Sigma-Delta ($\Sigma\Delta$) quantization and finite frames

John J. Benedetto, Alexander M. Powell, and Özgür Yılmaz

***Abstract*— The K-level Sigma-Delta ($\Sigma\Delta$) scheme with step size δ is introduced as a technique for quantizing finite frame expansions for $\mathbb{R}^d$. Error estimates for various quantized frame expansions are derived, and, in particular, it is shown that $\Sigma\Delta$ quantization of a normalized finite frame expansion in $\mathbb{R}^d$ achieves approximation error $||x - \widetilde{x}|| \leq \frac{\delta d}{2N}(\sigma(F,p)+2)$, where N is the frame size, and the frame variation $\sigma(F,p)$ is a quantity which reflects the dependence of the $\Sigma\Delta$ scheme on the frame. Here $||\cdot||$ is the d-dimensional Euclidean 2-norm. Lower bounds and refined upper bounds are derived for certain specific cases. As a direct consequence of these error bounds one is able to bound the mean squared error (MSE) by an order of $1/N^2$. When dealing with sufficiently redundant frame expansions, this represents a significant improvement over classical PCM quantization, which only has MSE of order $1/N$ under certain nonrigorous statistical assumptions. $\Sigma\Delta$ also achieves the optimal MSE order for PCM with consistent reconstruction.**

***Index Terms*— Sigma-Delta quantization, finite frames.**

I. Introduction

In signal processing, one of the primary goals is to obtain a digital representation of the signal of interest that is suitable for storage, transmission, and recovery. In general, the first step towards this objective is finding an atomic decomposition of the signal. More precisely, one expands a given signal x over an at most countable dictionary $\{e_n\}_{n\in\Lambda}$ such that

$$x = \sum_{n\in\Lambda} c_n e_n, \tag{1}$$

where c_n are real or complex numbers. Such an expansion is said to be *redundant* if the choice of c_n in (1) is not unique.

Although (1) is a discrete representation, it is certainly not "digital" since the coefficient sequence $\{c_n\}_{n\in\Lambda}$ is real or complex valued. Therefore, a second step is needed to reduce the continuous range of this sequence to a discrete, and preferably finite, set. This second step is called *quantization*. A *quantizer* maps each expansion (1) to an element of

$$\Gamma_{\mathcal{A}} = \{\sum_{n\in\Lambda} q_n e_n : q_n \in \mathcal{A}\},$$

where the *quantization alphabet* $\mathcal{A}$ is a given discrete, and preferably finite, set. The performance of a quantizer is reflected in the approximation error $\|x - \widetilde{x}\|$, where $\|\cdot\|$ is a suitable norm, and

$$\widetilde{x} = \sum_{n\in\Lambda} q_n e_n \tag{2}$$

is the quantized expansion.

The process of reconstructing $\widetilde{x}$ in (2) from the quantized coefficients, $q_n, n \in \Lambda$, is called *linear reconstruction*. More general approaches to quantization, such as consistent reconstruction, e.g., [1], [2], use nonlinear reconstruction, but unless otherwise mentioned, we shall focus on quantization using linear reconstruction, as in (2).

A simple example of quantization, for a given expansion (1), is to choose q_n to be the closest point in the alphabet $\mathcal{A}$ to c_n. Quantizers defined this way are usually called *pulse code modulation* (PCM) algorithms. If $\{e_n\}_{n\in\Lambda}$ is an orthonormal basis for a Hilbert space H, then PCM algorithms provide the optimal quantizers in that they minimize $\|x - \widetilde{x}\|$ for every x in H, where $\|\cdot\|$ is the Hilbert space norm. On the other hand, PCM can perform poorly if the set $\{e_n\}_{n\in\Lambda}$ is redundant. We shall discuss this in detail in Section II-B.

In this paper we shall examine the quantization of redundant real finite atomic decompositions (1) for $\mathbb{R}^d$. The signal, x, and dictionary elements, $e_n, n \in \Lambda$, are elements of $\mathbb{R}^d$, the index set Λ is finite, and the coefficients, $c_n, n \in \Lambda$, are real numbers.

A. Frames, redundancy, and robustness

In various applications it is convenient to assume that the signals of interest are elements of a Hilbert space H, e.g., $H = L^2(\mathbb{R}^d)$, or $H = \mathbb{R}^d$, or H is a space of bandlimited functions. In this case, one can consider more structured dictionaries, such as frames. Frames are a special type of dictionary which can be used to give stable redundant decompositions (1). Redundant frames are used in signal processsing because they yield representations that are robust under

- additive noise [3] (in the setting of Gabor and wavelet frames for $L^2(\mathbb{R})$), [4] (in the setting of oversampled bandlimited functions), and [5] (in the setting of tight Gabor frames),
- quantization [6], [7], [8] (in the setting of oversampled bandlimited functions), [9] (in the setting of tight Gabor frames), and [2] (in the setting of finite frames for $\mathbb{R}^d$), and
- partial data loss [10], [11] (in the setting of finite frames for $\mathbb{R}^d$).

Although redundant frame expansions use a larger than necessary bit-budget to represent a signal (and hence are not preferred for storage purposes where data compression is the main goal), the robustness properties listed above make them ideal for applications where data is to be transferred over noisy channels, or to be quantized very coarsely. In particular, in the case of Sigma-Delta ($\Sigma\Delta$) modulation of *oversampled bandlimited functions* x, one has very good reconstruction

This work was supported by NSF DMS Grant 0139759 and ONR Grant N000140210398.

J. J. Benedetto is with the Department of Mathematics, University of Maryland, College Park, MD 20742.

A. M. Powell is with the Program in Applied and Computational Mathematics, Princeton University, Princeton, NJ 08544.

Ö. Yılmaz is with the Department of Mathematics, University of British Columbia, Vancouver, B.C. Canada V6T 1Z2.

using only 1-bit quantized values of the frame coefficients [6], [12], [13]. Moreover, the resulting approximation $\widetilde{x}$ is robust under quantizer imperfections as well as bit-flips [6], [7], [8].

Another example where redundant frames are used, this time to ensure robust transmission, can be found in the works of Goyal, Kovačević, Kelner, and Vetterli [10], [14], cf., [15]. They propose using *finite tight frames* for $\mathbb{R}^d$ to transmit data over erasure channels; these are channels over which transmission errors can be modeled in terms of the loss (erasure) of certain packets of data. They show that the redundancy of these frames can be used to "mitigate the effect of the losses in packet-based communication systems"[16], cf., [17]. Further, the use of finite frames has been proposed for generalized multiple description coding [18], [11], for multiple-antenna code design [19], and for solving modified quantum detection problems [20]. Thus, finite frames for $\mathbb{R}^d$ or $\mathbb{C}^d$ are emerging as a natural mathematical model and tool for many applications.

B. Redundancy and quantization

A key property of frames is that greater frame redundancy translates into more robust frame expansions. For example, given a normalized tight frame for $\mathbb{R}^d$ with frame bound A, any transmission error that is caused by the erasure of e coefficients can be corrected as long as $e < A$ [10]. In other words, increasing the frame bound, i.e., the redundancy of the frame, makes the representation more robust with respect to erasures. However, increasing redundancy also increases the number of coefficients to be transmitted. If one has a fixed bit-budget, a consequence is that one has fewer bits to spend for each coefficient and hence needs to be more resourceful in *how* one allocates the available bits.

- When dealing with PCM, using linear reconstruction, for finite frame expansions in $\mathbb{R}^d$, a longstanding analysis with certain assumptions on quantization "noise" bounds the resulting mean square approximation error by $C_1\delta^2/A$ where C_1 is a constant, A is the frame bound, and δ is the quantizer step size [21], see Section II-B.
- On the other hand, for 1-bit first order $\Sigma\Delta$ quantization of oversampled *bandlimited* functions, the approximation error is bounded by C_2/A pointwise [6], and the mean square approximation error is bounded by C_3/A^3 [12], [13].

