SOBOLEV DUALS IN FRAME THEORY AND SIGMA-DELTA QUANTIZATION

JAMES BLUM, MARK LAMMERS, ALEXANDER M. POWELL, AND ÖZGÜR YILMAZ

ABSTRACT. A new class of alternative dual frames is introduced in the setting of finite frames for $\mathbb{R}^d$. These dual frames, called Sobolev duals, provide a high precision linear reconstruction procedure for Sigma-Delta ($\Sigma\Delta$) quantization of finite frames. The main result is summarized as follows: reconstruction with Sobolev duals enables stable rth order Sigma-Delta schemes to achieve deterministic approximation error of order $\mathcal{O}(N^{-r})$ for a wide class of finite frames of size N. This asymptotic order is generally not achievable with canonical dual frames. Moreover, Sobolev dual reconstruction leads to minimal mean squared error under the classical white noise assumption.

1. INTRODUCTION

In signal processing, an essential problem is to encode a signal, typically an analog object, with finitely many bits in an efficient and robust manner. Efficiency of an encoding is reflected by the associated approximation error: given a bit-budget, one seeks the error to be as small as possible in an appropriate norm. Noting that the encoding process typically involves an analog-to-digital (A/D) conversion step, it is also essential that the required *digital* representation be robust, for example, with respect to errors in arithmetic operations during the encoding process, and partial loss of information. This robustness requirement is frequently satisfied by redundantly representing signals before the A/D conversion or *quantization* stage. For example, one oversamples if the signals to be encoded are bandlimited functions, or one uses overcomplete frame expansions if the signals of interest are vectors in a finite dimensional Hilbert space.

Redundantly representing a signal provides robustness, but it also makes A/D conversion more complex since coefficients in the representation are correlated. In particular, componentwise scalar quantization, also known as *pulse code modulation* (PCM), is generally suboptimal for redundant expansions. $\Sigma\Delta$ quantization was devised as an alternative to PCM for quantizing bandlimited sampling expansions, [19]. Since then, the practice has matured and $\Sigma\Delta$ quantizers have established themselves as a preferred method for quantizing oversampled bandlimited signals, e.g., see [23]. The mathematical theory of $\Sigma\Delta$ quantization on the other hand is quite recent. After work by Gray, e.g., [14, 16], Daubechies and DeVore [10] established an approximation theoretical framework and provided rigorous results on the relationship between robustness of $\Sigma\Delta$ schemes, redundancy of the associated representations, and the approximation error. The work in [10] gives an explicit construction of stable 1-bit $\Sigma\Delta$ schemes of general order r. Moreover, [10] shows

that in the bandlimited setting the L^∞ approximation error associated with a stable rth order $\Sigma\Delta$ scheme behaves like $1/\lambda^r$ where λ is the oversampling ratio.

Inspired by the work of Daubechies and DeVore, in particular that $\Sigma\Delta$ quantization exploits redundancy more efficiently than PCM for bandlimited functions, [2] investigated the use of $\Sigma\Delta$ quantization for redundant finite frame expansions. The performance of PCM for quantizing finite frames is well investigated, e.g., [13, 12]. It was shown in [2] that even a first order $\Sigma\Delta$ scheme outperforms PCM (with linear reconstruction) in terms of approximation error, at least for certain families of frames with sufficiently high redundancy.

A main focus of the subsequent work in [1, 7, 4, 5, 21] is to achieve better approximation rates (as the redundancy of the frame increases) for higher order $\Sigma\Delta$ schemes. In particular, one desires an approximation error of order $\mathcal{O}(N^{-r})$ with an rth order $\Sigma\Delta$ scheme, where N is the number of frame vectors. It was proved in [21] that for $\Sigma\Delta$ schemes of order 3 or higher one cannot obtain these sought-for approximation rates, at least in the case of harmonic frames, if one uses canonical dual frames in the reconstruction. As a remedy, two main approaches have been developed. One approach concentrates on choosing special Parseval frames and uses the same frame in both analysis and reconstruction after quantization, e.g., [5]. This approach is less computationally intensive as there is no need to compute a dual frame, but it only works for rather specific analysis frames. A different approach used in [21], and in this paper, allows a more flexible choice of analysis frame and chooses a non-canonical dual frame that is specially designed for the particular quantization scheme.

Overview. The work in [21] identifies general conditions under which alternative dual frames can be used in rth order $\Sigma\Delta$ quantization to achieve pointwise reconstruction error of order $\mathcal{O}(N^{-r})$, where N is the frame size. However, [21] does not indicate how to construct such alternative dual frames. Indeed, concrete examples are only given for harmonic frames and the methods are decidedly ad hoc.

The main result of this paper, Theorem 5.4, provides a general constructive method for finding alternative dual frames that are suitable for reconstruction in higher order $\Sigma\Delta$ quantization of finite frames. Specifically, we define a new class of alternative dual frames, the Sobolev duals, and prove that the associated reconstruction error for rth order $\Sigma\Delta$ quantization satisfies $||x - \widetilde{x}|| = \mathcal{O}(N^{-r})$.

The paper is organized as follows. Section 2 recalls basic frame theory. Section 3 briefly reviews $\Sigma\Delta$ quantization for finite frames. In Section 4 we define Sobolev dual frames. Our main result, Theorem 5.4, on pointwise $\Sigma\Delta$ error bounds with Sobolev duals is proven in Section 5. Section 6 addresses mean squared error with Sobolev duals. Section 7 contains numerical results.

2. FINITE FRAMES

Definition 2.1. A finite collection of vectors $\{e_n\}_{n=1}^N \subseteq \mathbb{R}^d$ is a *frame* for $\mathbb{R}^d$ with frames bounds $0 < A \leq B < \infty$ if

$$(2.1) \qquad \forall x \in \mathbb{R}^d, \quad A\|x\|^2 \leq \sum_{n=1}^N |\langle x, e_n \rangle|^2 \leq B\|x\|^2,$$

where $\|\cdot\| = \|\cdot\|_2$ denotes the Euclidean norm. If $A = B$ then the frame is said to be *tight*, and it is *unit-norm* if $\|e_n\| = 1$ for each n. The frame bounds are taken to be the respective largest and smallest values of A and B such that (2.1) holds.

