ALTERNATIVE DUAL FRAMES FOR DIGITAL-TO-ANALOG CONVERSION IN SIGMA-DELTA QUANTIZATION

MARK LAMMERS, ALEXANDER M. POWELL, AND ÖZGÜR YILMAZ

Abstract. We design alternative dual frames for linearly reconstructing signals from Sigma-Delta ($\Sigma\Delta$) quantized finite frame coefficients. In the setting of sampling expansions for bandlimited functions, it is known that a stable rth order Sigma-Delta quantizer produces approximations where the approximation error is at most of order $1/\lambda^r$, and $\lambda > 1$ is the oversampling ratio. We show that the counterpart of this result is not true for several families of redundant finite frames for $\mathbb{R}^d$ when the canonical dual frame is used in linear reconstruction. As a remedy, we construct alternative dual frame sequences which enable an rth order Sigma-Delta quantizer to achieve approximation error of order $1/N^r$ for certain sequences of frames where N is the frame size. We also present several numerical examples regarding the constructions.

1. Introduction

Finite frames constitute a natural tool for providing stable signal decompositions in $\mathbb{R}^d$. A frame $\{e_n\}_{n=1}^N \subset \mathbb{R}^d$ for $\mathbb{R}^d$ gives the signal expansions

$$\forall x \in \mathbb{R}^d, \quad x = \sum_{n=1}^{N} \langle x, e_n \rangle f_n, \tag{1.1}$$

where $\{f_n\}_{n=1}^N \subset \mathbb{R}^d$ is a generally non-unique dual frame. An important practical feature of frames is that they can be chosen to be redundant, i.e., one can have $N > d$, which in turn leads to favorable robustness properties in many settings, e.g., [13, 15].

Quantization is the intrinsically lossy process of encoding the "analog" coefficients $\langle x, e_n \rangle$ in (1.1), by a set of "digital" coefficients. This is achieved by replacing each frame coefficient $\langle x, e_n \rangle \in \mathbb{R}$ by some $q_n = q_n(x) \in \mathcal{A}$ where $\mathcal{A} \subset \mathbb{R}$ is a finite set, called the *quantization alphabet*. This process of *encoding* is also referred to as *analog-to-digital (A/D) conversion*. The process of *decoding*, i.e., reconstructing a signal $\widetilde{x} \in \mathbb{R}^d$ from the quantized coefficients q_n, is called *digital-to-analog (D/A)* conversion.

Key words and phrases. Sigma-Delta quantization, finite frames, alternative dual frames.
AMS Subject Classification: 41A99,94A34.

The first author was supported in part by a Cahill Grant from the University of North Carolina at Wilmington.

The second author was supported in part by NSF DMS Grant 0219233, NSF DMS Grant 0504924, and by an Erwin Schrödinger Institute (ESI) Junior Research Fellowship.

The third author was supported in part by the Natural Sciences and Engineering Research Council of Canada.

For the A/D and D/A steps to be meaningful and practical the quantization error $||x-\widetilde{x}||$ should be small, and the process of reconstructing $\widetilde{x}$ should be of reasonable complexity. This makes the following *linear reconstruction* a natural choice:

$$\widetilde{x} = \sum_{n=1}^{N} q_n f_n. \tag{1.2}$$

In this paper, we shall restrict our attention to linear reconstruction methods. It is important to note that linear reconstruction does not provide the most accurate estimate $\widetilde{x}$ that can be obtained from the quantized coefficients. There are alternative methods to linear reconstruction, e.g., consistent reconstruction [16, 29, 13], which give smaller reconstruction error at the cost of greater complexity.

The main goal of this paper is to address the following problem. Suppose one is given a finite frame $\{e_n\}_{n=1}^N$ for $\mathbb{R}^d$ and that for each input signal x the frame coefficients in the associated frame expansion are quantized using a particular quantization scheme. We wish to construct a dual frame $\{f_n\}_{n=1}^N$ that is tailored to the given quantization scheme and improves the approximation when used in place of the canonical dual frame for linearly reconstructing $\widetilde{x}$. Our emphasis will be on the class of higher order Sigma-Delta ($\Sigma\Delta$) quantization schemes.

Overview. The paper is organized as follows. In Section 2 we give necessary definitions and background on finite frames. In Section 3, we define the quantization problem. In Section 4 we discuss background on PCM quantization for finite frames, including basic deterministic estimates, the role of the white noise assumption, and the issue of optimal dual frames. Section 5 contains background on Sigma-Delta quantization in the setting of finite frames, including basic error estimates. In Section 6, we prove a lower bound on the approximation error associated with $\Sigma\Delta$ schemes of order r. This shows that for certain natural choices of unit-norm frames for $\mathbb{R}^d$, if $r \geq 3$ then the approximation error cannot robustly be of order $1/N^r$, where N is the frame size. Section 7 provides a remedy to the obstruction of Section 6, and for a large class of frames establishes sufficient conditions for obtaining an approximation rate of order $1/N^r$ when an rth order $\Sigma\Delta$ scheme is used to quantize finite frame expansions in $\mathbb{R}^d$. Section 8 concludes with specific constructions of alternative dual frames for higher order $\Sigma\Delta$ quantization of the roots-of-unity frames and the harmonic frames.

2. Finite Frames for $\mathbb{R}^d$

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

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

Here and throughout the paper, $||\cdot||$ denotes the Euclidean norm. If $A = B$ then the frame is *tight*. If $||e_n|| = 1$ holds for each $n = 1, \cdots, N$, then the frame is said to be *unit-norm*. We shall fix the convention that all vectors in $\mathbb{R}^d$, in particular the frame vectors, are column vectors.

Given a frame $\{e_n\}_{n=1}^N$ for $\mathbb{R}^d$ the associated *frame operator*, $S : \mathbb{R}^d \longrightarrow \mathbb{R}^d$, is defined by $S(x) = \sum_{n=1}^N \langle x, e_n\rangle e_n$. If $\{e_n\}_{n=1}^N$ is a frame then its frame operator is positive and invertible, e.g., see [1, 2, 11], and one has following canonical frame decompositions

$$\forall x \in \mathbb{R}^d, \quad x = \sum_{n=1}^{N} \langle x, \widetilde{e}_n \rangle e_n = \sum_{n=1}^{N} \langle x, e_n \rangle \widetilde{e}_n, \tag{2.1}$$

where $\widetilde{e}_n = S^{-1} e_n$. In this paper we shall primarily work with the latter of the two decompositions in (2.1). The collection $\{\widetilde{e}_n\}_{n=1}^N$ is a frame for $\mathbb{R}^d$ and is referred to as the *canonical dual frame* to $\{e_n\}_{n=1}^N$. For example, if $\{e_n\}_{n=1}^N$ is a unit-norm tight frame for $\mathbb{R}^d$ then $\widetilde{e}_n = \frac{d}{N} e_n$, e.g., [15, 2]. There are generally, but not always, other frames besides $\{\widetilde{e}_n\}_{n=1}^N$ which lead to decompositions as in (2.1). Any frame $\{f_n\}_{n=1}^N$ for $\mathbb{R}^d$ satisfying

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

is called a *dual frame* to $\{e_n\}_{n=1}^N$. If $\{f_n\}_{n=1}^N$ is not the canonical dual frame then we refer to it as an *alternative dual frame*.

There are many examples of unit-norm tight frames for $\mathbb{R}^d$.

Example 2.1 (Roots-of-unity frames)**.** If $N \geq 3$ then the collection of vectors $\{e_n^N\}_{n=1}^N \subset \mathbb{R}^2$ given by

$$e_n^N = [\cos(2\pi n/N), \sin(2\pi n/N)]^T, \quad n = 1, \cdots, N, \tag{2.3}$$

is a unit-norm tight frame for $\mathbb{R}^2$ with frame bound $A = N/2$ and $\widetilde{e}_n^N = (2/N) e_n$. This family of frames is often referred to as the *roots-of-unity frames*.

Example 2.2 (Harmonic frames)**.** The *harmonic frames* are another important family of frames for $\mathbb{R}^d$. These frames are constructed using columns of the Fourier matrix, e.g., see [15, 32, 16]. The definition of the harmonic frame $H_N^d = \{h_n^N\}_{n=1}^N$, $N \geq d$, depends on whether the dimension d is even or odd.

If d is even let

$$h_n^N = \sqrt{\frac{2}{d}} \left[\cos\frac{2\pi n}{N}, \sin\frac{2\pi n}{N}, \cos\frac{4\pi n}{N}, \sin\frac{4\pi n}{N}, \cos\frac{6\pi n}{N}, \right. \tag{2.4}$$

$$\left. \sin\frac{6\pi n}{N}, \cdots, \cos\frac{2\pi \frac{d}{2} n}{N}, \sin\frac{2\pi \frac{d}{2} n}{N} \right]^T$$

for $n = 1, 2, \cdots, N$. If d is odd let

$$h_n^N = \sqrt{\frac{2}{d}} \left[\frac{1}{\sqrt{2}}, \cos\frac{2\pi n}{N}, \sin\frac{2\pi n}{N}, \cos\frac{4\pi n}{N}, \sin\frac{4\pi n}{N}, \right. \tag{2.5}$$

$$\left. \cos\frac{6\pi n}{N}, \sin\frac{6\pi n}{N}, \cdots, \cos\frac{2\pi \frac{d-1}{2} n}{N}, \sin\frac{2\pi \frac{d-1}{2} n}{N} \right]^T \tag{2.6}$$

for $n = 1, 2, \cdots, N$. It is shown in [32] that H_N^d, as defined above, is a unit-norm tight frame for $\mathbb{R}^d$ with frame bound $A = N/d$ and $\widetilde{e}_n^N = (d/N) e_n^N$.

Given a finite frame $\{e_n\}_{n=1}^N$ for $\mathbb{R}^d$, it is often convenient to work with the associated $d \times N$ matrix $E = [e_1, e_2, \cdots, e_N]$, that has the vectors e_n as its columns. The collection $\{e_n\}_{n=1}^N$ is a frame if and only if the associated matrix E has rank d, and in this case we refer to E as the *frame matrix* associated to $\{e_n\}_{n=1}^N$. If

E is a frame matrix then the associated canonical dual frame has frame matrix $\widetilde{E} = (EE^*)^{-1}E$. In particular, $\widetilde{E}E^* = I_d$, where I_d is the $d \times d$ identity matrix. Moreover, an alternative dual frame to $\{e_n\}_{n=1}^N$ is simply a set of frame vectors $\{f_n\}_{n=1}^N$ whose associated frame matrix F satisfies $FE^* = I_d$. See [23], and also [12, 24], for background material on dual frames. The following result, whose proof we include for the sake of completeness, shows that the canonical dual frame has minimal Frobenius norm among all dual frames.

Lemma 2.3. *Let $\{e_n\}_{n=1}^N$ be a frame for $\mathbb{R}^d$ with frame matrix E. The canonical dual frame $\{\widetilde{e}_n\}_{n=1}^N$ is the dual frame to $\{e_n\}_{n=1}^N$ whose frame matrix $\widetilde{E}$ uniquely has minimal Frobenius norm. That is, $F = \widetilde{E}$ is the matrix satisfying $FE^* = I_d$ for which $||F||_{\text{Frob}} = \sqrt{\text{trace}(F^*F)} = \sqrt{\text{trace}(FF^*)}$ is minimal.*

Proof. Let F be a dual frame. By Theorem 3.6 of [23], F has the form $F = \widetilde{E} + Z$, where $\widetilde{E}$ is the canonical dual frame, and where $ZE^* = 0$. It follows that

$$Z\widetilde{E}^* = 0 \quad \text{and} \quad \widetilde{E}Z^* = 0.$$

Thus,

$$\begin{aligned} ||F||_{\text{Frob}}^2 = ||\widetilde{E} + Z||_{\text{Frob}}^2 &= \text{trace}(\widetilde{E}\widetilde{E}^* + Z\widetilde{E}^* + \widetilde{E}Z^* + ZZ^*) \\ &= \text{trace}(\widetilde{E}\widetilde{E}^* + ZZ^*) = ||\widetilde{E}||_{\text{Frob}}^2 + ||Z||_{\text{Frob}}^2. \end{aligned}$$

Thus, $||F||_{\text{Frob}}^2$ is minimal when $Z = 0$ and $F = \widetilde{E}$ is the canonical dual frame. □

3. The quantization problem

Let $E = \{e_n\}_{n=1}^N \subset \mathbb{R}^d$ be a finite frame for $\mathbb{R}^d$ and fix a finite set $\mathcal{A} \subset \mathbb{R}$ called a *quantization alphabet*.

