(r+1/2)}r^{r+C_{22}}\big(1+O(N^{-1})\big).$$

■

We are now ready to prove the main result of this section.

Theorem 18. *For $0<K\in\mathbb{Z}$ and $0<\delta\in\mathbb{R}$, let $x\in\mathbb{R}^d$ be such that $\|x\|_2\leq\mu<\delta\left(K-\frac{1}{2}\right)$, and suppose that we wish to quantize the harmonic frame expansion of x with oversampling rate $\lambda=N/d$ using the alphabet $\mathcal{A}=\mathcal{A}_K^\delta$. There exists a Sigma-Delta quantization scheme $Q^{\Sigma\Delta}$ of order $r^\#=r(\lambda)$ such that $\|x-FQ^{\Sigma\Delta}(Ex)\|_2\leq C_{23}e^{-C_{24}\sqrt{\lambda}}$. Here, E is the harmonic frame as above, F its $r^\#$-th order Sobolev dual, and C_{23}, C_{24} are constants, depending on d but independent of x.*

Proof: Again, we use the $\Sigma\Delta$ schemes given in (7) with g, h as in [4] or [5]. Let $\widetilde{x}=FQ^{\Sigma\Delta}(Ex)$ be the linear reconstruction of x from its quantization. Then by Proposition 2 we have that $\|u\|_\infty\leq C_8C_9^r r^r$, and using Lemma 17, we conclude that

$$\begin{aligned}\|x-\widetilde{x}\|_2 &= \|FD^r u\|_2\\ &\leq \|FD^r\|_{2\to2}\|u\|_\infty N^{1/2}\\ &\leq C_{25}C_{26}^r N^{-r}r^{2r+C_{27}}\end{aligned}$$

As in Theorem 9, the optimal order will be of the form $r=\lfloor C_{28}N^{1/2}\rfloor$, yielding

$$\|x-\widetilde{x}\|_2\leq C_{23}\cdot e^{-C_{24}\lambda^{1/2}}$$

as desired. ■

V. Numerical Experiments

In this section, we illustrate our results with some numerical experiments. First, for $d=2$ and $d=6$, we generate 100 random vectors $x\in\mathbb{R}^d$ (from the Gaussian ensemble) and normalize their magnitude to $\|x\|_2=$

$2 - cosh(\pi/\sqrt{6}) \approx 0.0584$. For each x, we obtain the redundant representation $y = Ex$ where $E \in \mathbb{R}^{N \times d}$ is the harmonic frame or the Sobolev self-dual frame of order r. For $r \in \{1, ..., 10\}$ and several values of N, we perform 1-bit $\Sigma\Delta$ quantization on y according to the schemes in Proposition 2. Subsequently, an approximation of x is obtained by linear reconstruction using the r-th-order Sobolev dual of E, and the approximation error is computed. For each N, the smallest (over r) of the maximum error (over the 100 runs) is computed. The resulting error curves for $d = 2$ and 6 are illustrated in Figure 1(a) and 1(b) respectively. Similarly, the smallest (over r) of the mean error (over the 100 runs) is reported in Figure 2 (a) and 2(b). Next, the same experiment is repeated with $d = 20$, this time with 3-bit $\Sigma\Delta$ quantization and 1500 random vectors. In particular, we increase the number of vectors to compensate for the larger size of d in the hope that we can capture the true behavior of the error curves. As before, for each N, the smallest (over r) of the maximum error (over the 1500 runs) is computed and the resulting error curves are illustrated in Figure 3.

From all these experiments, we see that the observed performance indeed matches our predictions both for Sobolev self-dual and harmonic frames. In particular, we observe the root exponential error decay (both for the worst-case and average error). This indicates that at least for these frames, one cannot hope to derive exponential error bounds in the framework of $\Sigma\Delta$ quantization and linear reconstruction via Sobolev duals.

Acknowledgments

The authors thank Sinan Güntürk and Özgur Yılmaz for many useful conversations.

References

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

[2] H. Inose and Y. Yasuda, “A unity bit coding method by negative feedback,” *Proceedings of the IEEE*, vol. 51, no. 11, pp. 1524–1535, 1963.

[3] 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.

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

[5] 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.