If $\{e_n\}_{n=1}^N \subseteq \mathbb{R}^d$ is a frame then there exists a *dual frame* $\{f_n\}_{n=1}^N \subseteq \mathbb{R}^d$ such that

$$(2.2) \qquad \forall x \in \mathbb{R}^d, \quad x = \sum_{n=1}^N \langle x, e_n \rangle f_n = \sum_{n=1}^N \langle x, f_n \rangle e_n.$$

If $N > d$ then the frame $\{e_n\}_{n=1}^N \subset \mathbb{R}^d$ is an overcomplete collection and the choice of dual frame $\{f_n\}_{n=1}^N$ in (2.2) is not unique. We will work with the left expansion in (2.2), and hence refer to $\{e_n\}_{n=1}^N$ as the analysis or encoding frame, and the dual frame $\{f_n\}_{n=1}^N$ as the synthesis or decoding frame. Tight frames have the property that the dual frame can be chosen as $f_n = A^{-1}e_n$, where A is the frame bound. For more background on frames see [8, 9].

For finite frames in $\mathbb{R}^d$, the basic definitions can be conveniently reformulated in terms of matrices. Given a frame $\{e_n\}_{n=1}^N \subset \mathbb{R}^d$ we define the associated *frame matrix* E to be the $d \times N$ matrix with e_j as its jth column. In particular, the columns of a $d \times N$ matrix E form a frame for $\mathbb{R}^d$ if and only if E has rank d. In view of this equivalence, we shall identify a frame $\{e_n\}_{n=1}^N$ with its frame matrix E and simply refer to both as frames for $\mathbb{R}^d$.

Let $\{e_n\}_{n=1}^N \subset \mathbb{R}^d$ be a frame for $\mathbb{R}^d$ with dual frame $\{f_n\}_{n=1}^N \subset \mathbb{R}^d$, and let E and F be the corresponding $d \times N$ frame matrices. The frame expansions (2.2) can be equivalently expressed in terms of E and F as

$$FE^* = EF^* = I_d,$$

where I_d is the $d \times d$ identity matrix and S^* denotes the adjoint of a matrix S. Namely, F is a dual frame for E if and only if F is a left inverse to E^*. This naturally leads one to define the *canonical dual frame* associated to E by $F = (EE^*)^{-1}E$. The canonical dual frame is a minimal norm dual frame with respect to various matrix norms described below.

Given a $d \times N$ matrix E with the vectors $\{e_n\}_{n=1}^N$ as its columns

$$\|E\|_{\mathcal{F}} = \sqrt{tr(EE^*)} = \left(\sum_{n=1}^N \|e_n\|^2 \right)^{1/2} \quad \text{and} \quad \|E\|_{op} = \sup_{\|x\|=1} \|Ex\|$$

will respectively denote the Frobenius norm and operator norm of E. Here $tr(\cdot)$ denotes the trace of a square matrix. We will also use the following norm

$$t(E) = \sup\{\|\sum_{n=1}^N u_n e_n\| : u_n \in \{-1, 1\}\}$$
$$= \sup\{\|Eu\| : u = (u_1, \cdots, u_N) \text{ and } u_n \in \{-1, 1\}\}.$$

This is simply the operator norm of E as a mapping from $(\mathbb{R}^N, \|\cdot\|_\infty)$ to $(\mathbb{R}^d, \|\cdot\|_2)$. The notation $t(\cdot)$ is used in view of the work in [4] on total variation of frame paths.

The next lemma is standard. The proof follows easily from Theorem 3.6 in [22].

Lemma 2.2. *Let E be a $d \times N$ frame matrix and let F be an arbitrary dual frame to E. The three quantities $\|F\|_{op}, \|F\|_{\mathcal{F}}$ and $t(F)$ are minimized when F is taken to be the canonical dual frame of E, namely $F = (EE^*)^{-1}E$.*

The following standard lemma relates the lower frame bound of a frame to the operator norm of its canonical dual. We include a proof for the sake of completeness.

Lemma 2.3. *Let E be an $d \times N$ frame matrix. If F is the canonical dual frame matrix of E then $\|F\|_{op} = A^{-1/2}$ where A is the lower frame bound for E.*

Proof. Note that the frame inequality (2.1) can be rewritten as

$$\forall x \in \mathbb{R}^d, \quad A\|x\|^2 \le \|EE^*x\|^2 \le B\|x\|^2.$$

Also, the canonical dual frame is given by $F = (EE^*)^{-1}E$ so that

$$FF^*EE^* = (EE^*)^{-1}EE^*(EE^*)^{-1}EE^* = I.$$

Since FF^* and EE^* are positive and are inverses of each other the result follows. □

3. Sigma-Delta quantization of finite frame expansions

Quantization is the process of digitally encoding frame coefficients $\langle x, e_n \rangle$ by replacing them with elements of a finite set of numbers $\mathcal{A}$ known as a *quantization alphabet*. Given a finite set $\mathcal{A} \subset \mathbb{R}$, the associated *scalar quantizer* is defined by

$$(3.1) \qquad\qquad Q(u) = \arg\min_{q \in \mathcal{A}} |u - q|.$$

If $Q(u)$ has two possible values for a specific u then arbitrarily specify one.

Sigma-Delta ($\Sigma\Delta$) schemes are a class of recursive algorithms that directly make use of dependencies among frame vectors in the quantization process. We give a brief description of $\Sigma\Delta$ schemes, but refer the reader to [10, 23] for more detailed background. Suppose that $\{e_n\}_{n=1}^N \subset \mathbb{R}^d$ is a frame for $\mathbb{R}^d$ and that $x \in \mathbb{R}^d$ has frame coefficients $x_n = \langle x, e_n \rangle$. An rth order $\Sigma\Delta$ scheme with alphabet $\mathcal{A}$ runs the following iteration

$$(3.2) \qquad q_n = Q(\rho(u_{n-1}^1, u_{n-1}^2, \cdots, u_{n-1}^r, x_n)), \quad \Delta^r u_n^r = x_n - q_n \quad \text{with}$$
$$u_n^j = \Delta u_n^{j+1}, \quad j = 1, \ldots, r-1,$$

where $u_0^1 = u_0^2 = \cdots, u_0^r = 0$, and where the iteration runs for $n = 1, \cdots, N$. Here $\rho : \mathbb{R}^{r+1} \to \mathbb{R}$ is a fixed function, called the *quantization rule*, and Δ^r is the rth order backwards difference operator defined by $\Delta w_n = w_n - w_{n-1}$ and $\Delta^r = \Delta^{r-1}\Delta$. The u_n^j are internal state variables in the algorithm and the $q_n \in \mathcal{A}$ are the desired output coefficients. One can linearly reconstruct a signal $\widetilde{x}$ from the q_n with a dual frame $\{f_n\}_{n=1}^N$ by

$$(3.3) \qquad\qquad \widetilde{x} = \sum_{n=1}^N q_n f_n.$$

The main issue of this paper concerns how to construct a suitable dual frame $\{f_n\}_{n=1}^N$ so that the reconstruction error $\|x - \widetilde{x}\|$ is small.