Thus, if we momentarily "compare apples with oranges", we see that $\Sigma\Delta$ quantization algorithms for bandlimited functions utilize the redundancy of the expansion more efficiently than PCM algorithms for $\mathbb{R}^d$.

C. Overview of the paper and main results

Section II discusses necessary background material. In particular, Section II-A gives basic definitions and theorems from frame theory, and Section II-B presents basic error estimates for PCM quantization of finite frame expansions for $\mathbb{R}^d$.

In Section III, we introduce the K-level $\Sigma\Delta$ scheme with step size δ as a new technique for quantizing normalized finite frame expansions. A main theme of this paper is to show that the $\Sigma\Delta$ scheme outperforms linearly reconstructed PCM quantization of finite frame expansions. In Section III-A, we introduce the notion of *frame variation*, $\sigma(F,p)$, as a quantity which reflects the dependence of the $\Sigma\Delta$ scheme's performance on properties of the frame. Section III-B uses the frame variation, $\sigma(F,p)$, to derive basic approximation error estimates for the $\Sigma\Delta$ scheme. For example, we prove that if F is a normalized tight frame for $\mathbb{R}^d$ of cardinality $N \geq d$, then the K-level $\Sigma\Delta$ scheme with quantization step size δ gives approximation error

$$||x - \widetilde{x}|| \leq \frac{\delta d}{2N}(\sigma(F,p)+2),$$

where $||\cdot||$ is the d-dimensional Euclidean 2-norm.

Section IV is devoted primarily to examples. We give examples of infinite families of frames with uniformly bounded frame variation. We compare the error bounds of Section III with the numerically observed error for these families of frames. Since $\Sigma\Delta$ schemes are iterative, they require one to choose a *quantization order*, p, in which frame coefficients are given as input to the scheme. We present a numerical example which shows the importance of carefully choosing the quantization order to ensure good approximations.

In Section V, we derive lower bounds and refined upper bounds for the $\Sigma\Delta$ scheme. This partially explains properties of the approximation error which are experimentally observed in Section IV. In particular, we show that in certain situations, if the frame size N is even, then one has the improved approximation error bound $||x - \widetilde{x}|| \leq C_1/N^{5/4}$ for an x-dependent constant C_1. On the other hand, if N is odd we prove the lower bound $C_2/N \leq ||x - \widetilde{x}||$ for an x-dependent constant C_2. In both cases $||\cdot||$ is the Euclidean norm.

In Section VI, we compare the mean square (approximation) error (MSE) for the $\Sigma\Delta$ scheme with PCM using linear reconstruction. If we have an harmonic frame for $\mathbb{R}^d$ of cardinality $N \geq d$, then we show that the MSE for the $\Sigma\Delta$ scheme is bounded by an order of $1/N^2$, whereas the standard MSE estimates for PCM are only of order $1/N$. Thus, if the frame redundancy is large enough then $\Sigma\Delta$ outperforms PCM. We present numerical examples to illustrate this. This also shows that $\Sigma\Delta$ quantization achieves the optimal approximation order for PCM with consistent reconstruction.

II. Background

A. Frame theory

The theory of frames in harmonic analysis is due to Duffin and Schaeffer [22]. Modern expositions on frame theory can be found in [3], [23], [24]. In the following definitions, Λ is an at most countable index set.

Definition II.1. *A collection $F = \{e_n\}_{n\in\Lambda}$ in a Hilbert space H is a frame for H if there exists $0 < A \leq B < \infty$ such that*

$$\forall x \in H, \quad A||x||^2 \leq \sum_{n\in\Lambda} |\langle x, e_n\rangle|^2 \leq B||x||^2.$$

The constants A and B are called the frame bounds.

A frame is *tight* if $A = B$. An important remark is that the size of the frame bound of a *normalized* or *uniform* tight frame, i.e., a tight frame with $||e_n|| = 1$ for all n, "measures" the redundancy of the system. For example, if $A = 1$ then a normalized tight frame must be an orthonormal basis and there is no redundancy, see Proposition 3.2.1 of [3]. The larger the frame bound $A \geq 1$ is, the more redundant a normalized tight frame is.

Definition II.2. *Let $\{e_n\}_{n\in\Lambda}$ be a frame for a Hilbert space H with frame bounds A and B. The* analysis operator

$$L : H \to l^2(\Lambda)$$

is defined by $(Lx)_k = \langle x, e_k\rangle$. The operator $S = L^*L$ is called the frame operator*, and it satisfies*

$$AI \leq S \leq BI, \tag{3}$$

where I is the identity operator on H. The inverse of S, S^{-1}, is called the dual frame operator*, and it satisfies*

$$B^{-1}I \leq S^{-1} \leq A^{-1}I. \tag{4}$$

The following theorem illustrates why frames can be useful in signal processing.

Theorem II.3. *Let $\{e_n\}_{n\in\Lambda}$ be a frame for H with frame bounds A and B, and let S be the corresponding frame operator. Then $\{S^{-1}e_n\}_{n\in\Lambda}$ is a frame for H with frame bounds B^{-1} and A^{-1}. Further, for all $x \in H$*

$$x = \sum_{n\in\Lambda} \langle x, e_n\rangle (S^{-1}e_n) \tag{5}$$

$$= \sum_{n\in\Lambda} \langle x, (S^{-1}e_n)\rangle e_n, \tag{6}$$

with unconditional convergence of both sums.

The atomic decompositions in (5) and (6) are the first step towards a digital representation. If the frame is tight with frame bound A, then both frame expansions are equivalent and we have

$$\forall x \in H, \quad x = A^{-1} \sum_{n\in\Lambda} \langle x, e_n\rangle e_n. \tag{7}$$

When the Hilbert space H is $\mathbb{R}^d$ or $\mathbb{C}^d$, and Λ is finite, the frame is referred to as a *finite frame* for H. In this case, it is straightforward to check if a set of vectors is a tight frame. Given a set of N vectors, $\{v_n\}_{n=1}^N$, in $\mathbb{R}^d$ or $\mathbb{C}^d$, define the associated matrix L to be the $N \times d$ matrix whose rows are the $\overline{v}_n$. The following lemma can be found in [25].

Lemma II.4. *A set of vectors $\{v_n\}_{n=1}^N$ in $H = \mathbb{R}^d$ or $H = \mathbb{C}^d$ is a tight frame with frame bound A if and only if its associated matrix L satisfies $S = L^*L = AI_d$, where L^* is the conjugate transpose of L, and I_d is the $d \times d$ identity matrix.*

For the important case of *finite normalized tight frames* for $\mathbb{R}^d$ and $\mathbb{C}^d$, the frame constant A is N/d, where N is the cardinality of the frame [26], [10], [2], [25].

B. PCM algorithms and Bennett's white noise assumption

Let $\{e_n\}_{n=1}^N$ be a normalized tight frame for $\mathbb{R}^d$, so that each $x \in \mathbb{R}^d$ has the frame expansion

$$x = \frac{d}{N} \sum_{n=1}^N x_n e_n, \quad x_n = \langle x, e_n\rangle.$$

Given $\delta > 0$, the 2 $\lceil 1/\delta\rceil$-level PCM quantizer with step size δ replaces each $x_n \in \mathbb{R}$ with

$$q_n = q_n(x) = \begin{cases} \delta(\lceil x_n/\delta\rceil - 1/2) & \text{if } |x_n| < 1, \\ \delta(\lceil 1/\delta\rceil - 1/2) & \text{if } x_n \geq 1, \\ -\delta(\lceil 1/\delta\rceil - 1/2) & \text{if } x_n \leq -1, \end{cases} \tag{8}$$

where $\lceil\cdot\rceil$ denotes the *ceiling function*. Thus, PCM quantizes x by

$$\widetilde{x} = \frac{d}{N} \sum_{n=1}^N q_n e_n.$$

It is easy to see that

$$\forall n, \quad |x_n| < 1 \quad \Longrightarrow \quad |x_n - q_n| \leq \delta/2. \tag{9}$$

PCM quantization as defined above assumes linear reconstruction from the PCM quantized coefficients, q_n. We very briefly address the nonlinear technique of consistent reconstuction in Section VI.