Definition 3.1. We shall call a map $\mathcal{Q}$ an $\mathcal{A}$-*quantizer* associated to the frame $E = \{e_n\}_{n=1}^N \subset \mathbb{R}^d$, if for each $x \in \mathbb{R}^d$, $\mathcal{Q}$ maps $\{\langle x, e_n\rangle\}_{n=1}^N$ to an element in $\mathcal{A}^N$.

To assess the performance of an $\mathcal{A}$-quantizer, one must compute an approximation to $x \in \mathbb{R}^d$ from the quantized coefficients $\{q_n\}_{n=1}^N = \mathcal{Q}(\{\langle x, e_n\rangle\}_{n=1}^N)$, and check if the approximation error is small. To this end, one must consider a *reconstruction map* $\mathcal{R} : \mathcal{A}^N \to \mathbb{R}^d$ that maps each set of quantized coefficients $\{q_n\}_{n=1}^N$ to a signal $x_{\mathcal{Q},\mathcal{R}} \in \mathbb{R}^d$. For a fixed reconstruction map $\mathcal{R}$, the performance of $\mathcal{Q}$ on a bounded set $B \subset \mathbb{R}^d$ is determined by the associated *distortion* or *approximation error* defined by

$$d_\infty(\mathcal{Q}, \mathcal{R}) = \sup_{x \in B} \|x - \mathcal{R}(\mathcal{Q}(\{\langle x, e_n\rangle\}_{n=1}^N))\|. \tag{3.1}$$

One can also consider the mean-square distortion on B defined by

$$d_{\text{MSE}}(\mathcal{Q}, \mathcal{R}) = \mathcal{E}\|x - \mathcal{R}(\mathcal{Q}(\{\langle x, e_n\rangle\}_{n=1}^N))\|^2, \tag{3.2}$$

where $\mathcal{E}$ denotes the expectation with respect to a Borel probability measure supported on B.

An important class of reconstruction maps consists of all maps $\mathcal{R}_F$ that are defined by linear reconstruction with a fixed dual frame $F = \{f_n\}_{n=1}^N$ of $E = \{e_n\}_{n=1}^N$. In particular, given $\{a_n\}_{n=1}^N \subset \mathbb{R}^N$, one has

$$\mathcal{R}_F : \{a_n\}_{n=1}^N \longmapsto \sum_{n=1}^N a_n f_n. \tag{3.3}$$

Note that for each $x \in \mathbb{R}^d$, one has the perfect reconstruction $x = \mathcal{R}_F(\{\langle x, e_n\rangle\}_{n=1}^N)$.

If $E = \{e_n\}_{n=1}^N$ is an orthonormal basis for $\mathbb{R}^d$, then $N = d$, the associated Bessel map $L : \mathbb{R}^d \to \mathbb{R}^d$, given by $L(x) = \{\langle x, e_n\rangle\}_{n=1}^d$, is a bijection, and there is a unique perfect reconstruction map for E given by $\mathcal{R}_E$. If, on the other hand, E is redundant, i.e., $N > d$, then the associated Bessel map is an injection from $\mathbb{R}^d$ into $\mathbb{R}^N$, and there are infinitely many linear perfect reconstruction maps. In this case, a typical choice is the canonical reconstruction map $\mathcal{R}_{\widetilde{E}}$, i.e., the linear reconstruction map (3.3) obtained by using the canonical dual frame $\widetilde{E}$ of E. The performance of a quantizer is often assessed according to the distortion associated with the canonical reconstruction map, e.g., $d_{\mathrm{MSE}}(\mathcal{Q}, \mathcal{R}_{\widetilde{E}})$ in the case of PCM quantizers, see [16, 15], and $d_\infty(\mathcal{Q}, \mathcal{R}_{\widetilde{E}})$ in the case of $\Sigma\Delta$ quantizers, see [4, 5, 8].

In Section 4 we show that among all linear reconstruction maps, $\mathcal{R}_{\widetilde{E}}$ minimizes the MSE approximation error for PCM schemes under Bennett's white noise assumption on the distribution of the quantization error. On the other hand, in Section 7 we show that in the case of $\Sigma\Delta$ schemes there are alternative dual frames G for which $d_\infty(\mathcal{Q}, G) \leq d_\infty(\mathcal{Q}, \widetilde{E})$, at least for certain classes of frames with sufficiently high redundancy. In Section 8, we construct such alternative dual frames for the roots-of-unity frames (2.3) and for the harmonic frames (2.4) and (2.5). In these constructions and the associated approximation error estimates, we do not make any assumptions on the distribution of the quantization error. See [7] for further analysis of the ramifications of Bennett's white noise assumption in the case of $\Sigma\Delta$ schemes, and in particular a construction of optimal alternative dual frames that minimize $d_{\mathrm{MSE}}(\mathcal{Q}, \mathcal{R}_G)$ for $\Sigma\Delta$ quantizers.

4. PCM Quantization

4.1. **PCM basics.** Let $\{e_n\}_{n=1}^N \subset \mathbb{R}^d$ be a unit-norm frame for $\mathbb{R}^d$, and let $\{f_n\}_{n=1}^N \subset \mathbb{R}^d$ be any dual frame. Fix $\delta > 0$ and $K \in \mathbb{N}$. Given the *$2K$-level midrise quantization alphabet with stepsize δ*,

$$\mathcal{A}_K^\delta = \{(-K+1/2)\delta, (-K+3/2)\delta, \cdots, (-1/2)\delta, (1/2)\delta, \cdots, (K-1/2)\delta\},$$

define the associated *scalar quantizer*

$$Q(u) = \arg\min_{q \in \mathcal{A}_K^\delta} |u - q|. \tag{4.1}$$

For a given $x \in \mathbb{R}^d$ with frame coefficients $\langle x, e_n\rangle$, PCM quantization replaces each $\langle x, e_n\rangle$ by $q_n = Q(\langle x, e_n\rangle)$. One can reconstruct a signal $\widetilde{x} \in \mathbb{R}^d$ using the linear reconstruction

$$\widetilde{x} = \sum_{n=1}^N q_n f_n. \tag{4.2}$$

If $||x|| \leq (K - 1/2)\delta$ then $|\langle x, e_n\rangle - q_n| \leq \delta/2$, and one has the basic PCM error estimate

$$||x - \widetilde{x}|| \leq \frac{\delta}{2} \sum_{n=1}^N ||f_n||. \tag{4.3}$$

4.2. **Bennett's white noise assumption for PCM.** Longstanding analysis dating back to Bennett [6] addresses the average error when a large collection of vectors is quantized with PCM. The fundamental hypothesis, known as *Bennett's white noise assumption*, is that, on average, the error sequence $\{\langle x, e_n\rangle - q_n\}_{n=1}^N$ is well approximated by a sequence of independent, identically distributed uniform random variables with mean 0 and variance $\delta^2/12$. In particular, each $\langle x, e_n\rangle - q_n$ is assumed to be a uniform random variable on $[-\delta/2, \delta/2]$. It is well known that Bennett's white noise assumption is not accurate in general, however it has been shown to be accurate in many circumstances, [6, 28, 30, 21]. For example, the white noise assumption is asymptotically correct as the step size δ approaches 0.

A simple consequence of Bennett's noise assumption, e.g., [16, 15], is that the expected norm squared of the quantization error, i.e., *mean squared error (MSE)*, for PCM quantization is given by

$$MSE_{PCM} = \mathcal{E}(||x - \widetilde{x}||^2) = \frac{\delta^2}{12}\sum_{n=1}^{N} ||f_n||^2, \tag{4.4}$$

where $\mathcal{E}(\cdot)$ denotes the expected value.

To compare (4.3) and (4.4) suppose that $\{e_n\}_{n=1}^N$ is a unit-norm tight frame for $\mathbb{R}^d$ and that $\{f_n\}_{n=1}^N$ is chosen to be the canonical dual frame $f_n = \widetilde{e}_n = \frac{d}{N}e_n$. Then the error estimates (4.3) and (4.4) respectively yield

$$||x - \widetilde{x}||^2 \leq \frac{d^2\delta^2}{4} \quad \text{and} \quad \mathcal{E}||x - \widetilde{x}||^2 = \frac{d^2\delta^2}{12N}.$$

In particular, Bennett's noise assumption implies that one should expect better average performance than predicted by the deterministic estimate (4.3). This improved average performance has been experimentally validated, e.g., [21].

Tight frames were shown to play an important role in PCM quantization under noise models such as Bennett's, see [15]. For example, it was shown in [15] that tight frames minimize MSE under certain assumptions.

4.3. **Optimal dual frames for PCM.** Under Bennett's noise assumption, one can show that the canonical dual frame is optimal for linear reconstruction, cf., [10, 15].

Theorem 4.1. *Let $\{e_n\}_{n=1}^N$ be a frame for $\mathbb{R}^d$. The canonical dual frame $\{\widetilde{e}_n\}_{n=1}^N$ is the dual frame that minimizes MSE_{PCM} in (4.4), the mean squared error for PCM quantization under Bennett's white noise assumption.*

Proof. Let $\{f_n\}_{n=1}^N$ be a dual frame to $\{e_n\}_{n=1}^N$, and let $F = [f_1, \cdots, f_N]$ be its frame matrix. Thus if $\{f_n\}_{n=1}^N$ is used to linearly reconstruct PCM quantized frame coefficients as in (4.2), then

$$MSE_{PCM} = \mathcal{E}||x - \widetilde{x}||^2 = \frac{\delta^2}{12}\sum_{n=1}^{N} ||f_n||^2 = \frac{\delta^2}{12}||F||^2_{\text{Frob}}.$$

Here, the second equality follows from Bennett's white noise assumption. The proof now follows, since by Lemma 2.3 the canonical dual frame is the dual frame whose frame matrix has minimal Frobenius norm. □

It is important to view Theorem 4.1 in the correct perspective. The theorem says that *under Bennett's white noise assumption* the canonical dual frame is optimal for linearly reconstructing PCM quantized frame coefficients. In particular, the theorem is only meaningful when the white noise assumption is reasonably accurate, e.g., when the stepsize $\delta > 0$ is small.

5. Sigma-Delta ($\Sigma\Delta$) quantization

5.1. **First order $\Sigma\Delta$ quantization.** Let $\mathcal{A}_K^\delta$ be the $2K$-level midrise quantization alphabet with stepsize δ, and let Q be the associated scalar quantizer from (4.1). Let $\{e_n\}_{n=1}^N \subset \mathbb{R}^d$ be a unit-norm frame for $\mathbb{R}^d$ with frame operator S. Let $\{f_n\}_{n=1}^N \subset \mathbb{R}^d$ be any, not necessarily unit-norm, dual frame, and let p be a fixed permutation of the index set $\{1, 2, \cdots, N\}$.