For a $\Sigma\Delta$ scheme to be useful in practice it should be *stable*. The scheme (3.2) is stable if there exist constants $C_1, C_2 > 0$, such that for any $N > 0$ and any $\{x_n\}_{n=1}^N \subset \mathbb{R}$,

$$(3.4) \quad \forall\, 1 \le n \le N,\ |x_n| \le C_1 \implies \forall\, 1 \le n \le N, \forall j = 1, \cdots, r, \quad |u_n^j| < C_2.$$

The stability constants C_1, C_2 depend on the quantization alphabet $\mathcal{A}$ and the quantization rule ρ. If the frame vectors e_n are uniformly bounded above in norm by a constant M then the frame coefficients satisfy $|x_n| = |\langle x, e_n \rangle| \le M\|x\|$. This allows one to consider stability in terms of the norm of the input signal $x \in \mathbb{R}^d$

since $\|x\| \le C_1/M = \eta$ implies $|x_n| \le C_1$. Since we will be working with uniformly norm-bounded families of frames, we shall refer to stability in terms of $\|x\|$ and say that the $\Sigma\Delta$ scheme is stable for inputs $\|x\| \le \eta$.

Constructing a stable $\Sigma\Delta$ scheme requires carefully choosing the quantization rule ρ in (3.2). Stability analysis of $\Sigma\Delta$ schemes can be quite complicated, especially for 1-bit higher order schemes, [10]. The following example shows a basic first order $\Sigma\Delta$ quantizer. For examples of stable higher order $\Sigma\Delta$ schemes, see [10, 18, 25].

Example 3.1. Let Q be the scalar quantizer associated to the alphabet $\mathcal{A}_K^\delta = \{\pm(k - \frac{1}{2})\delta : |k| \le K, \ k \in \mathbb{Z}\}$. The following first order $\Sigma\Delta$ scheme is stable:

$$q_n = Q(u_{n-1} + x_n), \quad u_n = u_{n-1} + x_n - q_n,$$

where $u_0 = 0$ and $n = 1, \cdots, N$. In particular, if the input sequence satisfies $|x_n| \le (K - 1/2)\delta$ then the state variables satisfy $|u_n| \le \delta/2$, e.g., see [2].

The following notation will help simplify the error analysis of $\Sigma\Delta$ schemes. Let D be the first-order difference matrix given by

$$(3.5) \qquad D = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & -1 & \cdots & 0 \\ & & \ddots & \ddots & \\ 0 & \cdots & 0 & 1 & -1 \\ 0 & \cdots & 0 & 0 & 1 \end{bmatrix}_{N \times N}$$

and define the discrete Laplacian $\nabla = D^*D$. Note that D and ∇ are invertible. If one linearly reconstructs from the $\Sigma\Delta$ quantized coefficients q_n, obtained via (3.2), as in (3.3) using a dual frame $\{f_n\}_{n=1}^N$, then the reconstruction error equals

$$(3.6) \qquad \|x - \widetilde{x}\| = \|\sum_{n=1}^N (x_n - q_n)f_n\| = \|\sum_{n=1}^N (\Delta^r u_n^r) \, f_n\| = \|FD^{r*}(u)\|,$$

where $u = [u_1^r, u_2^r, \cdots, u_N^r]^*$ and F is the frame matrix associated to $\{f_n\}_{n=1}^N$.

4. Sobolev dual frames

In this section we introduce the class of Sobolev dual frames.

Definition 4.1. Let F be a $d \times N$ matrix. Define the Sobolev-type matrix norms

$$\|F\|_{r,op} = \|D^r F^*\|_{op} = \|FD^{r*}\|_{op},$$
$$\|F\|_{r,\mathcal{F}} = \|D^r F^*\|_{\mathcal{F}} = \|FD^{r*}\|_{\mathcal{F}},$$

and the norm, e.g., [4, 5], given by $T_r(F) = t(FD^{r*})$.

Definition 4.2. (Sobolev dual) Fix a positive integer r. Let $\{e_n\}_{n=1}^N \subset \mathbb{R}^d$ be a frame for $\mathbb{R}^d$ and let E be the associated $d \times N$ frame matrix. The rth order *Sobolev dual* $\{f_n\}_{n=1}^N \subset \mathbb{R}^d$ of E is defined so that f_j is the jth column of the matrix

$$F = (ED^{-r}(D^*)^{-r}E^*)^{-1}ED^{-r}(D^*)^{-r},$$

where D is the invertible matrix defined by (3.5).

The next theorem is motivated by the work on alternative dual frames in [11].

Theorem 4.3. *Let E be an $d \times N$ frame matrix. The rth order Sobolev dual F is the dual frame of E for which $\|F\|_{r,op}, \|F\|_{r,\mathcal{F}}$ and $T_r(F)$ are minimal.*

Proof. Note that $FE^* = (ED^{-r}(D^*)^{-r}E^*)^{-1}ED^{-r}(D^*)^{-r}E^* = I$, since D is invertible and E has full rank. Thus F is a dual frame for E.

Since D is invertible, U is a dual frame to E if and only if $UD^{*r}(D^*)^{-r}E^* = UE^* = I$ if and only if UD^{*r} is a dual frame to ED^{-r}. If F is the rth order Sobolev dual of E then

$$(4.1) \qquad FD^{*r} = (ED^{-r}(D^*)^{-r}E^*)^{-1}ED^{-r}$$

is the canonical dual frame of ED^{-r}. By Lemma 2.2, FD^{*r} is the dual frame of ED^{-r} with minimal operator norm, Frobenius norm, and norm $t(\cdot)$. So, the Sobolev dual F is the dual frame of E minimizing $\|\cdot\|_{r,op}$ and $\|\cdot\|_{r,\mathcal{F}}$ and $T_r(\cdot)$. $\square$

5. Signal reconstruction in $\Sigma\Delta$ quantization with Sobolev duals

In this section we prove that Sobolev dual frames provide an effective reconstruction method for $\Sigma\Delta$ quantization of finite frame expansions, and enable rth order $\Sigma\Delta$ schemes to achieve pointwise error of order $1/N^r$, where N is the frame size. We focus on the class of frames associated with piecewise smooth paths in $\mathbb{R}^d$ and adopt the terminology of [5]. Frames with this property arise naturally in quantization problems, e.g., [2, 1, 4, 5]. We say that a function $f : [0,1] \to \mathbb{R}$ is piecewise-C^1 if it is C^1 except at finitely many points in $[0,1]$, and the left and right limits of f and f' exist at all of these points.