Fix $\delta > 0$, and let $\|\cdot\|$ be the d-dimensional Euclidean 2-norm. Let $x \in \mathbb{R}^d$ and let $\widetilde{x}$ be the quantized expansion given by $2\lceil 1/\delta\rceil$-level PCM quantization. If $||x|| < 1$ then by (9) the approximation error $||x - \widetilde{x}||$ satisfies

$$\|x - \widetilde{x}\| = \frac{d}{N} \|\sum_{n=1}^N (x_n - q_n)e_n\|$$

$$\leq \left(\frac{\delta}{2}\right)\left(\frac{d}{N}\right) \sum_{n=1}^N \|e_n\| = \left(\frac{d}{2}\right)\delta. \tag{10}$$

This error estimate does not utilize the redundancy of the frame. A common way to improve the estimate (10) is to make statistical assumptions on the differences $x_n - q_n$, e.g., [21], [2].

Example II.5 (Bennett's white noise assumption). *Let $\{e_n\}_{n=1}^N$ be a normalized tight frame for $\mathbb{R}^d$ with frame bound $A = N/d$, let $x \in \mathbb{R}^d$, and let x_n, q_n, and $\widetilde{x}$ be defined as above. Since the "pointwise" estimate (10) is unsatisfactory, a different idea is to derive better error estimates which hold "on average" under certain statistical assumptions.*

Let ν be a probability measure on $\mathbb{R}^d$, and consider the random variables $\eta_n = x_n - q_n$ with the probability distribution μ_n induced by ν as follows. For $\mathcal{B} \subseteq \mathbb{R}$ measurable,

$$\mu_n(\mathcal{B}) = \nu(\{x \in \mathbb{R}^d : \langle x, e_n\rangle - q_n(x) \in \mathcal{B}\}).$$

The classical approach dating back to Bennett, [21], is to assume that the quantization noise sequence*, $\{\eta_n\}_{n=1}^N$, is a sequence of independent, identically distributed random variables with mean 0 and variance σ^2. In other words, $\mu_n = \mu$ for $n = 1, \cdots, N$, and the joint probability distribution $\mu_{1,\cdots,N}$ of $\{\eta_n\}_{n=1}^N$ is given by $\mu_{1,\cdots,N} = \mu^N$. We shall refer* *to this statistical assumption on* $\{\eta_n\}_{n=1}^N$ *as* Bennett's white noise assumption.

It was shown in [2], that under Bennett's white noise assumption, the mean square (approximation) error (MSE) satisfies

$$MSE_{PCM} = E(\|x-\widetilde{x}\|^2) = \frac{d\sigma^2}{A} = \frac{d^2\sigma^2}{N}, \qquad (11)$$

where the expectation $E(||x-\widetilde{x}||^2)$ *is defined by*

$$E(||x-\tilde{x}||^2) = \int_{\mathbb{R}^d} ||x-\tilde{x}||^2 d\nu(x),$$

which can be rewritten using Bennett's white noise assumption as

$$E(||x-\tilde{x}||^2) = \int_{\mathbb{R}^N} \frac{d}{N} ||\sum_{n=1}^N \eta_n e_n||^2 d\mu^N(\eta_1, \cdots, \eta_N).$$

Since we are considering PCM quantization with stepsize δ*, and in view of (9), it is quite natural to assume that each* η_n *is a uniform random variable on* $[-\frac{\delta}{2}, \frac{\delta}{2}]$*, and hence has mean 0, and variance* $\sigma^2 = \delta^2/12$*, [27]. In this case one has*

$$MSE_{PCM} = \frac{d\delta^2}{12A} = \frac{d^2\delta^2}{12N}. \qquad (12)$$

Although (12) in Example II.5 represents an improvement over (10) it is still unsatisfactory for the following reasons:

(a) The MSE bound (12) only gives information about the average quantizer performance.
(b) As one increases the redundancy of the expansion, i.e., as the frame bound A increases, the MSE given in (12) decreases only as $1/A$, i.e., the redundancy of the expansion is not utilized very efficiently.
(c) (12) is computed under assumptions that are not rigorous and, at least in certain cases, not true. See [28] for an extensive discussion and a partial deterministic analysis of the quantizer error sequence $\{\eta_n\}$. In Example II.6, we show an elementary setting where Bennett's white noise assumption does not hold for PCM quantization of finite frame expansions.

Since a redundant frame has more elements than are necessary to span the signal space, there will be interdependencies between the frame elements, and hence between the frame coefficients. It is intuitively reasonable to expect that this redundancy and interdependency may violate the independence part of Bennett's white noise assumption. The following example makes this intuition precise.

Example II.6 (Shortcomings of the noise assumption). *Consider the normalized tight frame for* $\mathbb{R}^2$*, with frame bound* $A = N/2$*, given by*

$$\{e_n\}_{n=1}^N, \quad e_n = (\cos(2\pi n/N), \sin(2\pi n/N)),$$

where $N > 2$ *is assumed to be even. Given* $x \in \mathbb{R}^2$*, and let* $x_n = \langle x, e_n \rangle$ *be the corresponding* n*th frame coefficient. It is easy to see that since* N *is even*

$$\forall n, \ e_n = -e_{n+N/2},$$

and hence

$$\forall n, \ x_n = -x_{n+N/2}.$$

Next, note that for almost every $x \in \mathbb{R}^2$ *(with respect to Lebesgue measure) one has*

$$\forall n, \ x_n \notin \delta\mathbb{Z}.$$

By the definition of the PCM scheme, this implies that for almost every $x \in \mathbb{R}^2$ *with* $||x|| < 1$ *one has* $q_n = -q_{n+N/2}$*, and hence* $\eta_n = -\eta_{n+N/2}$*. This means that the quantization noise sequence* $\{\eta_n\}$ *is not independent and that Bennett's white noise assumption is violated. Thus, the MSE predicted by (12) will not be attained in this case. One can rectify the situation by applying the white noise assumption to the frame that is generated by deleting half of the points to ensure that only one of* e_n *and* $e_{n+N/2}$ *is left in the resulting set.*

In addition to the limitations of PCM mentioned above, it is also well known that PCM has poor robustness properties in the bandlimited setting, [6]. In view of these shortcomings of PCM quantization, we seek an alternate quantization scheme which is well suited to utilizing frame redundancy. We shall show that the class of Sigma-Delta ($\Sigma\Delta$) schemes perform exceedingly well when used to quantize redundant finite frame expansions.

III. $\Sigma\Delta$ ALGORITHMS FOR FRAMES FOR $\mathbb{R}^d$

Sigma-Delta ($\Sigma\Delta$) quantizers are widely implemented to quantize oversampled bandlimited functions [29], [6]. Here, we define the fundamental $\Sigma\Delta$ algorithm with the aim of using it to quantize finite frame expansions.