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

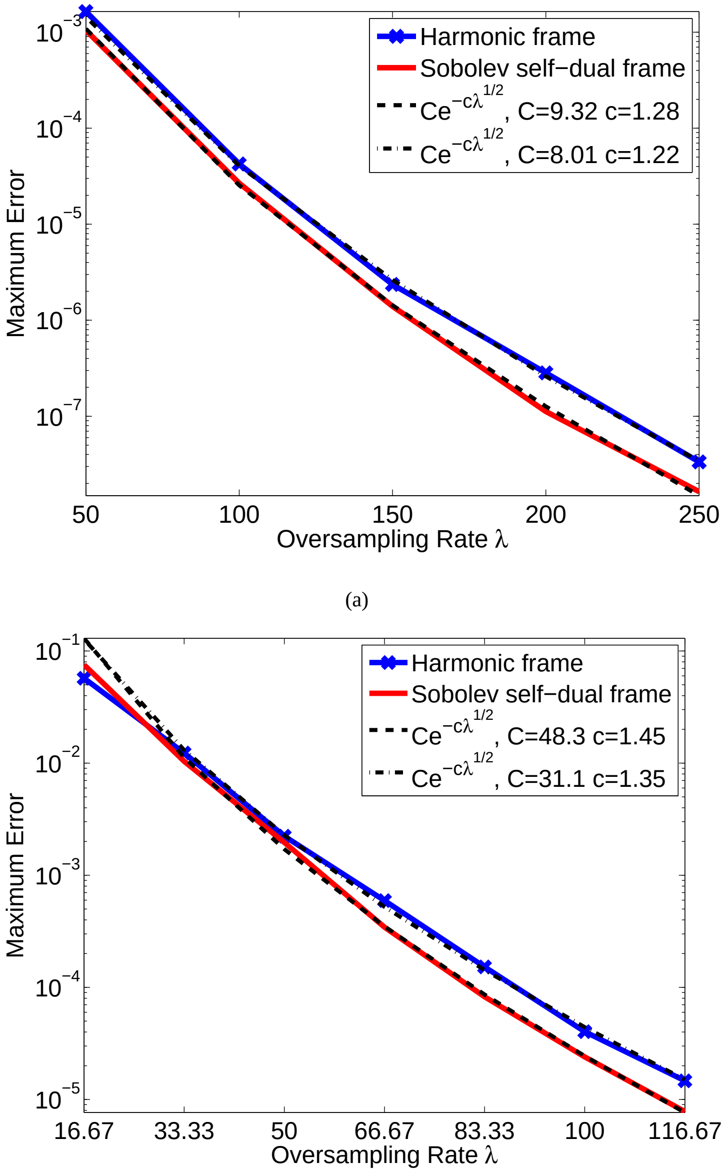
[7] B. Bodmann, V. Paulsen, and S. Abdulbaki, “Smooth frame-path termination for higher order sigma-delta quantization,” *J. Fourier Anal. and Appl.*, vol. 13, no. 3, pp. 285–307, 2007.

[8] 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.

[9] C. Güntürk, M. Lammers, A. Powell, R. Saab, and Ö. Yılmaz, “Sobolev duals for random frames and sigma-delta quantization of compressed sensing measurements,” *Submitted*.

[10] 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.