Definition 5.1. A vector valued function $E : [0,1] \to \mathbb{R}^d$ given by $E(t) = [e_1(t), e_2(t), \cdots e_d(t)]^*$ is a *piecewise-C^1 uniformly-sampled frame path* if the following three conditions hold:

   (i) For $1 \le n \le N$, $e_n : [0,1] \to \mathbb{R}$ is piecewise-C^1 on $[0,1]$.
   (ii) The functions $\{e_n\}_{n=1}^d$ are linearly independent.
   (iii) There exists N_0 such that for each $N \ge N_0$ the collection $\{E(n/N)\}_{n=1}^N$ is a frame for $\mathbb{R}^d$.

Frame vectors generated by such a frame path are uniformly bounded in norm. Namely, there exists M such that $\|E(n/N)\| \le M$ holds for all n and N.

Example 5.2. Roots-of-unity frame path. Consider the frame path defined by $E(t) = [\cos(2\pi t), \sin(2\pi t)]^*$. It is well known, e.g., [13], that for each $N \ge 3$, the collection $U_N = \{E(n/N)\}_{n=1}^N \subset \mathbb{R}^2$ given by

$$(5.1) \qquad E(n/N) = [\cos(2\pi n/N), \sin(2\pi n/N)]^*, \quad 1 \le n \le N,$$

is a unit-norm tight frame for $\mathbb{R}^2$. Figure 1 shows coordinate components of (a) the canonical dual of U_{200} for $\mathbb{R}^2$, and (b) the associated 2nd order Sobolev dual.

Example 5.3. Repetition frame path. Consider the frame path

$$E(t) = [\chi_{[0,\frac{1}{d}]}(t), \chi_{(\frac{1}{d},\frac{2}{d}]}(t), \cdots \chi_{(\frac{d-1}{d},1]}(t)]^*,$$

where χ_S denotes the characteristic function of S. Note that $R_N = \{E(n/N)\}_{n=1}^N$ is a frame for $\mathbb{R}^d$ for all $N \ge d$. When $N \in d\mathbb{N}$, then R_N is obtained by N-fold repetition of the standard basis for $\mathbb{R}^d$, hence the name *repetition frame*. For $d = 32$ and $N = 192$, Figure 2 shows (a) the components h_1, h_8, and h_{15} of the canonical dual of the repetition frame R_{192} in $\mathbb{R}^{32}$, and (b) the components f_1, f_8, and f_{15} of the associated 2nd order Sobolev dual.

For $s \in \mathbb{R}$, we say that $f(N) = \mathcal{O}(N^s)$ if $\limsup_{N\to\infty} N^{-s}|f(N)| < \infty$.
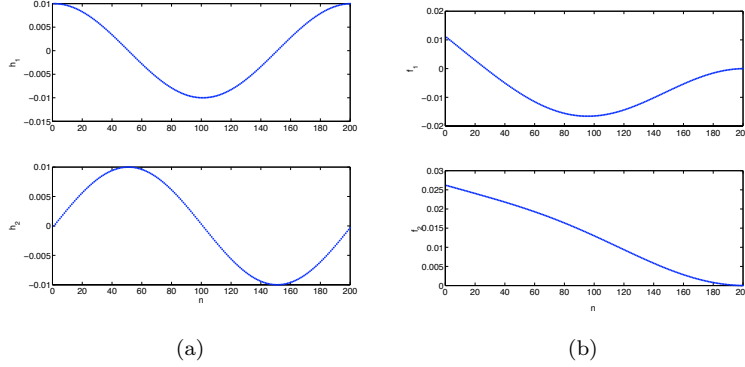
FIGURE 1. Components of (a) the canonical dual, and (b) the 2nd order Sobolev dual for the roots-of-unity frame, see Example 5.2.
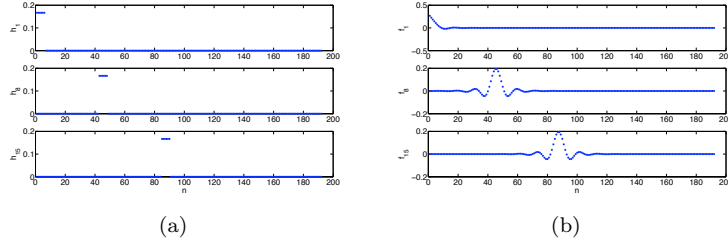

FIGURE 2. Components of (a) the canonical dual, and (b) the 2nd order Sobolev dual for the repetition frame, see Example 5.3.

Theorem 5.4. *Let r be a positive integer, and suppose that one is given an rth order $\Sigma\Delta$ scheme, with quantization alphabet $\mathcal{A}$, that is stable for all inputs $x \in \mathbb{R}^d$ with $\|x\| \leq \eta$ for some $\eta > 0$.*

Let $E(t)$ be a piecewise-C^1 uniformly-sampled frame path for $\mathbb{R}^d$, and suppose that N_0 is such that $E_N = \{E(n/N)\}_{n=1}^N$ is a frame for $\mathbb{R}^d$ for all $N \geq N_0$. Given $x \in \mathbb{R}^d$, $\|x\| \leq \eta$, with frame coefficients $\{\langle x, E(n/N)\rangle\}_{n=1}^N$, let $\{q_n^N\}_{n=1}^N \subset \mathcal{A}$ be the sequence of quantized frame coefficients that are generated by the rth order $\Sigma\Delta$ scheme. If one uses the rth order Sobolev dual frame F_N of E_N to linearly reconstruct an approximation $\widetilde{x}$ to x from the quantized frame coefficients via

$$\widetilde{x} = F_N q$$

where $q = [q_1^N, q_2^N, \cdots, q_N^N]^*$, then

(5.2) $$\|x - \widetilde{x}\| = \mathcal{O}(N^{-r}).$$

The implicit constant may be taken independent of x.

Proof. Since the $\Sigma\Delta$ scheme is stable there exists $C > 0$ such that the $\Sigma\Delta$ state variables satisfy $|u_n^j| \leq C$ for $1 \leq n \leq N$. Letting $u = [u_1^r, u_2^r, \cdots, u_N^r]^*$ gives $\|u\|_2 \leq C\sqrt{N}$, whereby it follows from (3.6) that

$$\|x - \widetilde{x}\|_2 = \|FD^{r*}u\|_2 \leq \|FD^{r*}\|_{op}\|u\|_2 \leq C\sqrt{N}\|FD^{r*}\|_{op}.$$

It therefore suffices to prove that $\|FD^{r*}\|_{op} = \mathcal{O}(N^{-r-1/2})$.