Let $K \in \mathbb{N}$ and $\delta > 0$. Given the *midrise* quantization alphabet

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

consisting of $2K$ elements, we define the *$2K$-level midrise uniform scalar quantizer* with stepsize δ by

$$Q(u) = \arg\min_{q \in \mathcal{A}_K^\delta} |u-q| \qquad (13)$$

Thus, $Q(u)$ is the element of the alphabet which is closest to u. If two elements of $\mathcal{A}_K^\delta$ are equally close to u then let $Q(u)$ be the larger of these two elements, i.e., the one larger than u. For simplicity, we only consider midrise quantizers, although many of our results are valid more generally.

Definition III.1. *Given* $K \in \mathbb{N}, \delta > 0$*, and the corresponding midrise quantization alphabet and* $2K$*-level midrise uniform scalar quantizer with stepsize* δ*. Let* $\{x_n\}_{n=1}^N \subseteq \mathbb{R}^d$*, and let* p *be a permutation of* $\{1, 2, \cdots, N\}$*. The associated* first order $\Sigma\Delta$ quantizer *is defined by the iteration*

$$\begin{aligned} u_n &= u_{n-1} + x_{p(n)} - q_n, \\ q_n &= Q(u_{n-1} + x_{p(n)}), \end{aligned} \qquad (14)$$

where u_0 *is a specified constant. The first order* $\Sigma\Delta$ *quantizer produces the* quantized sequence $\{q_n\}_{n=1}^N$*, and an auxiliary sequence* $\{u_n\}_{n=0}^N$ *of state variables.*

Thus, a first-order $\Sigma\Delta$ quantizer is a *$2K$-level first-order $\Sigma\Delta$ quantizer with step size δ* if it is defined by means of (14), where Q is defined by (13). We shall refer to the permutation p as the *quantization order*.

The following proposition, cf., [6], shows that the first-order $\Sigma\Delta$ quantizer is *stable*, i.e., the auxiliary sequence $\{u_n\}$ defined by (14) is uniformly bounded if the input sequence $\{x_n\}$ is appropriately uniformly bounded.

Proposition III.2. *Let K be a positive integer, let $\delta > 0$, and consider the $\Sigma\Delta$ system defined by (14) and (13). If $|u_0| \leq \delta/2$ and*

$$\forall n = 1, \cdots, N, \quad |x_n| \leq (K - 1/2)\delta,$$

then

$$\forall n = 1, \cdots, N, \quad |u_n| \leq \delta/2.$$

Proof: Without loss of generality assume that p is the identity permutation. The proof proceeds by induction. The base case, $|u_0| \leq \delta/2$, holds by assumption. Next, suppose that $|u_{j-1}| \leq \delta/2$. This implies that $|u_{j-1} - x_j| \leq K\delta$, and hence, by (14) and the definition of Q,

$$|u_j| = |u_{j-1} - x_j - Q(u_{j-1} - x_j)| \leq \delta/2.$$

A. Frame variation

Let $F = \{e_n\}_{n=1}^N$ be a finite frame for $\mathbb{R}^d$ and let

$$x = \sum_{n=1}^{N} x_n S^{-1} e_n, \quad x_n = \langle x, e_n \rangle, \tag{15}$$

be the corresponding frame expansion for some $x \in \mathbb{R}^d$. Since this frame expansion is a finite sum, the representation is independent of the order of summation. In fact, recall that by Theorem II.3, any frame expansion in a Hilbert space converges unconditionally.

Although frame expansions do not depend on the ordering of the frame, the $\Sigma\Delta$ scheme in Definition III.1 is iterative in nature, and *does* depend strongly on the order in which the frame coefficients are quantized. In particular, we shall show that changing the order in which frame coefficients are quantized can have a drastic effect on the performance of the $\Sigma\Delta$ scheme. This, of course, stands in stark contrast to PCM schemes which are order independent. The $\Sigma\Delta$ scheme (14) takes advantage of the fact that there are "interdependencies" between the frame elements in a redundant frame expansion. This is a main underlying reason why $\Sigma\Delta$ schemes outperform PCM schemes, which quantize frame coefficients without considering any "interdependencies".

We now introduce the notion of *frame variation*. This will play an important role in our error estimates and directly reflects the importance of carefully choosing the order in which frame coefficients are quantized.

Definition III.3. *Let $F = \{e_n\}_{n=1}^N$ be a finite frame for $\mathbb{R}^d$, and let p be a permutation of $\{1, 2, \ldots, N\}$. We define the* variation *of the frame F with respect to p as*

$$\sigma(F, p) := \sum_{n=1}^{N-1} \|e_{p(n)} - e_{p(n+1)}\|. \tag{16}$$

Roughly speaking, if a frame F has low variation with respect to p, then the frame elements will not oscillate too much in that ordering and there is more "interdependence" between succesive frame elements.

B. Basic error estimates

We now derive error estimates for the $\Sigma\Delta$ scheme in Definition III.1 for $K \in \mathbb{N}$ and $\delta > 0$. Given a frame $F = \{e_n\}_{n=1}^N$ for $\mathbb{R}^d$, a permutation p of $\{1, 2, \cdots, N\}$, and $x \in \mathbb{R}^d$, we shall calculate how well the quantized expansion

$$\widetilde{x} = \sum_{n=1}^{N} q_n S^{-1} e_{p(n)}$$

approximates the frame expansion

$$x = \sum_{n=1}^{N} x_{p(n)} S^{-1} e_{p(n)}, \quad x_{p(n)} = \langle x, e_{p(n)} \rangle.$$

Here, $\{q_n\}_{n=1}^N$ is the quantized sequence which is calculated using Definition III.1 and the sequence of frame coefficients, $\{x_{p(n)}\}_{n=1}^N$. We now state our first result on the *approximation error*, $||x - \widetilde{x}||$. We shall use $|| \cdot ||_{op}$ to denote the operator norm induced by the Euclidean norm, $|| \cdot ||$, for $\mathbb{R}^d$.

Theorem III.4. *Given the $\Sigma\Delta$ scheme of Definition III.1. Let $F = \{e_n\}_{n=1}^N$ be a finite normalized frame for $\mathbb{R}^d$, let p be a permutation of $\{1, 2, \ldots, N\}$, let $|u_0| \leq \delta/2$, and let $x \in \mathbb{R}^d$ satisfy $||x|| \leq (K - 1/2)\delta$. The approximation error $||x - \widetilde{x}||$ satisfies*

$$||x - \widetilde{x}|| \leq ||S^{-1}||_{op} \left(\sigma(F, p)\frac{\delta}{2} + |u_N| + |u_0| \right),$$

where S^{-1} is the inverse frame operator for F.

Proof:

$$\begin{aligned} x - \widetilde{x} &= \sum_{n=1}^{N} (x_{p(n)} - q_n) S^{-1} e_{p(n)} \\ &= \sum_{n=1}^{N} (u_n - u_{n-1}) S^{-1} e_{p(n)} \\ &= \sum_{n=1}^{N-1} u_n S^{-1} (e_{p(n)} - e_{p(n+1)}) \\ &\quad + u_N S^{-1} e_{p(N)} - u_0 S^{-1} e_{p(1)}. \end{aligned} \tag{17}$$

Since $||x|| \leq (K - 1/2)\delta$ it follows that

$$\forall\ 1 \leq n \leq N, \quad |x_n| = |\langle x, e_n \rangle| \leq (K - 1/2)\delta.$$

Thus, by Proposition III.2,

$$\begin{aligned} ||x - \widetilde{x}|| &\leq \sum_{n=1}^{N} \frac{\delta}{2}\ ||S^{-1}||_{op} ||e_{p(n)} - e_{p(n+1)}|| \\ &\quad + |u_N|\ ||S^{-1}||_{op} + |u_0|\ ||S^{-1}||_{op} \\ &= ||S^{-1}||_{op} \left(\sigma(F, p)\frac{\delta}{2} + |u_0| + |u_N| \right). \end{aligned}$$

Theorem III.4 is stated for general normalized frames, but since finite tight frames are especially desirable in applications, we shall restrict the remainder of our discussion to tight frames. The utility of finite normalized tight frames is apparent in the simple reconstruction formula (7). Note that general finite normalized frames for $\mathbb{R}^d$ are elementary to construct. In fact, any finite subset of $\mathbb{R}^d$ is a frame for its span. However, the construction and characterization of finite normalized tight frames is much more interesting due to the additional algebraic constraints involved [26].