Given $x \in \mathbb{R}^d$ satisfying $||x|| \leq (K - 1/2)\delta$, and having frame coefficients $x_n = \langle x, e_n \rangle$, the first order $\Sigma\Delta$ quantizer produces quantized frame coefficients q_n by running the iteration

$$\begin{aligned} q_n &= Q(u_{n-1} + x_{p(n)}), \\ u_n &= u_{n-1} + x_{p(n)} - q_n, \end{aligned} \tag{5.1}$$

for $n = 1, \cdots, N$, and with $u_0 = 0$. The u_n are internal state variables of the $\Sigma\Delta$ scheme, and the q_n are the quantized frame coefficients from which we linearly reconstruct

$$\widetilde{x} = \sum_{n=1}^{N} q_n f_{p(n)}. \tag{5.2}$$

The $\Sigma\Delta$ scheme is *stable*, [14, 4]. In particular,

$$\forall\ 1 \leq n \leq N, \quad |x_n| \leq (K - 1/2)\delta \implies \forall\ 1 \leq n \leq N, \quad |u_n| \leq \delta/2. \tag{5.3}$$

For unit-norm frames, the condition $||x|| \leq (K - 1/2)\delta$ implies $|x_n| = |\langle x, e_n \rangle| \leq (K - 1/2)\delta$, so that $|u_n| \leq \delta/2$ will always hold in our setting.

Error estimates for $\Sigma\Delta$ quantization in the setting of finite frames are given in [3, 4, 5], see also [8, 9]. For example, if the canonical dual frame is used in the reconstruction (5.2), then

$$||x - \widetilde{x}|| \leq \frac{\delta}{2}\ ||S^{-1}||_{\mathrm{op}} \left(\sigma(\{e_n\}_{n=1}^N, p) + 1\right), \tag{5.4}$$

where the *frame variation of* $\{e_n\}_{n=1}^N$ *with respect to* p is defined by $\sigma(\{e_n\}_{n=1}^N, p) = \sum_{n=1}^{N-1} ||e_{p(n)} - e_{p(n+1)}||$. Unlike PCM, the order in which frame coefficients are quantized is quite important in $\Sigma\Delta$ quantization, and this is reflected in the role of the permutation p in the above $\Sigma\Delta$ error estimates. Some examples of frames and good choices of p are given in [4]. For example, for the frame (2.3) in its natural ordering, the frame variation is bounded by 2π and the error estimate (5.4) yields

$$||x - \widetilde{x}|| \leq \frac{\delta}{N}(2\pi + 1), \tag{5.5}$$

which decreases as the frame size N, i.e., redundancy, increases. For the remainder of the paper, we shall work with the given order of a frame, and ignore the role of the permutation p by implicitly assuming that p is the identity.

5.2. **Higher order ΣΔ quantization.** The first order ΣΔ scheme works by controlling a first order difference operator. Higher order ΣΔ schemes work analogously by controlling higher order difference operators. A general rth order ΣΔ scheme runs the following iteration

$$\begin{aligned} q_n &= Q(F(u^1_{n-1}, u^2_{n-1}, \cdots, u^r_{n-1}, x_n)),\\ u^1_n &= u^1_{n-1} + x_n - q_n,\\ u^2_n &= u^2_{n-1} + u^1_n,\\ &\vdots\\ u^r_n &= u^r_{n-1} + u^{r-1}_n, \end{aligned} \tag{5.6}$$

where $u^1_0 = u^2_0 = \cdots, u^r_0 = 0$ and where the iteration runs for $n = 1, \cdots, N$. Here $F : \mathbb{R}^{r+1} \to \mathbb{R}$ is a fixed function, which we refer to as the *quantization rule*. The u^j_n are simply internal state variables in the algorithm and the $q_n \in \mathcal{A}^\delta_K$ are the desired output coefficients.

As with first order schemes, it is important for higher order ΣΔ schemes to be stable. The scheme (5.6) is stable if there exist constants $C_1, C_2 > 0$, such that for any $N > 0$ and any $\{x_n\}_{n=1}^N \subset \mathbb{R}^N$,

$$\forall\ 1 \le n \le N,\ |x_n| < C_1 \implies \forall\ 1 \le n \le N, \forall j = 1, \cdots, r, \quad |u^j_n| < C_2. \tag{5.7}$$

In other words, appropriately bounded input sequences lead to uniformly bounded state variable sequences. As in (5.3), the stability constants $C_1 = C_1(\delta, K)$ and $C_2 = C_2(\delta, K)$ depend on the quantization alphabet $\mathcal{A}^\delta_K$.

The construction of higher order 1-bit ΣΔ schemes, i.e., when $\mathcal{A}^\delta_K$ has $K = 1$, can be a difficult problem. In fact, the existence of arbitrary order stable 1-bit ΣΔ schemes was only recently proven by Daubechies and DeVore in [14]. An alternative family of stable 1-bit ΣΔ schemes of arbitrary order is constructed by Güntürk in [18]. For examples of stable second order ΣΔ schemes see [27, 31, 5].

Example 5.1. The following 1-bit second order ΣΔ scheme is stable, [27, 31, 5].

$$\begin{aligned} q_n &= \text{sign}(u^1_{n-1} + \frac{1}{2}\ u^2_{n-1}),\\ u^1_n &= u^1_{n-1} + x_n - q_n,\\ u^2_n &= u^2_{n-1} + u^1_n, \end{aligned} \tag{5.8}$$

where $u^1_0 = u^2_0 = 0$ and $n = 1, 2, \cdots, N$. Here, sign(x) $= 1$, if $x \ge 0$, and sign(x) $= -1$, if $0 > x$.

In contrast to the 1-bit case, it is elementary to construct stable higher order multi-bit ΣΔ schemes, i.e., where the alphabet $\mathcal{A}^\delta_K$ is sufficiently large, e.g., [20, 8].

Example 5.2. Consider the rth order scheme as in (5.6) with

$$q_n = Q(u^r_{n-1} + u^{r-1}_{n-1} + ... + u^1_{n-1} + x_n)$$

where Q, as in (4.1), is the midrise quantizer with step size 2δ and the number of levels are chosen so that $|u - Q_\delta(u)| < \delta$ for all $u \in [-1, 1]$. If $u^j_{n-1} < 2^{r-j}\delta$, and $|x_n| < 1 - (2^r - 1)\delta$, then one can easily check that $|u^j_n| < 2^{r-j}\delta$, so that the rth order scheme is stable. However, note that for this argument to work, it essential that the quantizer is non-overloading, i.e., δ is sufficiently small so that

$1-(2^r-1)\delta > 0$. Moreover, the scheme is only stable if $|x_n| < 1-(2^r-1)\delta$, that is, if x is in a proper subset of the unit ball in $\mathbb{R}^d$ that shrinks in volume as the order r increases.

Lemma 5.3. *Consider an rth order $\Sigma\Delta$ scheme (5.6) with stability constants $0 < C_1, C_2$ as in (5.7), and suppose that $||x|| < C_1$ has the frame expansion $x = \sum_{n=1}^{N} \langle x, e_n\rangle f_n$. If the frame coefficients $x_n = \langle x, e_n\rangle$ are the input to the $\Sigma\Delta$ scheme then the linear reconstruction $\widetilde{x} = \sum_{n=1}^{N} q_n f_n$ satisfies*

$$x - \widetilde{x} = \sum_{n=1}^{N-r} u_n^r \Delta^r f_n + \sum_{j=1}^{r} u_{N-j+1}^{j} \Delta^{j-1} f_{N-j+1}, \tag{5.9}$$

where $\Delta^0 e_n = e_n$, $\Delta e_n = e_n - e_{n+1}$, and $\Delta^j e_n = \Delta \cdot \Delta^{j-1} e_n$. It follows from (5.9) that for stable schemes

$$||x - \widetilde{x}|| \leq C_2 \sum_{n=1}^{N-r} ||\Delta^r f_n|| + C_2 \sum_{j=1}^{r} ||\Delta^{j-1} f_{N-j+1}||. \tag{5.10}$$

Proof. The proof follows by using that $u_n^j - u_{n-1}^j = u_n^{j-1}$, $u_0^j = 0$, $|u_n^j| \leq C_2$, and applying summation by parts several times to

$$x - \widetilde{x} = \sum_{n=1}^{N} (x_n - q_n) f_n = \sum_{n=1}^{N} (u_n^1 - u_{n-1}^1) f_n = \sum_{n=1}^{N-1} u_n^1 \Delta f_n + u_N^1 f_N.$$

□

We shall refer to the first sum in (5.9) as the *main error term*, and refer to the second sum as *boundary terms*. An analogous computation in the setting of bandlimited signals, see [14], gives a similar main error term. However, the boundary terms are a special consequence of the finite setting here, and are not present in the bandlimited setting.

Existing estimates focus on the case where the dual frame $\{f_n\}_{n=1}^N$ is chosen to be the canonical dual frame $\{\widetilde{e}_n\}_{n=1}^N$, cf., [3, 4, 5, 8, 9]. For example, let $\{e_n\}_{n=1}^N$ be one of the unit-norm tight frames from (2.3), (2.4), or (2.5), and take $\{f_n\}_{n=1}^N = \{\widetilde{e}_n\}_{n=1}^N$ to be the canonical dual frame given by $\widetilde{e}_n = \frac{d}{N} e_n$. For these examples it straightforward to show that $||\Delta^l f_j|| \lesssim 1/N^{l+1}$. This gives the following error estimate

$$||x - \widetilde{x}|| \lesssim \sum_{n=1}^{N-r} 1/N^{r+1} + \sum_{j=1}^{r} 1/N^j \lesssim 1/N^r + 1/N^{r-1} + \cdots 1/N \lesssim 1/N. \tag{5.11}$$

We use the notation $A \lesssim B$ to mean that there is an absolute constant C such that $A \leq CB$. For higher order schemes, the upper bound (5.11) is clearly unsatisfactory since one would like error estimates of order $1/N^r$, analogous to the setting of bandlimited $\Sigma\Delta$ quantization, [14].

Precise knowledge of the state variables in (5.9) can, of course, give better estimates than this, but there are unfortunately situations where the $1/N$ estimate cannot be improved. For example for the second order scheme (5.8) the following error estimate was proven in [5].

$$N \text{ even} \implies ||x - \widetilde{x}|| \leq \frac{4\pi C_2 (2\pi + 1)}{N^2} \lesssim \frac{1}{N^2}, \tag{5.12}$$

and

$$N \text{ odd} \implies \frac{1}{N} \lesssim \frac{2}{N} - \frac{4\pi C_2(2\pi+1)}{N^2} \leq ||x - \widetilde{x}||. \tag{5.13}$$

Note that (5.5) provides error estimates of order $1/N$ for finite frames for the first order $\Sigma\Delta$ scheme, analogous to the error estimates for the first-order scheme in the bandlimited setting. On the other hand, the estimates (5.12) and (5.13) show that a similar analogy for second order schemes is at best partially true. Still, the dichotomy between (5.12) and (5.13) does allow one to robustly obtain approximation of order $1/N^2$ by simply restricting to the case where the frame size N is even. In Section 6 we show that the situation for higher order $\Sigma\Delta$ schemes can be substantially worse if one reconstructs with canonical dual frames.

6. Canonical dual frames and error estimates for higher order $\Sigma\Delta$ schemes

In this section we prove that there are fundamental limitations on the performance of higher order $\Sigma\Delta$ schemes when the canonical dual frame is used for linear reconstruction. We will focus on the frames in Examples 2.1 and 2.2. The following lemma contains some useful lower bounds for approximation error in higher order $\Sigma\Delta$ quantization.