With a slight abuse of notation let $E = E_N$ be the frame matrix associated with $\{E(n/N)\}_{n=1}^N$. Since F is the rth order Sobolev dual of E, we have that FD^{r*} is the canonical dual frame of ED^{-r}, see (4.1). By Lemma 2.3 it is therefore sufficient to prove that the lower frame bound $A = A_N$ of ED^{-r} satisfies

$$(5.3) \qquad\qquad A = A_N \geq \alpha N^{2r+1} + \mathcal{O}(N^{2r}),$$

for some constant $\alpha > 0$.

To estimate the frame bound A_N associated with ED^{-r} first note that D^{-1} is the $N \times N$ upper triangular matrix with 1 in all entries on and above the main diagonal and 0 elsewhere. Thus in the case when $r = 1$ we have that the frame elements associated with ED^{-1} are given by $\{\sum_{j=1}^i E(j/N)\}_{i=1}^N$. Similarly, for general r we have that the frame elements associated with ED^{-r} are given by

$$\left\{ \sum_{i_{r-1}=1}^{i_r} \cdots \sum_{i_1=1}^{i_2} \sum_{i_0=1}^{i_1} E(i_0/N) \right\}_{i_r=1}^N .$$

Let $v = [v_1, v_2, \cdots v_d]^T$ be on the unit sphere of $\mathbb{R}^d$, i.e., $\|v\| = 1$, and let r be an arbitrary positive integer. We estimate the lower frame bound A_N of ED^{-r} as follows. Note that

$$\sum_{i_r=1}^N |\langle v, \sum_{i_{r-1}=1}^{i_r} \cdots \sum_{i_1=1}^{i_2} \sum_{i_0=1}^{i_1} E(\frac{i_0}{N}) \rangle|^2$$

$$= \sum_{i_r=1}^N | \sum_{i_{r-1}=1}^{i_r} \cdots \sum_{i_1=1}^{i_2} \sum_{i_0=1}^{i_1} \langle v, E(\frac{i_0}{N}) \rangle \frac{1}{N^r} |^2 \frac{1}{N} \; N^{2r+1}$$

$$(5.4) \qquad = N^{2r+1} \int_0^1 | \int_0^{t_r} \cdots \int_0^{t_2} \int_0^{t_1} \langle v, E(t_0) \rangle \, dt_0 \, dt_2 \cdots dt_{r-1} |^2 \, dt_r + \mathcal{O}(N^{2r}).$$

Lemma 7.2 in the Appendix provides complete details for the Riemann sum argument used to obtain (5.4).

We next show that if $v \in \mathbb{R}^d$, $\|v\| = 1$, then there exists a subset $S_v \subseteq [0,1]$ with positive measure $0 < |S_v|$ such that

$$t_r \in S_v \implies \gamma_v^{(-r)}(t_r) = \int_0^{t_r} \cdots \int_0^{t_2} \int_0^{t_1} \langle v, E(t_0) \rangle \, dt_0 \, dt_1 \cdots dt_{r-1} \neq 0.$$

Suppose to the contrary that there exists $\|v\| = 1$ such that $\gamma_v^{(-r)}(t) = 0$ for a.e. $t \in [0,1]$. Noting that $\gamma_v^{(-r)}$ is continuous (in fact it is at least $r-1$ times continuously differentiable) we then have that $\gamma_v^{(-r)}(t) = 0$ for all $t \in [0,1]$. Taking r derivatives of $\gamma_v^{(-r)}$ implies that $\langle v, E(t_0) \rangle = 0$ for all $t_0 \in [0,1]$. However this leads to $\sum_{n=1}^d v_n e_n(t_0) = 0$ for all $t_0 \in [0,1]$ which contradicts the linear independence of the component functions $\{e_n(t)\}_{n=1}^d$ of the frame path $E(t)$.

Therefore, for every $\|v\| = 1$ we have that the continuous function

$$B(v_1, v_2, \cdots v_d) = \int_0^1 | \int_0^{t_r} \cdots \int_0^{t_2} \int_0^{t_1} \langle v, E(t_0) \rangle \, dt_0 \, dt_1 \cdots dt_{r-1} |^2 dt_r > 0,$$

is nonzero for all $\|v\| = 1$, and therefore attains a positive minimum over the compact set $\{v \in \mathbb{R}^d : \|v\| = 1\}$. Thus there exists a constant $\alpha > 0$ such that for

all $v \in \mathbb{R}^d, \|v\| = 1$,

$$\sum_{i_r=1}^{N} |\langle v, \sum_{i_{r-1}=1}^{i_r} \cdots \sum_{i_1=1}^{i_2} \sum_{i_0=1}^{i_1} E(\frac{i_0}{N}) \rangle|^2 \geq \alpha N^{2r+1} + \mathcal{O}(N^{2r}).$$

It follows that (5.3) holds. This completes the proof. $\qquad\qquad\square$

Sobolev duals are hierarchical in the following sense: if quantization is performed with an rth order $\Sigma\Delta$ scheme, and if the decoding party uses a Sobolev dual of order $s < r$, then the approximation error will be of order $\mathcal{O}(N^{-s})$. This follows by using norm properties of the rth order Sobolev dual (see the proof of Theorem 5.4) and $\|D\|_{op} \leq 2$ to obtain

$$\|FD^{r*}\|_{op} \leq \|FD^{s*}\|_{op}\|D^{(r-s)*}\|_{op} \leq \mathcal{O}(N^{-s-1/2})\, 2^{(r-s)} = \mathcal{O}(N^{-s-1/2}),$$

which implies that the approximation error satisfies $\|x - \widetilde{x}\| = \mathcal{O}(N^{-s})$. Thus, depending on the required precision, one can use a lower order Sobolev dual, and thereby reduce the computational resources required to construct the dual.

6. Mean Squared Error for $\Sigma\Delta$ quantization with Sobolev duals

This section shows that linear reconstruction with Sobolev duals achieves minimal mean squared error (MSE) under the $\Sigma\Delta$ *white noise assumption* ($\Sigma\Delta$-WNA).

Statistical noise models have a long history in the analysis and design of quantization algorithms, with the general goal being to replace or approximate deterministic analysis with reasonable statistical models. Bennett's classical work [3] modeled individual coefficient quantization errors in simple scalar quantization as independent uniform random variables. While Bennett's noise model has mathematical shortcomings, it is known to be empirically accurate in many settings, especially when the quantizer step size is small, e.g., see [3, 20, 15]. Similar noise models have been applied to the analysis of $\Sigma\Delta$ algorithms, e.g., [15, 23, 24, 6, 2]. We shall consider the following $\Sigma\Delta$ noise model that treats the $\Sigma\Delta$ state variables u_n in (3.6) as independent identically distributed (i.i.d.) random variables.