Corollary III.5. *Given the $\Sigma\Delta$ scheme of Definition III.1. Let $F = \{e_n\}_{n=1}^N$ be a normalized tight frame for $\mathbb{R}^d$ with frame bound $A = N/d$, let p be a permutation of $\{1, 2, \ldots, N\}$, let $|u_0| \leq \delta/2$, and let $x \in \mathbb{R}^d$ satisfy $||x|| \leq (K - 1/2)\delta$. The approximation error $||x - \widetilde{x}||$ satisfies*

$$||x - \widetilde{x}|| \leq \frac{d}{N}\left(\sigma(F, p)\frac{\delta}{2} + |u_N| + |u_0|\right).$$

Proof: As discussed in Section II-A, a tight frame $F = \{e_n\}_{n=1}^N$ for $\mathbb{R}^d$ has frame bound $A = N/d$, and, by (4), and Lemma II.4,

$$||S^{-1}||_{op} = ||\frac{d}{N}I||_{op} = d/N.$$

The result now follows from Theorem III.4.

Corollary III.6. *Given the $\Sigma\Delta$ scheme of Definition III.1. Let $F = \{e_n\}_{n=1}^N$ be a normalized tight frame for $\mathbb{R}^d$ with frame bound $A = N/d$, let p be a permutation of $\{1, 2, \ldots, N\}$, let $|u_0| \leq \delta/2$, and let $x \in \mathbb{R}^d$ satisfy $||x|| \leq (K - 1/2)\delta$. The approximation error $||x - \widetilde{x}||$ satisfies*

$$||x - \widetilde{x}|| \leq \frac{\delta d}{2N}\left(\sigma(F, p) + 2\right).$$

Proof: Apply Corollary III.5 and Proposition III.2.

Recall that the initial state u_0 in (14) can be chosen arbitrarily. It is therefore convenient to take $u_0 = 0$, because this will give a smaller constant in the approximation error given by Theorem III.4. Likewise, the error constant can be improved if one has more information about the final state variable, $|u_N|$. It is somewhat surprising that for *zero sum* frames the value of $|u_N|$ is completely determined by whether the frame has an even or odd number of elements.

Theorem III.7. *Given the $\Sigma\Delta$ scheme of Definition III.1. Let $F = \{e_n\}_{n=1}^N$ be a normalized tight frame for $\mathbb{R}^d$ with frame bound $A = N/d$, and assume that F satisfies the zero sum condition*

$$\sum_{n=1}^{N} e_n = 0. \tag{18}$$

Additionally, set $u_0 = 0$ in (14). Then

$$|u_N| = \begin{cases} 0, & \text{if } N \text{ even;} \\ \delta/2, & \text{if } N \text{ odd.} \end{cases} \tag{19}$$

Proof: Note that (14) implies

$$u_N = u_0 + \sum_{n=1}^{N} x_n - \sum_{n=1}^{N} q_n = \sum_{n=1}^{N} x_n - \sum_{n=1}^{N} q_n. \tag{20}$$

Next, (18) implies

$$\sum_{n=1}^{N} x_n = \sum_{n=1}^{N} \langle x, e_n \rangle = \langle x, \sum_{n=1}^{N} e_n \rangle = 0. \tag{21}$$

By the definition of the midrise quantization alphabet $\mathcal{A}_K^\delta$ each q_n is an odd integer multiple of $\delta/2$.

If N is even it follows that $\sum_{n=1}^N q_n$ is an integer multiple of δ. Thus, by (20) and (21), u_N is an integer multiple of δ. However, $|u_N| \leq \delta/2$ by Proposition III.2, so that we have $u_N = 0$.

If N is odd it follows that $\sum_{n=1}^N q_n$ is an odd integer multiple of $\delta/2$. Thus, by (20) and (21), u_N is an odd integer multiple of $\delta/2$. However, $|u_N| \leq \delta/2$ by Proposition III.2, so that we have $|u_N| = \delta/2$.

Corollary III.8. *Given the $\Sigma\Delta$ scheme of Definition III.1. Let $F = \{e_n\}_{n=1}^N$ be a normalized tight frame for $\mathbb{R}^d$ with frame bound $A = N/d$, and assume that F satisfies the zero sum condition (18). Let p be a permutation of $\{1, \cdots, N\}$ and let $x \in \mathbb{R}^d$ satisfy $||x|| \leq (K - 1/2)\delta$. Additionally, set $u_0 = 0$ in (14). Then the approximation error $||x - \widetilde{x}||$ satisfies*

$$||x - \widetilde{x}|| \leq \begin{cases} \frac{\delta d}{2N}\sigma(F, p), & \text{if } N \text{ even;} \\ \frac{\delta d}{2N}\left(\sigma(F, p) + 1\right), & \text{if } N \text{ odd.} \end{cases} \tag{22}$$

Proof: Apply Corollary III.5, Theorem III.7, and Proposition III.2.

Corollary III.8 shows that as a consequence of Theorem III.7, one has smaller constants in the error estimate for $||x - \widetilde{x}||$ when the frame size N is even. Theorem III.7 makes an even bigger difference when deriving refined estimates as in Section V, or when dealing with higher order $\Sigma\Delta$ schemes [30].

IV. Families of frames with bounded variation

One way to obtain arbitrarily small approximation error, $||x - \widetilde{x}||$, using the estimates of the previous section is simply to fix a frame and decrease the quantizer step size δ towards zero, while letting $K = \lceil 1/\delta \rceil$. By Corollary III.6, as δ goes to 0, the approximation error goes to zero. However, this approach is not always be desirable. For example, in analog-to-digital (A/D) conversion of bandlimited signals, it can be quite costly to build quantizers with very high resolution, i.e. small δ and large K, e.g., [6]. Instead, many practical applications involving A/D and D/A converters make use of oversampling, i.e., redundant frames, and use low resolution quantizers, e.g., [31]. To be able to adopt this type of approach for the quantization of finite frame expansions, it is important to be able to construct families of frames with uniformly bounded frame variation.

Let us begin by making the observation that if $F = \{e_n\}_{n=1}^N$ is a finite normalized frame and p is any permutation of $\{1, 2, \cdots, N\}$ then $\sigma(F, p) \leq 2(N - 1)$. However, this bound is too weak to be of much use since substituting it into an error bound such as the even case of (22) only gives

$$||x - \widetilde{x}|| \leq \frac{\delta d(N - 1)}{N}.$$

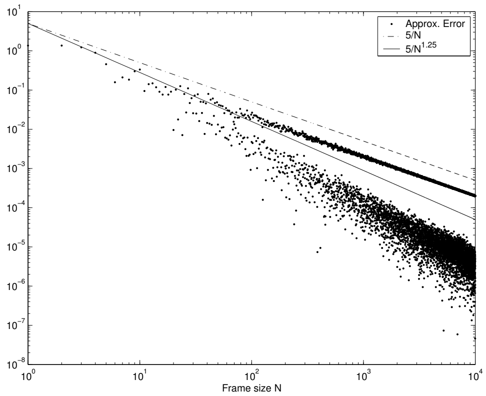

Fig. 1. The frame coefficients of $x = (1/\pi, \sqrt{3/17})$ with respect to the Nth roots of unity are quantized using the first order $\Sigma\Delta$ scheme. This log-log plot shows the approximation error $||x - \widetilde{x}||$ as a function of N compared with $5/N$ and $5/N^{1.25}$.