Lemma 6.1. *Suppose we are given a stable rth order $\Sigma\Delta$ scheme (5.6) with $3 \leq r$ and quantization alphabet $\mathcal{A}_K^\delta$, and let $0 < C_1, C_2$ be the associated stability constants as in (5.7). Let $\{e_n\}_{n=1}^N \subset \mathbb{R}^d$ be a unit-norm tight frame for $\mathbb{R}^d$ that satisfies the zero-sum condition*

$$\sum_{n=1}^{N} e_n = 0, \tag{6.1}$$

and also satisfies

$$\forall\ 1 \leq j \leq r,\ 1 \leq n \leq N-j, \qquad A/N^j \leq ||\Delta^j e_n|| \leq B/N^j. \tag{6.2}$$

Given $x \in \mathbb{R}^d$, $||x|| \leq C_1$, suppose that the frame coefficients $\{\langle x, e_n\rangle\}_{n=1}^N$ are quantized using the $\Sigma\Delta$ scheme to obtain quantized coefficients $\{q_n\}_{n=1}^N$. If one uses the canonical dual frame to linearly reconstruct $\widetilde{x} = \frac{d}{N}\sum_{n=1}^N q_n e_n$ then

$$N \textit{ odd} \implies \frac{d\delta}{2N} - \frac{3dC_2B}{N^2} \leq ||x - \widetilde{x}||, \tag{6.3}$$

and

$$N \textit{ even} \implies \frac{dA|u^2_{N-1}|}{N^2} - \frac{2dC_2B}{N^3} \leq ||x - \widetilde{x}||, \tag{6.4}$$

where u^2_{N-1} is the state variable as in (5.6).

Proof. Applying Lemma 5.3 with the canonical dual frame elements $\frac{d}{N}e_n$ gives

$$x - \widetilde{x} = \frac{d}{N}\sum_{n=1}^{N-r} u_n^r \Delta^r e_n + \frac{d}{N}\sum_{j=1}^{r} u^j_{N-j+1}\Delta^{j-1}e_{N-j+1}. \tag{6.5}$$

Next, note that

$$S_1 = ||\sum_{n=1}^{N-r} u_n^r \Delta^r e_n|| \leq \frac{C_2 B}{N^{r-1}} \quad \text{and} \quad S_2 = ||u^2_{N-1}\Delta^1 e_{N-1}|| \leq \frac{C_2 B}{N}, \tag{6.6}$$

$$S_3 = || \sum_{j=3}^{r} u^j_{N-j+1} \Delta^{j-1} e_{N-j+1} || \leq \frac{C_2 B}{N^2}. \tag{6.7}$$

A calculation as in [4] shows that for zero-sum frames

$$|u^1_N| \in \begin{cases} \delta\mathbb{Z}, & \text{if } N \text{ even;} \\ \delta\mathbb{Z} + \delta/2, & \text{if } N \text{ odd.} \end{cases} \tag{6.8}$$

Case 1. If $|u^1_N| \neq 0$ then by (6.8) one has $\delta/2 \leq |u^1_N|$. By (6.5), (6.6), (6.7) one has

$$\frac{d\delta}{2N} \leq \frac{d}{N}|u^1_N| \cdot ||e_N|| \leq ||x - \widetilde{x}|| + \frac{d}{N}(S_1 + S_2 + S_3) \leq ||x - \widetilde{x}|| + \frac{3dC_2B}{N^2}.$$

Case 2. If $|u^1_N| = 0$ then by (6.2), (6.5), (6.6), and (6.7) one has

$$\frac{dA|u^2_{N-1}|}{N^2} \leq \frac{d}{N}|u^2_{N-1}| \cdot ||\Delta^1 e_{N-1}|| \leq ||x - \widetilde{x}|| + \frac{d}{N}(S_1 + S_3) \leq ||x - \widetilde{x}|| + \frac{2dC_2B}{N^3}.$$

□

For the remainder of the section, we will emphasize 1-bit $\Sigma\Delta$ schemes and the roots-of-unity frames (2.3) for $\mathbb{R}^2$. The following theorem shows that there is no *robust* way for a stable rth order $\Sigma\Delta$ algorithm to achieve anything better than $||x - \widetilde{x}|| \lesssim 1/N^2$ when $\widetilde{x}$ is obtained by linearly reconstructing with the canonical dual. Let us clarify what we mean by robust here. Although (6.10) does not eliminate the possibility of estimates of order better than $1/N^2$, any such estimate would at best only be possible for a subsequence of integers $\{N_n\}_{n=1}^{\infty} \subset \mathbb{N}$ *which depends very sensitively on each different input x.* The uniform distribution properties of B_N introduced in (6.11) in the following theorem make it clear that, in practice, one can not reliably obtain approximations of order better than $1/N^2$.

Theorem 6.2. *Suppose we are given a stable 1-bit rth order $\Sigma\Delta$ scheme (5.6) with $3 \leq r$ and quantization alphabet $\mathcal{A}^2_1 = \{-1, 1\}$, and let $0 < C_1, C_2$ be the associated stability constants as in (5.7). Let $E_N = \{e^N_n\}_{n=1}^N \subset \mathbb{R}^2$ be the roots-of-unity frame from (2.3). Suppose $x = (a, b) \in \mathbb{R}^2$ satisfies $||x|| < C_1$, and that the frame coefficients $\langle x, e^N_n \rangle = a\cos(2\pi n/N) + b\sin(2\pi n/N)$ are quantized with the $\Sigma\Delta$ scheme to obtain quantized coefficients $\{q_n\}_{n=1}^N$. If one uses the canonical dual frame to linearly reconstruct $\widetilde{x} = \frac{2}{N}\sum_{n=1}^N q_n e^N_n$, then for $r < N$ one has*

$$N \text{ odd} \implies \frac{1}{N} \lesssim \frac{4}{N} - \frac{6(2\pi)^r C_2}{N^2} \leq ||x - \widetilde{x}||, \tag{6.9}$$

and

$$N \text{ even} \implies \frac{2\pi^r|B_N|}{N^2} - \frac{4(2\pi)^r C_2}{N^3} \leq ||x - \widetilde{x}||, \tag{6.10}$$

where $B_N \in [-1/2, 1/2)$ is defined by

$$B_N \equiv \frac{-aN}{2} + \frac{bN}{2\tan(\pi/N)} \quad \text{modulo} \quad 1. \tag{6.11}$$

In particular, for almost every $x = (a, b) \in \mathbb{R}^2$ satisfying $||x|| < C_1$, one has

$$\frac{2\pi^r}{8} \leq \lim_{M\to\infty} \frac{1}{M} \sum_{N=3}^{M} N^2 ||x - \widetilde{x}_N||.$$

Proof. For $2 \leq N$, one can show that E_N satisfies the requirements of Lemma 6.1 with $A = \pi^r$ and $B = (2\pi)^r$. This follows by noting that since E_N is given by (2.3) one has

$$\forall\ 1 \leq n \leq N-j, \qquad ||\Delta^j e_n|| = |e^{2\pi i/N} - 1|^j.$$

By Lemma 6.1, it suffices to compute the boundary term $|u^2_{N-1}|$. The definition (5.6) shows that

$$u^2_n = \sum_{j=1}^{n} u^1_j \quad \text{and} \quad u^1_n = \sum_{j=1}^{n} x_j - \sum_{j=1}^{n} q_j. \tag{6.12}$$

Since $x = (a,b)$ one has $x_n = \langle x, e^N_n \rangle = a\cos(2\pi n/N) + b\sin(2\pi n/N)$. This together with (6.12) and a direct calculation shows that

$$u^2_{N-1} = \frac{-aN}{2} + \frac{bN}{2\tan(\pi/N)} - \sum_{j=1}^{N-1}\sum_{n=1}^{j} q_n.$$

Since $q_n \in \{-1,1\}$ it follows that $|u^2_{N-1}| \geq |B_N|$, where $B_N \in [-1/2, 1/2)$ is defined by

$$B_N \equiv \frac{-aN}{2} + \frac{bN}{2\tan(\pi/N)} \quad \text{modulo} \quad 1.$$

Finally, note that Lemma 9.4 shows that for almost every $(a,b) \in \mathbb{R}^2$, the sequence $\{B_N\}_{N=1}^{\infty}$ is uniformly distributed modulo 1. See the Appendix for the definition and background results on uniform distribution. By Theorem 9.1 it follows that

$$\begin{aligned}\lim_{M\to\infty} \frac{1}{M}\sum_{N=3}^{M} N^2 ||x - \widetilde{x}_N|| &\gtrsim \lim_{M\to\infty}\frac{1}{M}\sum_{N=3}^{M}\left(2\pi^r|B_N| - \frac{6(2\pi)^2 C_2}{N}\right)\\ &= 2\pi^r \int_{-1/2}^{1/2} |x|\ dx = \frac{2\pi^r}{8}.\end{aligned}$$

□

One can extend the above results to multibit $\Sigma\Delta$ schemes with alphabet $\mathcal{A}^{\delta}_K$ and also to more general frames. For example, in even dimensions d the harmonic frames (2.4) satisfy the hypotheses of Lemma 6.1. If $x \in \mathbb{R}^d$ is of the form $x = (a, b, 0, 0, \cdots, 0)$ one then has an almost identical conclusion as in Theorem 6.2. However, it is perhaps most instructive to simply view Theorem 6.2 as an instance of the general moral that the boundary terms in $\Sigma\Delta$ quantization of unit-norm finite frame expansions can have a serious adverse effect on performance, and require special attention in the setting of finite frames.

7. Alternative dual frames for $\Sigma\Delta$ quantization

Theorem 6.2 shows that canonical dual frames are often not well suited for reconstructing $\Sigma\Delta$ quantized frame coefficients. In this section, we show that one can overcome the difficulties associated with canonical dual frames by instead using alternative dual frames for reconstruction. The main idea is to choose an alternative dual frame for which the associated boundary terms in (5.10) are sufficiently small. We show that this can improve the asymptotic order of the approximation for higher order $\Sigma\Delta$ schemes. In particular, we will specify sufficient conditions for an

rth order $\Sigma\Delta$ scheme to yield approximation error of order $1/N^r$ in the setting of unit-norm finite frames.

We shall focus on the class of frames for which there exist dual frames that can be obtained by sampling a smooth function, or a *frame path*, see [8, 9] for more on frame paths. More precisely, we consider frames with the following property.

Property 7.1. *Fix $r > 0$. Let $E = \{E_N\}_{N=d}^{\infty}$ be a collection of unit-norm frames for $\mathbb{R}^d$, where $E_N = \{e_n^N\}_{n=1}^N \subset \mathbb{R}^d$ has N elements. Suppose that there exists a family of frames $F = \{F_N\}_{N=d}^{\infty}$ with $F_N = \{f_n^N\}_{n=1}^N \subset \mathbb{R}^d$ such that F_N is a dual frame for E_N and satisfies*

$$f_n^N = \frac{1}{N}[\psi_1(n/N), \ldots, \psi_d(n/N)]^T,$$

for some real-valued functions $\psi_i \in C^r[0,1], i = 1, \cdots, d$. Moreover, suppose that there exists $C_\psi(r) > 0$, independent of N, such that the derivatives of ψ_i satisfy

$$\forall\ 1 \le i \le d,\ \forall\ 1 \le j \le r-1, \quad ||\psi_i^{(j)}||_{L^\infty[\frac{N-j}{N},1]} \le \frac{C_\psi(r)}{N^{r-j-1}}, \tag{7.1}$$

and

$$\forall\ 1 \le i \le d, \quad \psi_i(1) = 0. \tag{7.2}$$

One can more generally let the ψ_i depend on N in Property 7.1 and replace (7.2) with the condition

$$|\psi_{i,N}(1)| \le \frac{C_\psi(r)}{N^{r-1}}.$$

The subsequent theorems remain true with this more general condition, but for the sake of simplicity we restrict our discussion to Property 7.1 as originally stated above.

Theorem 7.2. *Fix $r > 0$, and let E_N and F_N be frames for $\mathbb{R}^d$ that satisfy the requirements of Property 7.1. Suppose we are given a stable rth order $\Sigma\Delta$ scheme (5.6) with the associated stability constants $0 < C_1, C_2$ as in (5.7). For $x \in \mathbb{R}^d$, $\|x\| < C_1$, let $\tilde{x} = \sum_{n=0}^{N-1} q_n(x) f_n$ where $q_n(x)$ is produced from the rth order $\Sigma\Delta$ scheme by quantizing $\langle x, e_n\rangle$. Then*

$$\|x - \tilde{x}\| \le \frac{C_{\Sigma\Delta}(r)}{N^r},$$

where $C_{\Sigma\Delta}(r) = C_2\,[C_F(r) + r(r+1)dC_\psi(r)/2]$ and $C_F(r) = \sum_{i=1}^d ||\psi_i^{(r)}||_{L^1[0,1]}$.