Design Criterion 6.1 ($\Sigma\Delta$-WNA)**.** Suppose that x, the vector to be quantized, is chosen randomly according to some probability measure supported on a compact set in $\mathbb{R}^d$, and quantized with an rth-order $\Sigma\Delta$ scheme. Model the corresponding $\Sigma\Delta$ state variables $\{u_n\}_{n=1}^{N}$ in (3.6) as independent, identically distributed random variables with mean zero and variance σ^2.

We previously saw that the state variables for the first order $\Sigma\Delta$ scheme in Example 3.1 satisfy $|u_n| \leq \delta/2$. In view of this, it is common to model the state variables u_n as uniform random variables on $[-\delta/2, \delta/2]$. Ergodic properties of the state variables, e.g., see [17, 18], are also closely related to the uniform distribution. We emphasize that the $\Sigma\Delta$-WNA is not mathematically rigorous but is nonetheless empirically reasonable in many settings. For this reason, we simply view the noise model as a useful design criterion, and refer to [24, 14, 16, 15] for discussion of its mathematical validity.

Theorem 6.2. *Let $\{e_n\}_{n=1}^{N} \subset \mathbb{R}^d$ be a frame for $\mathbb{R}^d$ with frame matrix E. Suppose that x is randomly drawn from a compact set B in $\mathbb{R}^d$ according to a Borel probability measure p, supported on B, and suppose that the frame coefficients $\{\langle x, e_n \rangle\}_{n=1}^{N}$ are*
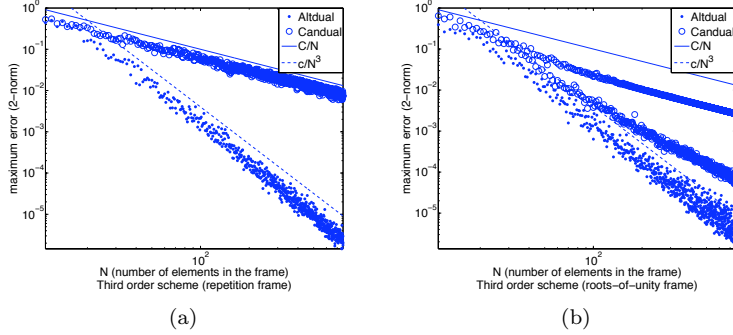
FIGURE 3. Comparison of worst case error in $\Sigma\Delta$ reconstruction with canonical duals and Sobolev duals, see Example 7.1. Errors for (a) repetition frame path, and (b) roots-of-unity frame path.

quantized using an rth order $\Sigma\Delta$ scheme that is stable on B. Under the $\Sigma\Delta$-WNA, among all dual frames F of E, the rth-order Sobolev dual minimizes the MSE

$$MSE = \mathbb{E}\|x - \widetilde{x}\|^2,$$

where $\widetilde{x} = \widetilde{x}_F$ is as in (3.3), and $\mathbb{E}$ denotes the expectation with respect to the probability measure p.

Proof. Let E be a frame and let F be any dual frame. It follows from (3.6) that the MSE is given by

$$MSE = \mathbb{E}\|x - \widetilde{x}\|^2 = \mathbb{E}[\langle FD^{r*}u, FD^{r*}u \rangle] = \sigma^2 \|FD^{r*}\|_{\mathcal{F}}^2.$$

By Theorem 4.3 the MSE is minimal when F is the rth order Sobolev dual of E. □

7. NUMERICAL EXAMPLE

The following experiment numerically compares the performance of Sobolev duals and canonical dual frames for reconstructing $\Sigma\Delta$ quantized frame coefficients.

Example 7.1. 30 points in $\mathbb{R}^2$ are randomly chosen according to the uniform distribution on the unit-square. For each of the 30 points, the corresponding frame coefficients with respect to the repetition frame R_N are quantized using the 3rd order $\Sigma\Delta$ scheme from [10]. Linear reconstruction is then performed with each of the 30 sets of quantized coefficients using both the canonical dual frame and the 3rd order Sobolev dual of R_N. Candual(N) and Altdual(N) will denote the largest of the 30 errors obtained using the canonical dual frame and Sobolev dual respectively. Part (a) in Figure 3 shows a log-log plot of Altdual(N) and Candual(N) against the frame size N. For comparison, log-log plots of $1/N^3$ and $1/N$ are also given.

Part (b) in Figure 3 repeats the above experiment using the roots-of-unity frame instead of the repetition frame. In this case, note the split behavior in the canonical reconstruction which was discussed in [21]. In both parts (a) and (b) Sobolev duals yield smaller reconstruction error than canonical dual frames.

<center>APPENDIX</center>

In this section, we provide the details of the Riemann sum estimates that were used in the proof of Theorem 5.4.

Lemma 7.2. *Let r be a positive integer, $E(t)$ a piecewise-C^1 uniformly-sampled frame path for $\mathbb{R}^d$, and suppose v is on the unit sphere of $\mathbb{R}^d$. Denote*

$$S_r = \sum_{i_r=1}^{N} |\langle v, \sum_{i_{r-1}=1}^{i_r} \cdots \sum_{i_1=1}^{i_2} \sum_{i_0=1}^{i_1} E(\frac{i_0}{N})\rangle|^2$$

Then

$$(7.1) \quad S_r = N^{2r+1} \int_0^1 |\int_0^{t_r} \cdots \int_0^{t_2} \int_0^{t_1} \langle v, E(t_0)\rangle \, dt_0 \, dt_2 \cdots dt_{r-1}|^2 \, dt_r + \mathcal{O}(N^{2r}).$$

Proof. We first show that (7.1) holds when the frame path E is C^1, i.e., the component functions are all continuously differentiable on $[0,1]$. Our proof shall be in several steps. First, we introduce notation to simplify our task. Let $\gamma(t) = \langle v, E(t)\rangle = \sum_{n=1}^d v_n e_n(t)$, and note that γ is continuously differentiable on $[0,1]$. In parts (i) through (v) of the proof, $\|\cdot\|_{L^\infty}$ denotes $\|\cdot\|_{L^\infty[0,1]}$. For a positive integer k, $\gamma^{(-k)}$ shall denote the kth antiderivative of γ with $d^j/(dx^j)[\gamma^{(-k)}](0) = 0$ for all $0 \le j < k$; we also define $\gamma^{(0)} = \gamma$ and $\gamma^{(1)} = \gamma'$.