In particular, this bound does not go to zero as N gets large, i.e., as one chooses more redundant frames. On the other hand, if one finds a family of frames and a sequence of permutations, such that the resulting frame variations are uniformly bounded, then one is able to obtain an approximation error of order $1/N$.

Example IV.1 (Roots of unity). *For $N > 2$, let $R_N = \{e_n^N\}_{n=1}^N$ be the Nth roots of unity viewed as vectors in $\mathbb{R}^2$, namely,*

$$\forall n = 1, \cdots, N, \quad e_n^N = (\cos(2\pi n/N), \sin(2\pi n/N)).$$

It is well known that R_N is a tight frame for $\mathbb{R}^2$ with frame bound $N/2$, e.g., see [26]. In this example, we shall always consider R_N in its natural ordering $\{e_n^N\}_{n=1}^N$. Note that $\sum_{n=1}^N e_n^N = 0$.

Since $||e_n - e_{n+1}|| \leq 2\pi/N$, it follows that

$$\forall N, \quad \sigma(R_N, p) \leq 2\pi, \tag{23}$$

where p is the identity permutation of $\{1, 2, \cdots, N\}$.

Thus, the error estimate of Corollary III.8 gives

$$||x - \widetilde{x}|| \leq \begin{cases} \frac{\delta}{N} 2\pi, & \text{if } N > 2 \text{ even;} \\ \frac{\delta}{N}(2\pi + 1), & \text{if } N > 2 \text{ odd.} \end{cases} \tag{24}$$

Figure 1 shows a log-log plot of the approximation error $||x - \tilde{x}_N||$ as a function of N, when the Nth roots of unity are used to quantize the input $x = (1/\pi, \sqrt{3/17})$. The figure also shows a log-log plot of $5/N$ and $5/N^{1.25}$for comparison. Note that the approximation error exhibits two very different types of behavior. In particular, for odd N the approximation error appears to behave like $1/N$ asymptotically, whereas for even N the approximation error is much smaller. We shall explain this phenomenon in Section V.

The most natural examples of normalized tight frames in $\mathbb{R}^d, d > 2$ are the harmonic frames, e.g., see [10], [25], [2]. These frames are constructed using rows of the Fourier matrix.

Example IV.2 (Harmonic frames). *We shall show that harmonic frames in their natural ordering have uniformly bounded frame variation. We follow the notation of [25], although the terminology "harmonic frame" is not specifically used there. The definition of the harmonic frame $H_N^d = \{e_j\}_{j=0}^{N-1}$, $N \geq d$, depends on whether the dimension d is even or odd.*

If $d \geq 2$ is even let

$$e_j = \sqrt{\frac{2}{d}} \left[\cos\frac{2\pi j}{N}, \sin\frac{2\pi j}{N}, \cos\frac{2\pi 2j}{N}, \sin\frac{2\pi 2j}{N}, \cos\frac{2\pi 3j}{N}, \sin\frac{2\pi 3j}{N}, \cdots, \cos\frac{2\pi \frac{d}{2} j}{N}, \sin\frac{2\pi \frac{d}{2} j}{N} \right]$$

for $j = 0, 1, \cdots, N-1$.

If $d > 1$ is odd let

$$e_j = \sqrt{\frac{2}{d}} \left[\frac{1}{\sqrt{2}}, \cos\frac{2\pi j}{N}, \sin\frac{2\pi j}{N}, \cos\frac{2\pi 2j}{N}, \sin\frac{2\pi 2j}{N}, \cos\frac{2\pi 3j}{N}, \sin\frac{2\pi 3j}{N}, \cdots, \cos\frac{2\pi \frac{d-1}{2} j}{N}, \sin\frac{2\pi \frac{d-1}{2} j}{N} \right]$$

for $j = 0, 1, \cdots, N-1$.

It is shown in [25] that H_N^d, as defined above, is a normalized tight frame for $\mathbb{R}^d$. If d is even then H_N^d satisfies the zero sum condition (18). If d is odd the frame is not zero sum, and, in fact,

$$\sum_{j=0}^{N-1} e_j = (\frac{N}{\sqrt{d}}, 0, 0, \cdots, 0) \in \mathbb{R}^d.$$

The verification of the zero sum condition for d even follows by noting that, for each $k \in \mathbb{Z}$ and not of the form $k = mN$, we have

$$\sum_{j=0}^{N-1} \cos\frac{2\pi k j}{N} = Re\left[\sum_{j=0}^{N-1} (e^{2\pi i k/N})^j\right] = 0.$$

and

$$\sum_{j=0}^{N-1} \sin\frac{2\pi k j}{N} = Im\left[\sum_{j=0}^{N-1} (e^{2\pi i k/N})^j\right] = 0.$$

Let us now estimate the frame variation for harmonic frames. First, suppose d even, and let p be the identity permutation. Calculating directly and using the mean value theorem in the first inequality, we have

$$\begin{aligned}
\sqrt{\frac{d}{2}}\sigma(H_N^d, p) &= \sqrt{\frac{d}{2}} \sum_{j=0}^{N-2} ||e_j - e_{j+1}|| \\
&= \sum_{j=0}^{N-2} \left[\sum_{k=1}^{d/2} \left(\cos\frac{2\pi k j}{N} - \cos\frac{2\pi k(j+1)}{N} \right)^2 \right. \\
&\quad \left. + \sum_{k=1}^{d/2} \left(\sin\frac{2\pi k j}{N} - \sin\frac{2\pi k(j+1)}{N} \right)^2 \right]^{\frac{1}{2}} \\
&\leq \sum_{j=0}^{N-2} \left[2\sum_{k=1}^{d/2} \left(\frac{2\pi k}{N}\right)^2 \right]^{\frac{1}{2}} \leq 2\pi\sqrt{2} \left[\sum_{k=1}^{d/2} k^2 \right]^{\frac{1}{2}} \\
&= 2\pi\sqrt{2} \left[\frac{d(d/2+1)(d+1)}{12} \right]^{\frac{1}{2}} \leq 2\pi \sqrt{\frac{d}{6}(d+1)}.
\end{aligned}$$

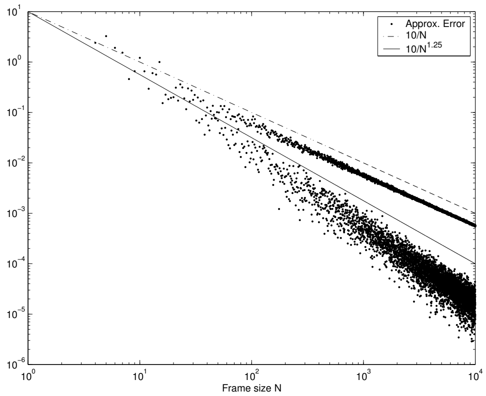

Fig. 2. The frame coefficients of $x = (1/\pi, 1/50, \sqrt{3/17}, e^{-1})$ with respect to the harmonic frame H_N^4 are quantized using the first order $\Sigma\Delta$ scheme. This log-log plot shows the approximation error $||x - \widetilde{x}||$ as a function of N compared with $10/N$ and $10/N^{1.25}$.

If d is odd then, proceeding as above, we have

$$\sqrt{\frac{d}{2}}\sigma(H_N^d, p) \leq 2\pi\sqrt{2}\left[\sum_{k=1}^{(d-1)/2} k^2\right]^{\frac{1}{2}} \leq 2\pi\sqrt{\frac{d}{6}}(d+1).$$

Thus

$$\sigma(H_N^d, p) \leq \frac{2\pi(d+1)}{\sqrt{3}}, \tag{25}$$

where p is the identity permutation, i.e., we consider the natural ordering as in the definition of H_N^d.