Proof. It suffices to use Lemma 5.3 and estimate the two sums in (5.10).

I. To estimate the sum $\sum_{n=1}^{N-r} \|\Delta^r f_n\|$, we use the technique employed in the proof of Proposition 3.1 in [14] and obtain

$$\begin{aligned} \sum_{n=1}^{N-r} \|\Delta^r f_n\| &\le \frac{1}{N}\sum_{i=1}^{d}\sum_{n=1}^{N-r} |\Delta^r \psi_i(n/N)| \\ &\le \frac{1}{N}\sum_{i=1}^{d}\left(\frac{1}{N^{r-1}}||\psi_i^{(r)}||_{L^1[0,1]}\right) = \frac{C_F(r)}{N^r}. \end{aligned} \tag{7.3}$$

II. To estimate the sum $\sum_{j=1}^{r}||\Delta^{j-1}f_{N-j+1}||$, we again use techniques from the proof of Proposition 3.1 in [14] to obtain that if $m+j\leq N$ then

$$\Delta^j\psi_i(\frac{N-m}{N})=(-1)^j\frac{1}{N^{j-1}}\int_0^{j/N}\psi_i^{(j)}(1-m/N+s)\phi_j(Ns)ds,$$

where the difference operator is with respect to m, and ϕ_j is the jth order B-spline. Since $\|\phi_j\|_{L^\infty}\leq 1$, it follows that

$$|\Delta^{j-1}\psi_i(\frac{N-j+1}{N})|\leq\frac{1}{N^{j-2}}\frac{j}{N}\frac{C_\psi(r)}{N^{r-(j-1)-1}}\leq j\frac{C_\psi(r)}{N^{r-1}}.$$

Thus,

$$\sum_{j=1}^{r}\|\Delta^{j-1}f_{N-j+1}\|\leq\frac{1}{N}\sum_{i=1}^{d}\sum_{j=1}^{r}|\Delta^{j-1}\psi_i(\frac{N-j+1}{N})|\leq d\frac{r(r+1)}{2}\frac{C_\psi(r)}{N^r}. \tag{7.4}$$

Combining the estimates (7.3) and (7.4) completes the proof. □

The constants in Theorem 7.2 can be refined, but such improvements do not affect the asymptotic approximation order and will be omitted in this paper. However, having small constants can be important in practice; see [4, 5] and especially [8, 9] for bounds on constants in error expressions for Sigma-Delta quantization of finite frame expansions.

The results of Section 6 show that higher order ΣΔ reconstruction with the canonical dual frame can lead to poor performance, whereas the results of this section, e.g., Theorem 7.2, show that such difficulties can be avoided by reconstructing with alternative dual frames. A main point of using alternative duals is that difficulties in higher order ΣΔ reconstruction are often solely a consequence of the reconstruction frame (e.g., see the proof of Theorem 6.2) and can have very little to do with the "encoding" frame used to compute frame coefficients. Alternative dual frames decouple the encoding and decoding/reconstruction and thereby provide extra flexibility to achieve accurate signal reconstruction.

For comparison, we also wish to point out the relevant work in [9] that is based on tight frames instead of alternative duals. That work has the benefit of essentially using the same frames for encoding and reconstruction, but thereby forces the encoding frame to have special properties that may only be desirable in the reconstruction frame. Consequently, the work in [9] only applies to very particular frames. For example, it does not apply to families of frames such as (2.3), (2.4) and (2.5), whose elements are unit-norm or uniformly bounded away from zero in norm.

8. Examples

8.1. **Roots-of-unity frames for $\mathbb{R}^2$.** It was shown in Section 6 that if one considers higher order ΣΔ quantization of frame expansions given by the roots-of-unity frames (2.3), then the canonical dual frame can perform poorly for linear reconstruction. In this section, we construct explicit alternative dual frames for the roots-of-unity frames E_N that are tailored to the order r of the ΣΔ quantization. We use the results from Section 7 to show that these alternative dual frames vastly improve the approximation order compared to the canonical reconstruction.

Let $E_N = \{e_n^N\}_{n=1}^N \subset \mathbb{R}^2$ be the unit-norm tight frame for $\mathbb{R}^2$ defined by (2.3), with canonical dual frame $\{\frac{2}{N}e_n^N\}_{n=1}^N$. Given $r \geq 1$, set

$$F_N = F_N(r) = \{f_n^N\}_{n=1}^N, \qquad f_n^N = \frac{1}{N}(2e_n^N + g_n^N), \tag{8.1}$$

where

$$g_n^N = \left[a_0 + \sum_{\ell=1}^k a_\ell \cos\left(2\pi(\ell+1)n/N\right), \ \sum_{\ell=1}^k b_\ell \sin\left(2\pi(\ell+1)2\pi n/N\right) \right]^T, \tag{8.2}$$

and $k = k_r > 0$, $\{a_l\}_{l=0}^k$ and $\{b_l\}_{l=1}^k$ are constants to be defined later. Next, we set

$$\psi_1(t) \quad = \quad 2\cos(2\pi t) + a_0 + \sum_{\ell=1}^k a_\ell \cos((\ell+1)2\pi t), \tag{8.3}$$

$$\psi_2(t) \quad = \quad 2\sin(2\pi t) + \sum_{\ell=1}^k b_\ell \sin((\ell+1)2\pi t). \tag{8.4}$$

With this notation, we have $f_n^N = \frac{1}{N}\left[\psi_1(n/N), \psi_2(n/N)\right]^T$.

Lemma 8.1. *If $1 \leq k < N-1$ is a positive integer then F_N, defined by (8.1) and (8.2), is a dual frame to E_N for every choice of $\{a_\ell\}_{\ell=0}^k, \{b_\ell\}_{\ell=1}^k \subset \mathbb{R}$.*

Proof. Let $x = (\alpha, \beta)$ be arbitrary. Since $\langle x, e_n^N \rangle = \alpha\cos(2\pi n/N) + \beta\sin(2\pi n/N)$, a computation using discrete orthogonality relations for cosine and sine shows that $\sum_{n=1}^N \langle x, e_n^N \rangle g_n^N = 0$. It follows that the dual frame relation (2.2) holds. □

Given $r > 0$, our goal is to choose $\{a_\ell\}_{\ell=0}^k$ and $\{b_\ell\}_{\ell=1}^k$ so that ψ_1 and ψ_2 satisfy the requirements of Property 7.1. Theorem 7.2 will then ensure that if rth order $\Sigma\Delta$ quantized frame coefficients from the frame (2.3) are linearly reconstructed with the dual frame F_N then the approximation error will be of order $1/N^r$.

We begin by computing values of $\{a_\ell\}_{\ell=0}^k$, $\{b_\ell\}_{\ell=1}^k$ for which the first $2k+1$ terms in the power series expansions about $t = 0$ of ψ_1 and ψ_2 vanish. Note that ψ_1 is an even function, ψ_2 is an odd function, and both are 1-periodic.

The power series expansions about $t = 0$ for ψ_1 and ψ_2 are given by

$$\psi_1(t) = \sum_{n=0}^\infty \beta_{2n} t^{2n} \quad \text{and} \quad \psi_2(t) = \sum_{n=0}^\infty \beta_{2n+1} t^{2n+1},$$

where

$$\beta_0 = 2 + \sum_{\ell=0}^k a_\ell \quad \text{and} \quad \forall n \geq 1, \ \beta_{2n} = \frac{(-1)^n (2\pi)^{2n}}{(2n)!}\left(2 + \sum_{\ell=1}^k a_\ell (\ell+1)^{2n}\right),$$

$$\forall n \geq 0, \ \beta_{2n+1} = \frac{(-1)^n (2\pi)^{2n+1}}{(2n+1)!}\left(2 + \sum_{\ell=1}^k b_\ell (\ell+1)^{2n+1}\right).$$

Let $\mathbf{V}_k$ be the following $k \times k$ Vandermonde matrix

$$\mathbf{V}_k = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 2^2 & 3^2 & \cdots & (k+1)^2 \\ 2^4 & 3^4 & \cdots & (k+1)^4 \\ \vdots & \vdots & \vdots & \vdots \\ 2^{2k-2} & 3^{2k-2} & \cdots & (k+1)^{2k-2} \end{bmatrix}, \tag{8.5}$$

and let $\mathbf{M}_k$ be the $k \times k$ diagonal matrix with $[2, 3, 4, \cdots, (k+1)]$ as its main diagonal. Note that since $\mathbf{V}_k$ is Vandermonde it is invertible with

$$\det(\mathbf{V}_k) = \prod_{1\leq i<j\leq k} (j+1)^2 - (i+1)^2 > 0.$$

The following lemma is a consequence of the above definitions.

Lemma 8.2. *Fix $k > 0$ and define $\{a_\ell\}_{\ell=0}^k$ and $\{b_\ell\}_{\ell=1}^k$ as follows. Let*

$$\mathbf{a} = [a_1, \cdots, a_k]^T \qquad \textit{and} \qquad \mathbf{b} = [b_1, b_2, \cdots, b_k]^T$$

be chosen as the unique solutions to

$$\mathbf{V}_k\mathbf{M}_k^2\mathbf{a} = -[2, 2, \cdots, 2]^T, \qquad \textit{and} \qquad \mathbf{V}_k\mathbf{M}_k\mathbf{b} = -[2, 2, \cdots, 2]^T,$$

and also let $a_0 = -2 - (a_1 + \cdots + a_k)$. Then $b_\ell = (\ell+1)a_\ell$ holds for $\ell = 1, \cdots, k$, and the functions ψ_1 and ψ_2 defined by (8.3) and (8.4) satisfy

$$\forall\ 0 \leq j \leq 2k, \quad \psi_1^{(j)}(0) = 0 \qquad \textit{and} \qquad \psi_2^{(j)}(0) = 0.$$

Equivalently, $\beta_n = 0$ for all $0 \leq n \leq 2k$.

In terms of Property 7.1, Lemma 8.2 says the following.

Lemma 8.3. *Given $k > 0$, let $\{a_\ell\}_{\ell=0}^k$ and $\{b_\ell\}_{\ell=1}^k$ be defined as in Lemma 8.2. Then $\psi_1, \psi_2 \in C^\infty(\mathbb{R})$ defined by (8.3) and (8.4) satisfy $\psi_1(1) = \psi_2(1) = 0$. Moreover, there exist positive constants $C_{\psi_1} = C_{\psi_1}(k)$ and $C_{\psi_2} = C_{\psi_2}(k)$ such that for all $1 \leq j \leq 2k$ and for all $N > 2k$ there holds*

$$\|\psi_1^{(j)}\|_{L^\infty[\frac{N-j}{N},1]} \leq \frac{C_{\psi_1}}{N^{2k+2-j}} \qquad \textit{and} \qquad \|\psi_2^{(j)}\|_{L^\infty[\frac{N-j}{N},1]} \leq \frac{C_{\psi_2}}{N^{2k+1-j}}. \tag{8.6}$$

Proof. Since ψ_1, ψ_2 are entire functions whose restrictions to $\mathbb{R}$ are 1-periodic, this follows from the power series properties provided by Lemma 8.2. □

Theorem 8.4. *Let $r \geq 3$ be a positive integer and let $E_N = \{e_n^N\}_{n=1}^N \subset \mathbb{R}^2$ be the roots-of-unity frame (2.3). Take $k = \lceil r/2 \rceil - 1$ and define the dual frame $F_N(r) = \{f_n^N\}_{n=1}^N$ by (8.1) and (8.2), where $\{a_\ell\}_{\ell=0}^k, \{b_\ell\}_{\ell=1}^k$ are as defined in Lemma 8.2.*

Suppose we are given a stable rth order $\Sigma\Delta$ scheme (5.6), and let $C_1, C_2 > 0$ be the associated stability constants as in (5.7). For $x \in \mathbb{R}^2$, $\|x\| < C_1$, let $\tilde{x} = \sum_{n=1}^N q_n(x) f_n^N$ where $q_n(x)$ is produced via the rth order $\Sigma\Delta$ scheme by quantizing $\langle x, e_n^N\rangle$. Then

$$\|x - \tilde{x}\| \leq \frac{C_{\Sigma\Delta}^{RU}(r)}{N^r},$$

where the constant $C_{\Sigma\Delta}^{RU}(r)$ is defined by $C_{\Sigma\Delta}^{RU}(r) = C_2\left(C_F^{RU}(r) + r(r+1)C_\Psi(r)\right)$, with $C_\Psi(r) = \max\{C_{\psi_1}, C_{\psi_2}\}$ and $C_F^{RU}(r) = ||\psi_1^{(r)}||_{L^1[0,1]} + ||\psi_2^{(r)}||_{L^1[0,1]}$.