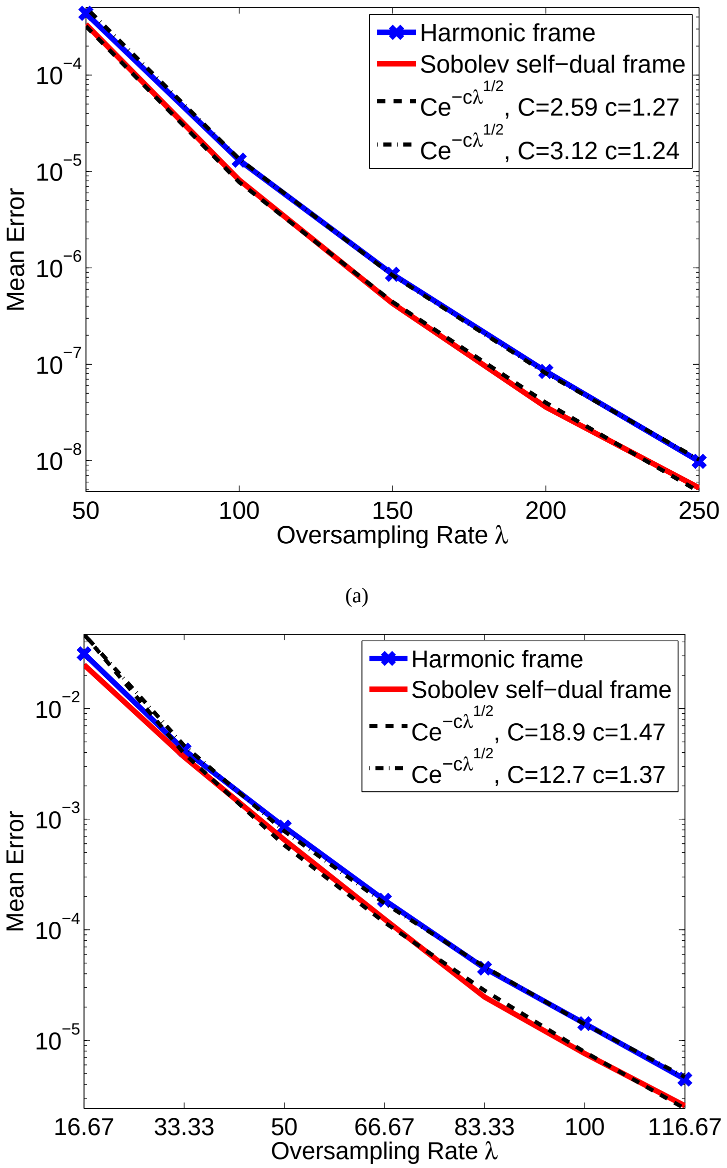
[11] D. Donoho, “Compressed sensing.” *IEEE Trans. Inform. Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.



(a)

(b)

Fig. 1. The maximum error from linear reconstruction of $\Sigma\Delta$ quantized redundant representations, with (a) $d = 2$, and (b) $d = 6$. The error is plotted (in log scale) as a function of the oversampling rate λ.



(a)

(b)

Fig. 2. The mean error from linear reconstruction of $\Sigma\Delta$ quantized redundant representations, with (a) $d = 2$, and (b) $d = 6$. The error is plotted (in log scale) as a function of the oversampling rate λ.

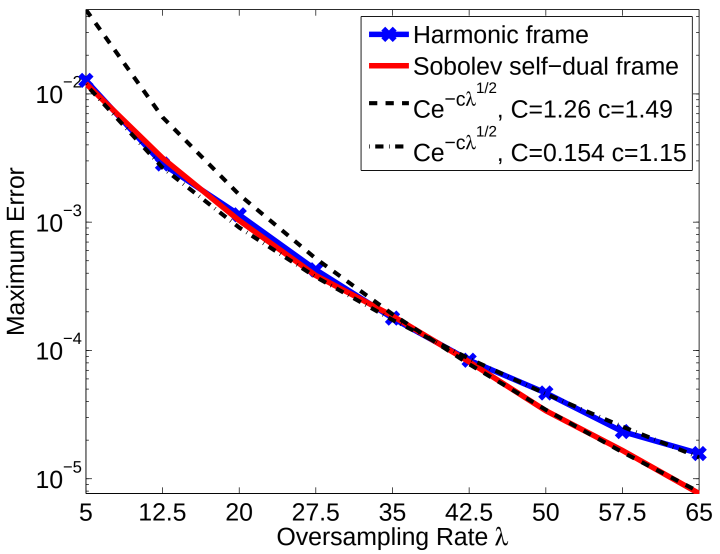


Fig. 3. The maximum error from linear reconstruction of $\Sigma\Delta$ quantized redundant representations, with $d = 20$. The error is plotted (in log scale) as a function of the oversampling rate λ.

[12] 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.

[13] S. R. Norsworthy, R. Schreier, and G. C. Temes, Eds., *Delta-Sigma-Converters: Theory, Design and Simulation*. Wiley-IEEE, 1996.

[14] J. von Neumann, “Distribution of the ratio of the mean-square successive difference to the variance,” *Ann. Math. Statistics*, vol. 12, no. 4, pp. 367–395, 1941.

[15] G. Strang, “The discrete cosine transform,” *SIAM review*, pp. 135–147, 1999.

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