We can now derive error estimates for $\Sigma\Delta$ quantization of harmonic frames in their natural order. If we set $u_0 = 0$ and assume that $x \in \mathbb{R}^d$ satisfies $||x|| \leq (K - 1/2)\delta$, then combining (25), Corollaries III.2, III.5, and III.8, and the fact that H_N^d satisfies (18) if N is even gives

$$||x - \widetilde{x}|| \leq \begin{cases} \frac{\delta d}{2N}\frac{2\pi(d+1)}{\sqrt{3}}, & \text{if } d \text{ is even and } N \text{ is even,} \\ \frac{\delta d}{2N}\left[\frac{2\pi(d+1)}{\sqrt{3}} + 1\right], & \text{otherwise.} \end{cases}$$

Figure 2 shows a log-log plot of the approximation error $||x - \tilde{x}_N||$ as a function of N, when the harmonic frame H_N^4 is used to quantize the input $x = (1/\pi, 1/50, \sqrt{3/17}, e^{-1})$. The figure also shows a log-log plot of $10/N$ and $10/N^{1.25}$for comparison.

As discussed earlier, the $\Sigma\Delta$ algorithm is quite sensitive to the ordering in which the frame coefficients are quantized. In Examples IV.1 and IV.2, the natural frame order gave uniformly bounded frame variation. Let us next consider an example where a bad choice of frame ordering leads to poor approximation error.

Example IV.3 (Order matters). *Consider the normalized tight frame for $\mathbb{R}^2$ which is given by the 7th roots of unity, viz., $R_7 = \{e_n\}_{n=1}^7$, where $e_n = (\cos(2\pi n/7), \sin(2\pi n/7))$. We randomly choose 10,000 points in the unit ball of $\mathbb{R}^2$. For each of these 10,000 points we first quantize the corresponding frame coefficients in their natural order using (14) with the alphabet*

$$\mathcal{A}_4^{1/4} = \{-7/8, -5/8, -3/8, -1/8, 1/8, 3/8, 5/8, 7/8\},$$

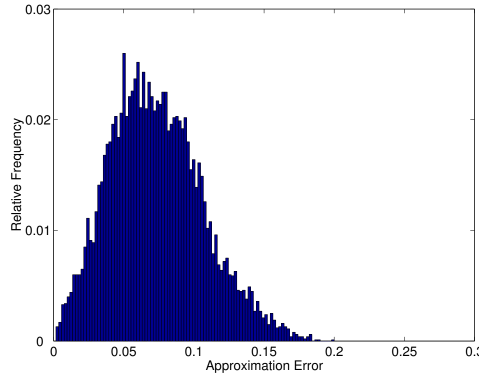

Fig. 3. Histogram of approximation error in Example IV.3 for the natural ordering.

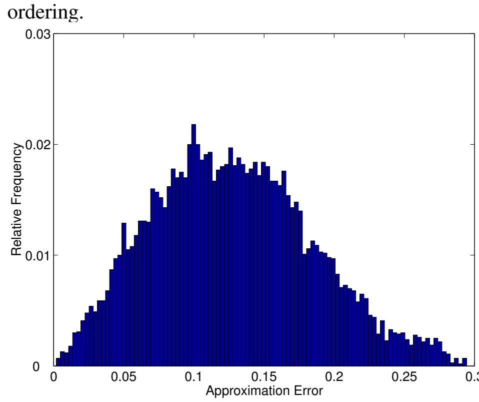

Fig. 4. Histogram of approximation error in Example IV.3 for an ordering giving higher variation.

and setting $x_n = \langle x, e_n\rangle$. Figure 3 shows the histogram of the corresponding approximation errors. Next, we quantize the frame coefficients of the same 10,000 points, only this time after reordering the frame coefficients as $x_1, x_4, x_7, x_3, x_6, x_2, x_5$. Figure 4 shows the histogram of the corresponding approximation errors in this case.

Clearly, the average approximation error for the new ordering is significantly larger than the average approximation error associated with the original ordering. This is intuitively explained by the fact that the natural ordering has significantly smaller frame variation than the other ordering. In particular, let p_1 be the identity permutation and let p_2 be the permutation corresponding the reordered frame coefficients used above. A direct calculation shows that

$$\sigma(F, p_1) \approx 5.2066 \quad \textit{and} \quad \sigma(F, p_2) \approx 11.6991.$$

In view of this example it is important to choose carefully the order in which frame coefficients are quantized. In $\mathbb{R}^2$ there is always a simple good choice.

Theorem IV.4. *Let $F_N = \{e_n\}_{n=1}^N$ be a normalized frame for $\mathbb{R}^2$, where $e_n = (\cos(\alpha_n), \sin(\alpha_n))$ and $0 \leq \alpha_n < 2\pi$. If p is a permutation of $\{1, 2, \cdots, N\}$ such that $\alpha_{p(n)} \leq \alpha_{p(n+1)}$ for all $n \in \{1, 2, \cdots, N-1\}$, then $\sigma(F_N, p) \leq 2\pi$.*

Proof: Is is easy to verify that

$$||e_{p(n)} - e_{p(n+1)}|| \leq |\alpha_{p(n)} - \alpha_{p(n+1)}|.$$

By the choice of p, and since $0 \leq \alpha_n < 2\pi$, it follows that

$$\sigma(F_N, p) = \sum_{n=1}^{N-1} ||e_{p(n)} - e_{p(n+1)}|| \leq 2\pi.$$

V. Refined estimates and lower bounds

In Figure 1 of Example IV.1, we saw that the approximation error appears to exhibit very different types of behavior depending on whether N is even or odd. In the even case the approximation error appears to decay better than the $1/N$ estimate given by the results in Section III-B; in the odd case it appears that the $1/N$ actually serves as a lower bound, as well as an upper bound, for the approximation error. This dichotomy goes beyond Corollary III.8, which only predicts different constants in the even/odd approximations as opposed to different orders of approximation. In this section we shall explain this phenomenon.

Let $\{F_N\}_{N=d}^{\infty}$ be a family of normalized tight frames for $\mathbb{R}^d$, with $F_N = \{e_n^N\}_{n=1}^N$, so that F_N has frame bound N/d. If $x \in \mathbb{R}^d$, then $\{x_n^N\}_{n=1}^N$ will denote the corresponding sequence of frame coefficients with respect to F_N, i.e., $x_n^N = \langle x, e_n^N \rangle$. Let $\{q_n^N\}_{n=1}^N$ be the quantized sequence which is obtained by running the $\Sigma\Delta$ scheme, (14), on the input sequence $\{x_n^N\}_{n=1}^N$, and let $\{u_n^N\}_{n=0}^N$ be the associated state sequence. Thus, if $x \in \mathbb{R}^d$ is expressed as a frame expansion with respect to F_N, and if this expansion is quantized by the first order $\Sigma\Delta$ scheme, then the resulting quantized expansion is

$$\widetilde{x}_N = \frac{d}{N} \sum_{n=1}^{N} q_n^N e_n^N.$$

Let us begin by rewriting the approximation error in a slightly more revealing form than in Section III-B. Starting with (17), specifiying $u_0^N = 0$, and specializing to the tight frame case where $S^{-1} = \frac{d}{N} I$, we have

$$\begin{aligned} x - \widetilde{x}_N &= \frac{d}{N} \left(\sum_{n=1}^{N-1} u_n^N (e_n^N - e_{n+1}^N) + u_N^N e_N^N \right) \\ &= \frac{d}{N} \left(\sum_{n=1}^{N-2} v_n^N (f_n^N - f_{n+1}^N) + v_{N-1}^N f_{N-1}^N + u_N^N e_N^N \right), \end{aligned} \tag{26}$$

where we have defined

$$f_n^N = e_n^N - e_{n+1}^N, \quad v_n^N = \sum_{j=1}^{n} u_j^N, \quad \text{and} \quad v_0^N = 0. \tag{27}$$

When working with the approximation error written as (26), the main step towards finding improved upper error bounds, as well as lower bounds, for $||x - \widetilde{x}||$, is to find a good estimate for $|v_n|$.