Proof. Since $k = \lceil r/2 \rceil - 1$, this follows from Lemma 8.3 and Theorem 7.2. □

Theorem 8.4 shows that if $r \geq 3$, then rth order $\Sigma\Delta$ quantization of the roots-of-unity frames has approximation error of the desired order $1/N^r$ if the alternative dual frames $F_N(r)$ are used for reconstruction. For second order schemes, one can easily modify the above construction, or simply use $F_N(r), r \geq 3$, to achieve approximation error of order $1/N^2$.

Numerical experiments. Figure 1 shows the canonical dual frame for E_{41} and two alternative dual frames $F_{41}(3)$ and $F_{41}(7)$ that are constructed for Theorem 8.4 for $\Sigma\Delta$ schemes of order $r = 3$ and $r = 7$, respectively.

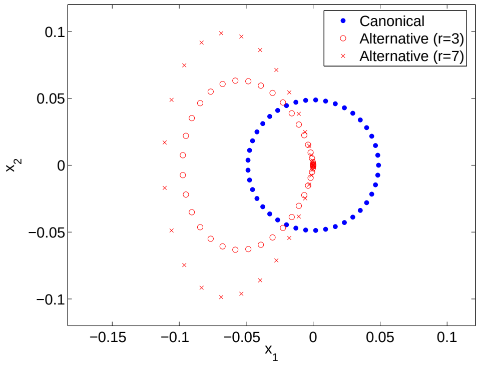
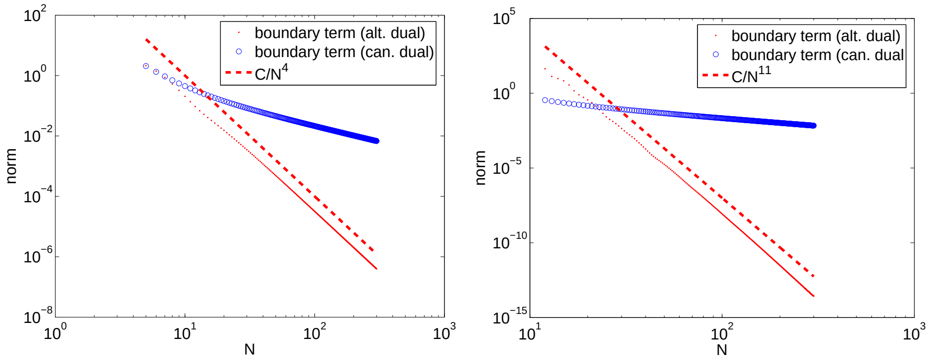


Figure 1. The canonical dual frame of E_{41} and two alternative dual frames $F_{41}(3)$ and $F_{41}(7)$.

As seen in Section 7, an advantage of using alternative dual frames in $\Sigma\Delta$ quantization is that they can be chosen so that the boundary terms in (6.5) are of small order. For example, Figure 2 shows that the boundary terms for the alternative dual frames $F_N(4)$ and $F_N(11)$ are of order $1/N^4$ and $1/N^{11}$, respectively.

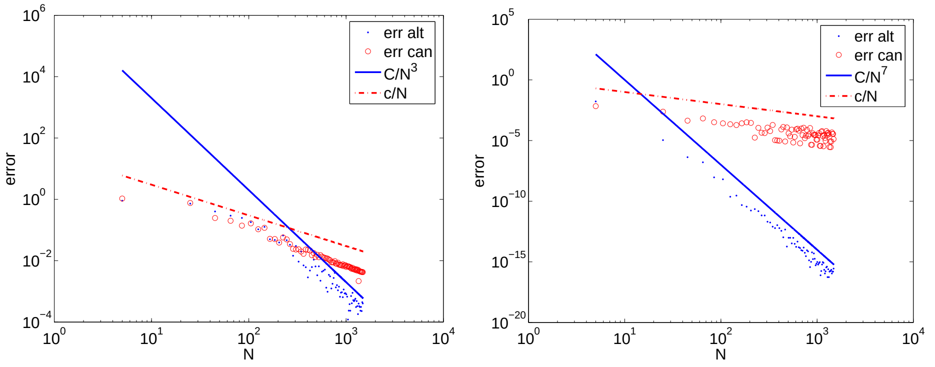


(a)

(b)

Figure 2. Parts (a) and (b) show log-log plots of the norm of the boundary terms in (6.5) for the frames $F_N(3)$ and $F_N(11)$, respectively. For comparison, boundary terms for the canonical dual frame of E_N are also plotted.

To illustrate Theorem 8.4, let $x = (1/\pi, \sqrt{3/17})$ and suppose that the frame coefficients of x with respect to the frame E_N are quantized with an rth order $\Sigma\Delta$ scheme. We compare the approximation error when the canonical dual frame and the alternative dual frames $F_N(r)$ from Theorem 8.4 are used to linearly reconstruct $\widetilde{x}$. Figure 3(a) is a log-log plot of the approximation error $||x-\widetilde{x}||$ as a function of N, if one uses the stable 3rd order $\Sigma\Delta$ scheme of [14] and reconstructs with the alternative dual frame $F_N(3)$. As predicted, the alternative reconstruction yields error of order $1/N^3$. Figure 3(b) is a log-log plot of the approximation error $||x-\widetilde{x}||$ as a function of N, of one uses the stable 7th order $\Sigma\Delta$ scheme from Example 5.2 with $\delta = 0.0039$, and reconstructs with the alternative dual frame $F_N(7)$. As predicted, the alternative reconstruction yields error of order $1/N^7$.



(a) (b)

FIGURE 3. The frame expansions of $x = (1/\pi, \sqrt{3/17})$ with respect to E_N are quantized using: (a) the 3rd order scheme of [14], (b) the 7th order $\Sigma\Delta$ scheme from Example 5.2 with $\delta = 0.0039$. Parts (a) and (b) show log-log plots of the approximation error $\|x-\widetilde{x}\|$ as a function of N, when $\widetilde{x}$ is reconstructed using the canonical dual frame ('err can') and the alternative dual frames $F_N(3)$ and $F_N(7)$, respectively ('err alt').

8.2. **Harmonic frames for $\mathbb{R}^d$.** In this section, we generalize the alternative dual construction of the previous section to harmonic frames for $\mathbb{R}^d$. Let $H_N^d = \{h_n^N\}_{n=1}^N$, be the harmonic frame for $\mathbb{R}^d$ defined in Example 2.2 by (2.4) and (2.5). For each $r \geq 3$, we construct an alternative dual frame to H_N^d which satisfies Property 7.1. The alternative dual frame F_N^d will be of the form

$$F_N^d = F_N^d(r) = \{f_n^N\}_{n=1}^N, \quad f_n^N = \frac{1}{N}(dh_n^N + g_n^N), \tag{8.7}$$

where the definition of $\{g_n^N\}_{n=1}^N \subset \mathbb{R}^d$ depends on whether d is even or odd.

Harmonic frames in even dimension d. Suppose that the dimension $d = 2d_0$ is even. Let $u[j]$ denote the jth component of the vector u and define F_N^d by (8.7) where

$$g_n^N[2s-1] = a_{s,0} + \sum_{\ell=1}^{k} a_{s,\ell}\cos\left(2\pi(d_0+\ell)n/N\right), \quad s = 1, 2. \ldots, d_0, \tag{8.8}$$

$$g_n^N[2s] = \sum_{\ell=1}^{k} b_{s,\ell} \sin\left(2\pi(d_0+\ell)n/N\right), \quad s = 1, 2. \dots, d_0, \tag{8.9}$$

and $k = k(r) > 0$, $\{a_{s,\ell}\}_{s=1,l=0}^{d_0,k}$, $\{b_{s,\ell}\}_{s=1,l=1}^{d_0,k}$ are constants to be defined later. Next, for $s = 1, 2, \dots, d_0$, we set

$$\psi_{2s-1}(t) \;=\; \sqrt{2d}\cos(2s\pi t) + a_{s,0} + \sum_{\ell=1}^{k} a_{s,\ell}\cos((d_0+\ell)2\pi t), \tag{8.10}$$

$$\psi_{2s}(t) \;=\; \sqrt{2d}\sin(2s\pi t) + \sum_{\ell=1}^{k} b_{s,\ell}\sin((d_0+\ell)2\pi t). \tag{8.11}$$

With this notation, we have $f_n^N = \frac{1}{N}\left[\psi_1(n/N), \psi_2(n/N), \dots, \psi_d(n/N)\right]^T$. As in Lemma 8.1, one has the following dual frame lemma for $F_N^d = \{f_n^N\}_{n=1}^N$.

Lemma 8.5. *If $1 \le k < N - d_0$ then F_N^d, defined by (8.7), (8.8) and (8.9), is a dual frame to H_N^d for every choice of $\{a_{s,\ell}\}_{s=1,\ell=0}^{d_0,k}$, $\{b_{s,\ell}\}_{s=1,\ell=1}^{d_0,k} \subset \mathbb{R}$.*

Proof. Similar to Lemma 8.1, this result follows from discrete orthogonality relations for cosine and sine. We omit the details. □

Given $r > 0$, our goal is to choose $\{a_{s,\ell}\}_{s=1,\ell=0}^{d_0,k}, \{b_{s,\ell}\}_{s=1,\ell=1}^{d_0,k}$, so that the functions $\{\psi_s\}_{s=1}^{d}$ satisfy the requirements of Property 7.1. As in Section 8.1, this is achieved by ensuring that lower order terms in the power series expansion of each ψ_s vanish. Note that that the ψ_{2s-1} are even functions, the ψ_{2s} are odd functions, and all are 1-periodic.

The power series expansions about $t = 0$ for the ψ_s, $1 \le s \le d_0$, are given by

$$\psi_{2s-1}(t) = \sum_{n=0}^{\infty} \beta_{2s-1,2n}\ t^{2n} \quad \text{and} \quad \psi_{2s}(t) = \sum_{n=0}^{\infty} \beta_{2s,2n+1}\ t^{2n+1}, \tag{8.12}$$

where

$$\beta_{2s-1,0} = \sqrt{2d} + \sum_{\ell=0}^{k} a_{s,\ell},$$

$$\forall n \ge 1,\ \beta_{2s-1,2n} = \frac{(-1)^n(2\pi)^{2n}}{(2n)!}\left(s^{2n}\sqrt{2d} + \sum_{\ell=1}^{k} a_{s,\ell}(d_0+\ell)^{2n}\right),$$

$$\forall n \ge 0, \quad \beta_{2s,2n+1} = \frac{(-1)^n(2\pi)^{2n+1}}{(2n+1)!}\left(s^{2n+1}\sqrt{2d} + \sum_{\ell=1}^{k} b_{s,\ell}(d_0+\ell)^{2n+1}\right).$$

Let $\mathbf{V}_{d_0,k}$ be the following $k \times k$ Vandermonde matrix

$$\mathbf{V}_{d_0,k} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ (d_0+1)^2 & (d_0+2)^2 & \cdots & (d_0+k)^2 \\ (d_0+1)^4 & (d_0+2)^4 & \cdots & (d_0+k)^4 \\ \vdots & \vdots & \vdots & \vdots \\ (d_0+1)^{2k-2} & (d_0+2)^{2k-2} & \cdots & (d_0+k)^{2k-2} \end{bmatrix}, \tag{8.13}$$

and let $\mathbf{M}_{d_0,k}$ be the $k \times k$ diagonal matrix with $[d_0+1, d_0+2, \cdots, d_0+k]$ on its main diagonal.