(i) Using the new notation, we have

$$S_r = \sum_{i_r=1}^{N} |\sum_{i_{r-1}=1}^{i_r} \cdots \sum_{i_1=1}^{i_2} \sum_{i_0=1}^{i_1} \gamma(i_0/N)|^2,$$

and the claim of the lemma can be rewritten as

$$S_r = N^{2r+1} \int_0^1 |\gamma^{(-r)}(t)|^2 dt + \mathcal{O}(N^{2r}).$$

(ii) For any function g that is continuously differentiable on $[0,1]$, and for any positive integer $M \le N$, we have

$$(7.2) \quad |Ng^{(-1)}(M/N) - \sum_{m=1}^{M} g(m/N)| \le \frac{M}{N}\|g'\|_{L^\infty}.$$

This is a simple consequence of the mean value theorem.

(iii) We have

$$L(i_r) := \sum_{i_{r-1}=1}^{i_r} \cdots \sum_{i_1=1}^{i_2} \sum_{i_0=1}^{i_1} \gamma(i_0/N) = N^r \gamma^{(-r)}(i_r/N) + \mathcal{O}(N^{r-1}).$$

To see this, note that by (7.2) we have

$$(7.3) \quad |N\gamma^{(-1)}(i_1/N) - \sum_{i_0=1}^{i_1} \gamma(i_0/N)| \le \frac{i_1}{N}\|\gamma'\|_{L^\infty}, \quad \text{and}$$

$$(7.4) \quad |N^2\gamma^{(-2)}(i_2/N) - \sum_{i_1=1}^{i_2} N\gamma^{(-1)}(i_1/N)| \le i_2\|\gamma\|_{L^\infty},$$

which imply

$$|N^2\gamma^{(-2)}(i_2/N) - \sum_{i_1=1}^{i_2}\sum_{i_0=1}^{i_1}\gamma(i_0/N)| \le \sum_{i_1=1}^{i_2}\frac{i_1}{N}\|\gamma'\|_{L^\infty} + i_2\|\gamma\|_{L^\infty}$$

(7.5)
$$\le \frac{i_2^2}{N}\|\gamma'\|_{L^\infty} + i_2\|\gamma\|_{L^\infty}.$$

By applying this argument repeatedly, it is easy to see that

$$|N^r\gamma^{(-r)}(i_r/N) - L(i_r)| \le \sum_{m=1}^{r} i_r^m N^{r-m-1}\|\gamma^{(m-r+1)}\|_{L^\infty} \le C_r N^{r-1}$$

where we used the fact that $0 \le i_r \le N$. Here C_r is a constant that only depends on r and on E (recall that $\|v\|_2 = 1$, consequently $\|\gamma^{(m)}\|_{L^\infty} \le \sum_{n=1}^{d}\|e_n^{(m)}\|_{L^\infty}$ for any integer $m \le 1$).

(iv) We next observe

$$|(N^r\gamma^{(-r)}(i_r/N))^2 - |L(i_r)|^2| = |N^r\gamma^{(-r)}(i_r/N) - L(i_r)||N^r\gamma^{(-r)}(i_r/N) + L(i_r)|$$
$$\le 2C_r N^{2r-1}\|\gamma^{(-r)}\|_{L^\infty} + C_r^2 N^{2r-2}$$
$$\le N^{2r-1}(2C_r\|\gamma^{(-r)}\|_{L^\infty} + C_r^2/N)$$

(7.6)
$$\le \widetilde{C}_r N^{2r-1}$$

where $\widetilde{C}_r = 2C_r\|\gamma^{(-r)}\|_{L^\infty} + C_r^2$. Consequently, we get

$$|\sum_{i_r=1}^{N} N^{2r}|\gamma^{(-r)}(i_r/N)|^2 - S_r| \quad = \quad |\sum_{i_r=1}^{N} N^{2r}|\gamma^{(-r)}(i_r/N)|^2 - \sum_{i_r=1}^{N}|L(i_r)|^2|$$

$$\le \quad \sum_{i_r=1}^{N}\widetilde{C}_r N^{2r-1} = \widetilde{C}_r N^{2r}.$$

(v) We finally approximate $\sum_{i_r=1}^{N} N^{2r}|\gamma^{(-r)}(i_r/N)|^2$ by an integral. In particular,

$$\sum_{i_r=1}^{N} N^{2r}|\gamma^{(-r)}(i_r/N)|^2 = N^{2r+1}\underbrace{\sum_{i_r=1}^{N}|\gamma^{(-r)}(i_r/N)|^2\frac{1}{N}}_{I_r},$$

and, as I_r is a Riemann sum,

$$|I_r - \int_0^1 |\gamma^{(-r)}(t)|^2 dt| \le \frac{\overline{C}_r}{N}$$

where $\overline{C}_r$ is the L^∞-norm of the derivative of $[\gamma^{(-r)}]^2$. It follows

$$|N^{2r+1}\int_0^1 |\gamma^{(-r)}(t)|^2 dt - S_r| \le (\overline{C}_r + \widetilde{C}_r)N^{2r}$$

which concludes the proof when the frame path is C^1.

(vi) For the extension of this result to the case when E is piecewise C^1 one needs only to modify (7.2), which changes the value of $\widetilde{C}_r$. For simplicity, consider the case $g(x) = f(x)\chi_{[0,a)}(x) + h(x)\chi_{[a,1]}(x)$, where $0 < a < 1$, and f, h are C^1 on the appropriate domains. Clearly, for $M \in \mathbb{N}$ such that $\frac{M}{N} < a$, (7.2) holds. For the

case $\frac{M}{N} \geq a$, we adjust the constant on the right-hand side of (7.2) as follows. Let $k \in \mathbb{N}$ be so that $\frac{k}{N} \leq a < \frac{k+1}{N}$. Hence, $(a - \frac{k}{N}) < \frac{1}{N}$. Thus

$$|N \int_0^{\frac{M}{N}} g(t)dt - \sum_{m=1}^M g(m/N)|$$

$$= |N \int_0^a f(t)dt - \sum_{m=1}^k f(m/N) + N \int_a^{\frac{M}{N}} h(t)dt - \sum_{m=k+1}^M h(m/N)|$$

$$\leq |N \int_0^{k/N} f(t)dt - \sum_{m=1}^k f(m/N)| + |N \int_{k/N}^a f(t)dt|$$

$$+ |N \int_a^{\frac{k+1}{N}} h(t)dt - h(\frac{k+1}{N}) + N \int_{\frac{k+1}{N}}^{\frac{M}{N}} h(t)dt - \sum_{m=k+2}^M h(m/N)|$$

$$\leq \frac{k}{N} \|f'\|_{L^\infty[0,a)} + \|f\|_{L^\infty[0,a)} + |N \int_a^{\frac{k+1}{N}} h(t)dt - (k+1-aN)h(\frac{k+1}{N})|$$

$$+ |(k-aN)h(\frac{k+1}{N})| + \left(\frac{M}{N} - \frac{k+1}{N}\right) \|h'\|_{L^\infty[a,1]}$$