Lemma 8.6. *Fix $k > 0$ and define $\{a_{s,\ell}\}_{s=1,\ell=0}^{d_0,k}$ and $\{b_{s,\ell}\}_{s=1,\ell=1}^{d_0,k}$ as follows. Let*

$$\mathbf{a}_s = [a_{s,1}, \cdots, a_{s,k}]^T \quad \text{and} \quad \mathbf{b}_s = [b_{s,1}, \cdots, b_{s,k}]^T$$

be chosen as the unique solutions to

$$V_{d_0,k} M_{d_0,k}^2 \mathbf{a}_s = -\sqrt{2d}\ [s^2, s^4, s^6, \cdots, s^{2k}]^T,$$

$$V_{d_0,k} M_{d_0,k} \mathbf{b}_s = -\sqrt{2d}\ [s, s^3, s^5, \cdots, s^{2k-1}]^T,$$

and also let

$$a_{s,0} = -\sqrt{2d} - \sum_{\ell=1}^{k} a_{s,\ell}.$$

Note that $b_{s,\ell} = (\frac{d_0+\ell}{s}) a_{s,\ell}$, for $1 \leq \ell \leq k$. If the functions $\{\psi_s\}_{s=1}^d$ are defined as in (8.10) and (8.11) with the above choice of constants, then, for each $1 \leq s \leq d_0$,

$$\forall\ 0 \leq j \leq 2k, \quad \psi_s^{(j)}(0) = 0.$$

Equivalently, for each $1 \leq s \leq d_0$, the power series coefficients in (8.12) satisfy $\beta_{2s-1,2n} = \beta_{2s,2n+1} = 0$ for all $0 \leq n \leq k-1$.

In particular, if one is given $r \geq 3$, takes $k = \lceil r/2 \rceil - 1$, and lets F_N^d be the alternative dual frame for H_N^d given by (8.7),(8.8) and (8.9), with $\{a_{s,\ell}\}, \{b_{s,\ell}\}$ as above, then H_N^d, F_N^d satisfy Property 7.1.

The choice of $\{a_{s,\ell}\}, \{b_{s,\ell}\}$ in Lemma 8.6 was made to ensure that appropriately many lower order power series coefficients of ψ_s are zero. Similar to the results for roots-of-unity frames in Section 8.1, Property 7.1 follows here since the ψ_s are entire functions whose restrictions to $\mathbb{R}$ are 1-periodic.

Harmonic frames in odd dimension d. When the dimension $d = 2d_0 + 1$ is odd, we modify the previous construction as follows. For $s = 1, \cdots, d_0$ define

$$\psi_{2s-1}(t) = \sqrt{2d} \cos(2s\pi t) + \sum_{\ell=1}^{k+1} a_{s,\ell} \cos((d_0 + \ell)2\pi t), \tag{8.14}$$

$$\psi_{2s}(t) = \sqrt{2d} \sin(2s\pi t) + \sum_{\ell=1}^{k} b_{s,\ell} \sin((d_0 + \ell)2\pi t), \tag{8.15}$$

$$\psi_0(t) = \sqrt{d} + \sum_{\ell=1}^{k+1} a_{0,\ell} \cos((d_0 + \ell)2\pi t). \tag{8.16}$$

where $k = k(r)$ and $\{a_{s,\ell}\}_{s=0,\ell=1}^{d_0,k+1}$, $\{b_{s,\ell}\}_{s=1,\ell=1}^{d_0,k}$ are constants to be defined later. Note that if $d > 2$ is an odd integer then $F_N^d = \{f_n^N\}_{n=1}^N$ defined by

$$f_n^N = \frac{1}{N} [\psi_0(n/N), \psi_1(n/N), \ldots, \psi_{d-1}(n/N)]^T \tag{8.17}$$

is an alternative dual frame for H_N^d (in the same manner as in Lemma 8.5). Let the matrices $\mathbf{V}_{d_0,k}$ and $\mathbf{M}_{d_0,k}$ be defined as in (8.13).

Lemma 8.7. *Fix $k > 0$ and define $\{a_{s,\ell}\}_{s=0,\ell=1}^{d_0,k+1}$ and $\{b_{s,\ell}\}_{s=1,\ell=1}^{d_0,k}$ as follows. Let*

$$\forall\ 0 \leq s \leq d_0, \quad \mathbf{a}_s = [a_{s,1}, \cdots, a_{s,k+1}]^T \ \text{and} \quad \forall\ 1 \leq s \leq d_0, \quad \mathbf{b}_s = [b_{s,1}, \cdots, b_{s,k}]^T$$

be chosen as the unique solutions to

$$\mathbf{V}_{d_0,k+1}\mathbf{a}_0 = -\sqrt{d}\ [1, 0, 0, \cdots, 0]^T,$$
$$\forall\ 1 \leq s \leq d_0, \quad \mathbf{V}_{d_0,k+1}\mathbf{a}_s = -\sqrt{2d}\ [1, s^2, s^4, s^6, \cdots, s^{2k}]^T,$$
$$\forall\ 1 \leq s \leq d_0, \quad \mathbf{V}_{d_0,k}\mathbf{M}_{d_0,k}\mathbf{b}_s = -\sqrt{2d}\ [s, s^3, s^5, \cdots, s^{2k-1}]^T.$$

If the functions $\{\psi_s\}_{s=0}^{d-1}$ are defined as in (8.14), (8.15) and (8.16) with the above choice of constants, then, for each $0 \leq s \leq d-1$,

$$\forall\ 0 \leq j \leq 2k, \qquad \psi_s^{(j)}(0) = 0.$$

In particular, if one is given $r \geq 3$, takes $k = \lceil r/2 \rceil - 1$, and lets F_N^d be the alternative dual frame for H_N^d given by (8.14), (8.15), (8.16) and (8.17) with $\{a_{s,\ell}\}, \{b_{s,\ell}\}$ as above, then H_N^d, F_N^d satisfy Property 7.1.

Similar to Lemma 8.6, writing out the power series expansions for ψ_s shows that the choice of $\{a_{s,\ell}\}$ and $\{b_{s,\ell}\}$ in Lemma 8.7 is made to ensure that appropriately many lower order power series coefficients of the ψ_s are zero. As in Lemma 8.6, Property 7.1 follows from power series properties of the ψ_s since each ψ_s is an entire function whose restriction to $\mathbb{R}$ is 1-periodic.

The following theorem provides error estimates when the alternative dual frames from Lemmas 8.6 and 8.7 are used to linearly reconstruct $\Sigma\Delta$ quantized harmonic frame coefficients.

Theorem 8.8. *Let $r \geq 3$ be a positive integer and let $H_N^d = \{h_n^N\}_{n=1}^N \subset \mathbb{R}^d$ be the harmonic frame for $\mathbb{R}^d$ defined by (2.4) and (2.5). Take $k = \lceil r/2 \rceil - 1$ and define the dual frame $F_N^d(r) = \{f_n^N\}_{n=1}^N$ through Lemmas 8.6 or 8.7 depending on whether d is even or odd.*

Suppose we are given a stable rth order $\Sigma\Delta$ scheme (5.6), and let $C_1, C_2 > 0$ be the associated stability constants as in (5.7). For $x \in \mathbb{R}^d$, $\|x\| < C_1$, let $\widetilde{x}_N = \sum_{n=1}^N q_n(x) f_n^N(r)$ where $q_n(x)$ is produced via the rth order $\Sigma\Delta$ scheme by quantizing $\langle x, h_n^N \rangle$. Then

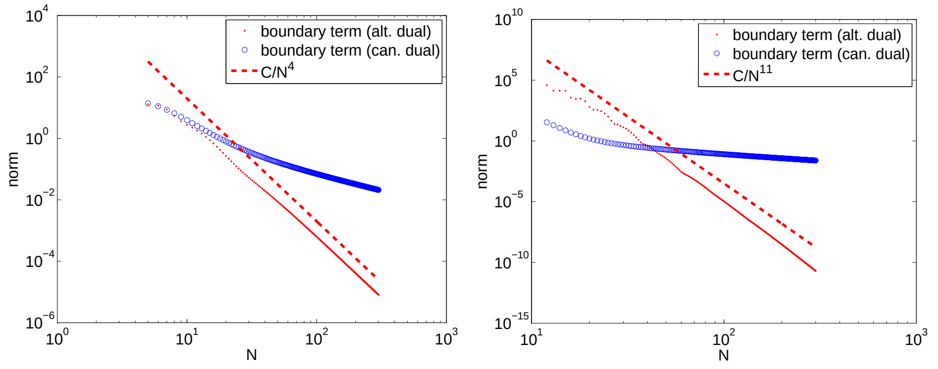
$$\|x - \widetilde{x}_N\| \leq \frac{C_{\Sigma\Delta}^{HF}(r,d)}{N^r},$$

where $C_{\Sigma\Delta}^{HF}(r,d)$ is the corresponding constant from Theorem 7.2.

Proof. Lemmas 8.6 and 8.7 show that Property 7.1 is satisfied by $F_N^d(r)$. Thus the result follows from Theorem 7.2. □

Numerical experiments. Figure 4(a) and (b), plot the norm of the boundary term in (5.10) as a function of N, for frames $F_N^6(4)$ and $F_N^7(11)$, respectively.

To illustrate Theorem 8.8 in dimension 4, let $x = (1/2)(1/\pi, \sqrt{3/17}, -1/2, e^{-1/2})$ and suppose that the frame coefficients of x with respect to the frame H_N^4 are quantized with the 3rd order $\Sigma\Delta$ scheme from [14]. We compare the approximation error when the canonical dual frame and the alternative dual frame $F^4(3)$ are used to reconstruct the quantized frame coefficients. Part (a) of Figure 5 shows a log-log plot of both corresponding approximation errors $||x - \widetilde{x}_N||$ as a function N. As predicted, the error is of order $1/N^3$.



(a)

(b)

FIGURE 4. Parts (a) and (b) show log-log plots of the norm of the boundary terms in (6.5) for the frames $F^6(4)$ and $F^7(11)$, respectively. For comparison, boundary terms for the canonical dual frame of H_N^7 are also shown.

To illustrate Theorem 8.8 in dimension 5, let

$$y = (1/3)(1/\pi, \sqrt{3/17}, -1/2, e^{-1/2}, \sqrt{1/2})$$

and suppose that the frame coefficients of y with respect to H_N^5 are quantized using the 7th order $\Sigma\Delta$ scheme from Example 5.2 with $\delta = 2^{-8} \approx 0.0039$ which guarantees that the scheme is stable. Part (b) of Figure 5 shows a log-log plot of the corresponding approximation errors $||y - \widetilde{y}_N||$ as a function N. As predicted, the error is of order $1/N^7$.

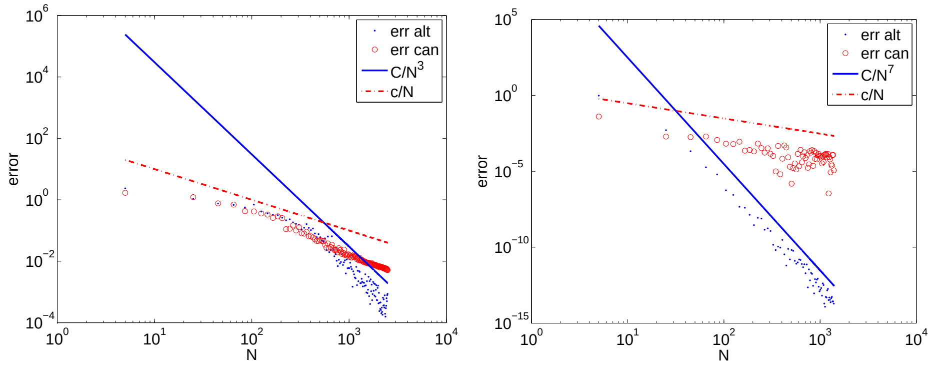
ACKNOWLEDGMENTS

The authors wish to thank John Benedetto, Ingrid Daubechies, Sinan Günturk, Nguyen Thao, and Yang Wang for valuable discussions related to the material.

A portion of this work was completed while the authors were at the Erwin Schrödinger Institute (ESI) during the Special Semester on "Modern methods of time-frequency analysis." Part of this work was presented in May 2005 at the associated International Conference on "Modern methods of time-frequency analysis" in Strobl, Austria. The authors thank ESI for its hospitality, and are especially grateful to Hans Feichtinger for having helped to arrange their visits.

9. APPENDIX: UNIFORM DISTRIBUTION

This appendix provides some necessary background on uniform distribution which is used in the previous sections. The discussion follows the references [22, 25], except that we choose to work with the interval $[-1/2, 1/2)$, instead of $[0, 1)$. We identify $[-1/2, 1/2)$ with the torus $\mathbb{T}$. Given $x \in \mathbb{R}$, we define $[[x]]$ to be the unique number in the interval $[-1/2, 1/2)$ such that $x \equiv [[x]]$ modulo 1, i.e, $x - [[x]] \in \mathbb{Z}$.



(a) (b)

Figure 5. The approximation error for two $\Sigma\Delta$ schemes. The frame expansions of (a) $x = (1/2)(1/\pi, \sqrt{3/17}, -1/2, e^{-1/2})$ with respect to H_N^4, and (b) $y = (1/3)(1/\pi, \sqrt{3/17}, -1/2, e^{-1/2}, \sqrt{1/2})$ with respect to H_N^5 are quantized using (a) the 3rd-order scheme of [14], and (b) the 7th-order $\Sigma\Delta$ scheme, as described in Example 5.2 with $\delta = 2^{-8} \approx 0.0039$. In both (a) and (b), we show the approximation error $\|x - \widetilde{x}\|$ where $\widetilde{x}$ is obtained using the canonical dual ('err can'), along with the approximation error that is obtained when the alternative duals $F_N^4(3)$ and $F_N^5(7)$ were used, respectively ('err alt').

The sequence $\{u_n\}_{n=1}^{\infty} \subset \mathbb{R}$ is *uniformly distributed modulo 1* if

$$\forall \ \text{interval} \ \ I \subseteq [-1/2, 1/2), \qquad \lim_{N\to\infty} \frac{\operatorname{card}\{1 \le n \le N : [[u_n]] \in I\}}{N} = |I|.$$

The classical theorem of Weyl gives useful equivalent conditions for a sequence to be uniformly distributed, [22, 25].

Theorem 9.1 (Weyl)**.** *Let $\{u_n\}_{n=1}^{\infty} \subset \mathbb{R}$. The following are equivalent:*

(1) *$\{u_n\}_{n=1}^{\infty}$ is uniformly distributed modulo 1,*
(2) *For every Riemann integrable function f on $\mathbb{T}$,*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} f(u_n) = \int_{\mathbb{T}} f(x)dx.$$

Polynomials with irrational coefficients provide basic examples of uniformly distributed sequences, e.g., see Theorem 3.2 in [22].

Lemma 9.2. *If $P_k(x) = c_k x^k + c_{k-1}x^{k-1} + \cdots + c_1 x + c_0$ is a polynomial of degree $k \ge 1$ for which at least one $c_j \in \mathbb{R}\backslash\mathbb{Q}$, with $j \ge 1$, then $\{P_k(n)\}_{n=1}^{\infty}$ is uniformly distributed modulo 1.*

The next lemma shows that certain perturbations of uniformly distributed sequences remain uniformly distributed, e.g., see page 23 of [22].

Lemma 9.3. *If $\{u_n\}_{n=1}^\infty \subset \mathbb{R}$, satisfies $u_n = v_n + w_n$, where $\{v_n\}_{n=1}^\infty$ is uniformly distributed modulo 1 and*

$$\lim_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} |w_n| = 0,$$

then $\{u_n\}_{n=1}^\infty \subset \mathbb{T}$ is uniformly distributed modulo 1.

The following lemma shows that the sequence $\{B_N\}_{N=3}^\infty$ resulting from the boundary terms in the proof of Theorem 6.2 is indeed uniformly distributed.

Lemma 9.4. *For almost every $(a,b) \in \mathbb{R}^2$, the sequence $\{B_N\}_{N=3}^\infty$ defined by*

$$B_N = \frac{-aN}{2} + \frac{bN}{2\tan(\pi/N)},$$

is uniformly distributed modulo 1.

Proof. Note that $B_N = V_N + W_N$, where

$$V_N = \left(\frac{bN^2}{2\pi} - \frac{aN}{2} - \frac{b\pi}{6}\right) + \frac{b\pi^3}{18(N^2+\pi^2/3)},$$

and

$$W_N = \frac{bN}{2}\left(\frac{1}{\tan(\frac{\pi}{N})} - \frac{1}{\frac{\pi}{N} + \frac{\pi^3}{3N^3}}\right).$$

A direct calculation shows that $|W_N| \lesssim 1/N^2$, so that $\lim_{M\to\infty} \frac{1}{M}\sum_{N=1}^{M} |W_N| = 0$. Lemma 9.2 and Lemma 9.3 show that $\{V_N\}_{N=1}^\infty$ is uniformly distributed modulo 1. Applying Lemma 9.3 a second time shows that $\{B_N\}_{N=1}^\infty$ is uniformly distributed modulo 1 and completes the proof. □

References

1. A. Aldroubi, Portraits of Frames, *Proceedings of AMS,* **123** (1995), 357–385.
2. J.J. Benedetto, M. Fickus, Finite normalized tight frames, *Advances in Computational Mathematics,* **18** (2003), no. 2-4, 357–385.
3. J.J. Benedetto, A.M. Powell, Ö. Yılmaz, Sigma-Delta quantization and finite frames, *Proceedings of the International Conference on Acoustics, Speech and Signal Processing,* vol. 3, Montreal, QC, Canada, May 2004, pp. 937–940.
4. J.J. Benedetto, A.M. Powell, Ö. Yılmaz, Sigma-Delta ($\Sigma\Delta$) quantization and finite frames, *IEEE Transactions on Information Theory,* **52** (2006), no.5, 1990–2005.
5. J.J. Benedetto, A.M. Powell, Ö. Yılmaz, Second order Sigma-Delta ($\Sigma\Delta$) quantization of finite frame expansions, *Applied and Computational Harmonic Analysis,* **20** (2006), no.1, 126–148.
6. W. Bennett, Spectra of quantized signals, *Bell System Technical Journal,* **27** (1949), 446–472.
7. J. Blum, M.C. Lammers, A.M. Powell, Ö. Yılmaz, Sobolev duals in frame theory and Sigma-Delta quantization, *Preprint.*
8. B.G. Bodmann, V.I. Paulsen, Frame paths and error bounds for sigma-delta quantization *Applied and Computational Harmonic Analysis,* **22** (2007), no.2, 176–197.
9. B.G. Bodmann, V.I. Paulsen, S. Abdulbaki, Smooth frame path termination for higher order sigma-delta quantization *Journal of Fourier Analysis and Applications,* **13** (2007), 285–307.
10. H. Bölcskei, F. Hlawatsch, Noise reduction in oversampled filter banks using predictive quantization, *IEEE Transactions on Information Theory,* **47** (2001), no.1, 155–172.
11. P. Casazza, The art of frame theory, *Taiwanese Journal of Mathematics,* **4** (2000), no. 2, 129–210.
12. O. Christensen, Y. Eldar, Characterization of oblique dual frame pairs, *EURASIP J. Applied Signal Proc.,* 2006, Frames and overcomplete representations in signal processing, communications, and information theory, Art. ID 92674, 11 pages.
13. Z. Cvetković, Resilience properties of redundant expansions under additive noise quantization, *IEEE Transactions on Information Theory,* **49** (2003), no. 3, 644–656.

14. I. Daubechies, R. DeVore, Approximating a bandlimited function from very coarsely quantized data: a family of stable sigma-delta modulators of arbitrary order, *Annals of Mathematics (2),* **158** (2003), no. 2, 679–710.
15. V. Goyal, J. Kovačević, J. Kelner, Quantized frame expansions with erasures, *Applied and Computational Harmonic Analysis,* **10** (2001), 203–233.
16. V. Goyal, M. Vetterli, N. Thao, Quantized overcomplete expansions in $\mathbb{R}^n$, *IEEE Transactions on Information Theory,* **44** (1998), no.1, 16–31.
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, One-bit sigma-delta quantization with exponential accuracy, *Communications on Pure and Applied Mathematics,* **56** (2003), no.11, 1608–1630.
19. R. Gray, Quantization noise spectra, *IEEE Transactions on Information Theory,* **36** (1990), no.6, 1220–1244.
20. N. He, F. Kuhlmann, and A. Buzo, Multi-loop sigma-delta quantization, *IEEE Transactions on Information Theory,* **38** (1992), no.3, 1015–1028.
21. D. Jimenez, L. Wang, Y. Wang, White noise hypothesis for uniform quantization errors, *SIAM J. Math. Anal.* **38** (2007) no.6, 2042-2056.
22. L. Kuipers, H. Niederrieter, *Uniform distribution of sequences,* Pure and Applied Mathematics, Wiley-Interscience, New York-London-Sydney, 1974.
23. S. Li, On general frame decompositions, *Numerical Functional Analysis and Optimization,* **16** (1995), no. 9 & 10, 1181–1191.
24. S. Li, H. Ogawa, Pseudoframes for subspaces with applications, *Journal of Fourier Analysis and Applications,* **10** (2004), no.4, 409–431.
25. H.L. Montgomery, *Ten Lectures on the Interface Between Analytic Number Theory and Harmonic Analysis,* CBMS Regional Conference Series in Mathematics, 84, American Mathematical Society, Providence, RI, 1994.
26. D. Marco, D. Neuhoff, The validity of the additive noise model for uniform scalar quantizers, *IEEE Transactions on Information Theory,* **51** (2005), no.5, 1739–1755.
27. S.C. Pinault, P.V. Lopresti, On the behavior of the double-loop sigma-delta modulator, *IEEE Transactions on Circuits and Systems II,* **40** (1993), no.8, 467–479.
28. A. Sripad, D. Snyder, A necessary and sufficient condition for quantization errors to be uniform and white, *IEEE Transactions on Acoustics, Speech and Signal Processing,* **25** (1977), no. 5, 442–448.
29. N. Thao, M. Vetterli, Deterministic analysis of oversampled A/D conversion and decoding improvement based on consistent estimates, *IEEE Transactions on Information Theory,* **42** (1994), no. 3, 519–531.
30. H. Viswanathan, R. Zamir, On the whiteness of high-resolution quantization errors, *IEEE Transactions on Information Theory,* **47** (2001), no. 5, 2029–2038.
31. 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.
32. G. Zimmermann, Normalized tight frames in finite dimensions, in: K. Jetter, W. Haussmann, M. Reimer (Eds.), Recent Progress in Multivariate Approximation, Birkhäuser, 2001.

Department of Mathematics, University of North Carolina at Wilmington, Wilmington, NC 28403, USA
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`