$$\leq \frac{M}{N} \|f'\|_{L^\infty[0,a)} + \frac{M}{Na} \|f\|_{L^\infty[0,a)} + |k+1-aN| \frac{1}{N} \|h'\|_{L^\infty[a,1]}$$

$$+ \|h\|_{L^\infty[a,1]} + \left(\frac{M}{N} - \frac{k+1}{N}\right) \|h'\|_{L^\infty[a,1]}$$

$$\leq \frac{M}{N} \|f'\|_{L^\infty[0,a)} + \frac{M}{Na} \left(\|f\|_{L^\infty[0,a)} + \|h\|_{L^\infty[a,1]}\right) + \left(\frac{M}{N} - a\right) \|h'\|_{L^\infty[a,1]} \leq K \frac{M}{N},$$

where $K = \|f'\|_{L^\infty[0,a)} + (2/a)\|g\|_{L^\infty[0,1]} + \|h'\|_{L^\infty[a,1]}$. So, on the right side of (7.2), we replace $\|g'\|_{L^\infty}$ with a constant K that depends only on the frame path. The remainder of the proof proceeds as in parts (iii) through (v), with the modified constant K changing the constant on the right side of (7.3), and consequently $\widetilde{C}_r$. This argument similarly extends to general functions g that are piecewise C^1. $\square$

Acknowledgments

References

1. J.J. Benedetto, A.M. Powell, and Ö. Yılmaz, *Second order sigma–delta (ΣΔ) quantization of finite frame expansions*, Appl. Comput. Harmon. Anal **20** (2006), 126–148.

2. _____, *Sigma-delta (ΣΔ) quantization and finite frames*, IEEE Transactions on Information Theory **52** (2006), no. 5, 1990–2005.

3. W.R. Bennett, *Spectra of quantized signals*, Bell Syst. Tech. J **27** (1948), no. 3, 446–472.

4. B.G. Bodmann and V.I. Paulsen, *Frame paths and error bounds for sigma–delta quantization*, Applied and Computational Harmonic Analysis **22** (2007), no. 2, 176–197.

5. B.G. Bodmann, V.I. Paulsen, and S.A. Abdulbaki, *Smooth Frame-Path Termination for Higher Order Sigma-Delta Quantization*, Journal of Fourier Analysis and Applications **13** (2007), no. 3, 285–307.

6. H. Bolcskei and F. Hlawatsch, *Noise reduction in oversampled filter banks using predictive quantization*, IEEE Transactions on Information Theory **47** (2001), no. 1, 155–172.

7. P.T. Boufounos and A.V. Oppenheim, *Quantization Noise Shaping on Arbitrary Frame Expansions*, EURASIP Journal on Applied Signal Processing **2006**, Article ID 53807.

8. P.G. Casazza, *The art of frame theory*, Taiwanese J. Math **4** (2000), no. 2, 129–201.

9. O. Christensen, *An Introduction to Frames and Riesz Bases*, Birkhäuser, 2003.

10. I. Daubechies and R. DeVore, *Approximating a bandlimited function using very coarsely quantized data: A family of stable sigma-delta modulators of arbitrary order*, Annals of Mathematics **158** (2003), no. 2, 679–710.

11. I. Daubechies, H.J. Landau, and Z. Landau, *Gabor Time-Frequency Lattices and the Wexler-Raz Identity*, Journal of Fourier Analysis and Applications **1** (1994), no. 4, 437–478.

12. V.K. Goyal, J. Kovačević, and J.A. Kelner, *Quantized Frame Expansions with Erasures*, Applied and Computational Harmonic Analysis **10** (2001), no. 3, 203–233.

13. V.K. Goyal, M. Vetterli, and N.T. Thao, *Quantized overcomplete expansions in $\mathbb{R}^N$: analysis, synthesis, and algorithms*, IEEE Transactions on Information Theory **44** (1998), no. 1, 16–31.

14. R.M. Gray, *Spectral analysis of quantization noise in a single-loop sigma-delta modulator with DC input*, IEEE Transactions on Communications **37** (1989), no. 6, 588–599.

15. ———, *Quantization noise spectra*, IEEE Transactions on Information Theory **36** (1990), no. 6, 1220–1244.

16. R.M. Gray, W. Chou, and P.W. Wong, *Quantization noise in single-loop sigma-delta modulation with sinusoidal inputs*, IEEE Transactions on Communications **37** (1989), no. 9, 956–968.

17. C.S. Güntürk, *Approximating a bandlimited function using very coarsely quantized data: improved error estimates in sigma-delta modulation*, Journal of the American Mathematical Society **17** (2004), no. 1, 229–242.

18. C.S. Güntürk and N.T. Thao, *Ergodic dynamics in sigma-delta quantization: tiling invariant sets and spectral analysis of error*, Adv. in Appl. Math. **34** (2005), no. 3, 523–560.

19. H. Inose and Y. Yasuda, *A unity bit coding method by negative feedback*, Proceedings of the IEEE **51** (1963), no. 11, 1524–1535.

20. D. Jimenez, L. Wang, and Y. Wang, *White noise hypothesis for uniform quantization errors*, SIAM Journal on Mathematical Analysis **38** (2007), 2042–2056.

21. M. Lammers, A.M. Powell, and Ö. Yılmaz, *Alternative dual frames for digital-to-analog conversion in Sigma-Delta quantization*, Advances in Computational Mathematics (2008), To appear.

22. S. Li, *On general frame decompositions*, Numerical Functional Analysis and Optimization **16** (1995), no. 9, 1181–1191.

23. R. Schreier and G.C. Temes, *Understanding delta-sigma data converters*, John Wiley & Sons, 2004.

24. Y. Wang, *Sigma-delta quantization errors and the traveling salesman problem*, Adv. Comput. Math. **28** (2008), no. 2, 101–118.

25. Ö. Yılmaz, *Stability analysis for several second-order sigma-delta methods of coarse quantization of bandlimited functions*, Constructive approximation **18** (2002), no. 4, 599–623.

DEPARTMENT OF MATHEMATICS, UNIVERSITY OF NORTH CAROLINA AT WILMINGTON, WILMINGTON, NC 28403, USA

E-mail address: `blumj@uncw.edu`

E-mail address: `lammersm@uncw.edu`

VANDERBILT UNIVERSITY, DEPARTMENT OF MATHEMATICS, NASHVILLE, TN 37240, USA

E-mail address: `alexander.m.powell@vanderbilt.edu`

DEPARTMENT OF MATHEMATICS, UNIVERSITY OF BRITISH COLUMBIA, VANCOUVER, B.C. CANADA V6T 1Z2

E-mail address: `oyilmaz@math.ubc.ca`