Let $\mathcal{B}_\Omega$ be the class of Ω-*bandlimited functions* consisting of all functions in $L^\infty(\mathbb{R})$ whose Fourier transforms (as distributions) are supported in $[-\Omega, \Omega]$. We shall work with the Fourier transform which is formally defined by $\widehat{f}(\gamma) = \int f(t) e^{-2\pi i t \gamma} dt$. By the Paley-Wiener theorem, elements of $\mathcal{B}_\Omega$ are restrictions of entire functions to the real line.

Definition V.1. *Let $f \in \mathcal{B}_\Omega$ and let $\{z_j\}_{j=1}^{n^*}$ be the finite set of zeros of f' contained in $[0,1]$. We say that $f \in \mathcal{M}_\Omega$ if $f' \in L^\infty(\mathbb{R})$, and if*

$$\forall j = 1, \cdots, n^*, \quad f''(z_j) \neq 0.$$

For simplicity and to avoid having to keep track of too many different constants, we shall use the notation $A \lesssim B$ to mean that there exists an absolute constant $C > 0$ such that $A \leq CB$. The following theorem relies on the uniform distribution techniques utilized by Sinan Güntürk in [12]. We briefly collect the necessary background on discrepancy and uniform distribution in Appendix I.

Theorem V.2. *Let $\{F_N\}_{N=d}^{\infty}$ be a family of normalized tight frames for $\mathbb{R}^d$, with $F_N = \{e_n^N\}_{n=1}^N$. Suppose $x \in \mathbb{R}^d$ satisfies $||x|| \leq (K - 1/2)\delta$, and let $\{x_n^N\}_{n=1}^N$ be the sequence of frame coefficients of x with respect to F_N. If, for some $\Omega > 0$, there exists $h \in \mathcal{M}_\Omega$ such that*

$$\forall N \text{ and } 1 \leq n \leq N, \quad x_n^N = h(n/N),$$

and if N is sufficiently large, then

$$|v_n^N| \lesssim \delta \left(\frac{n}{N^{1/4}} + N^{3/4} \log N \right) \lesssim \delta N^{3/4} \log N. \tag{28}$$

The implicit constants are independent of N and δ, but they do depend on x and hence h. The value of what constitutes a sufficiently large N depends on δ.

Proof: Let u_n^N be the state variable of the $\Sigma\Delta$ scheme and define $\widetilde{u}_n^N = u_n^N / \delta$. By the definition of v_n^N (see (27)), and by applying Koksma's inequality (see Appendix I), one has

$$\begin{aligned} |v_j^N| &= \delta \left| \sum_{n=1}^{j} \widetilde{u}_n^N \right| = j\delta \left| \frac{1}{j} \sum_{n=1}^{j} \widetilde{u}_n^N - \int_{-1/2}^{1/2} y \; dy \right| \\ &\leq j\delta \; \mathrm{Var}(x) \; \mathrm{Disc}(\{\widetilde{u}_n^N\}_{n=1}^j), \end{aligned} \tag{29}$$

where $\mathrm{Disc}(\cdot)$ denotes the *discrepancy* of a sequence as defined in Appendix I. Therefore, we need to estimate $D_j^N = \mathrm{Disc}(\{\widetilde{u}_n^N\}_{n=1}^j)$. Using the Erdös-Turán inequality (see Appendix I),

$$\forall K, \quad D_j^N \leq \frac{1}{K} + \frac{1}{j} \sum_{k=1}^{K} \frac{1}{k} \left| \sum_{n=1}^{j} e^{2\pi i k \widetilde{u}_n^N} \right|, \tag{30}$$

we see that it suffices for us to estimate $\left| \sum_{n=1}^{j} e^{2\pi i k \widetilde{u}_n^N} \right|$.

By Proposition 1 in [12], for each N there exists an analytic function $X_N \in \mathcal{B}_\Omega$ such that

$$u_n^N = X_N(n) \quad \text{modulo } [-\delta/2, \delta/2], \tag{31}$$

and

$$|X_N'(t) - h(t/N)| \lesssim 1/N. \tag{32}$$

Bernstein's inequality gives

$$|X_N''(t) - \frac{1}{N} h'(t/N)| \lesssim 1/N^2. \tag{33}$$

By hypothesis there exists $h \in \mathcal{M}_\Omega$ such that $x_n^N = h(n/N)$. Let $\{z_j\}_{n=1}^{n^*}$ be the set of zeros of h in $[0,1]$, and let $0 < \alpha < 1$ be a fixed constant to be specified later. Define the intervals I_j and J_j by

$$\forall j = 1, \cdots, n^*, \quad I_j = [Nz_j - N^\alpha, Nz_j + N^\alpha],$$
$$\forall j = 1, \cdots, n^* - 1, \quad J_j = [Nz_j + N^\alpha, Nz_{j+1} - N^\alpha],$$

and

$$J_0 = [1, Nz_1 - N^\alpha] \text{ and } J_{n^*} = [Nz_{n^*} + N^\alpha, N].$$

In the case where either 0 or 1 is a zero of h', one no longer needs the corresponding endpoint interval, J_i, but needs to modify the corresponding interval I_j to have 0 or 1 as its appropriate endpoint. Note that if N is sufficiently large then

$$J_0 \cup I_1 \cup J_1 \cup \cdots \cup I_{n_z} \cup J_{n_z} = [1, N].$$

It follows from the properties of $h \in \mathcal{M}_\Omega$ that if N is sufficiently large then

$$\forall n \in \mathbb{N} \cap J_j, \quad \frac{1}{N^{1-\alpha}} = \frac{N^\alpha}{N} \lesssim |h'(n/N)|.$$

Thus, by (33) we have that

$$\forall n \in \mathbb{N} \cap J_j, \quad \frac{k}{\delta N^{2-\alpha}} \lesssim |\frac{k}{\delta} X_N''(n)|. \tag{34}$$

Also, since $h \in \mathcal{M}_\Omega \subseteq L^\infty(\mathbb{R})$, and by (32), we obtain

$$\forall n \in \mathbb{N} \cap J_j, \quad |\frac{k}{\delta} X_N'(n)| \ \lesssim \frac{k}{\delta}. \tag{35}$$

Using (34), (35), Theorem 2.7 of [32], and since $0 < \delta < 1$, we have that for $1 \leq k$

$$\begin{aligned} \left| \sum_{n \in \mathbb{N} \cap J_j} e^{2\pi i k \widetilde{u}_n^N} \right| &= \left| \sum_{n \in \mathbb{N} \cap J_j} e^{2\pi i (k/\delta) X_N(n)} \right| \\ &\lesssim (2k/\delta + 2) \left(\frac{\delta^{1/2} N^{1-\frac{\alpha}{2}}}{k^{1/2}} + 1 \right) \\ &\lesssim \frac{k^{1/2} N^{1-\frac{\alpha}{2}}}{\delta^{1/2}} + \frac{k}{\delta}. \end{aligned}$$

Also, we have the trivial estimate

$$\left| \sum_{n \in \mathbb{N} \cap I_j} e^{2\pi i k \widetilde{u}_n^N} \right| \leq 2N^\alpha.$$

Thus,

$$\left| \sum_{n=1}^{j} e^{2\pi i k \widetilde{u}_n^N} \right| \lesssim N^\alpha + \frac{k^{1/2} N^{1-\frac{\alpha}{2}}}{\delta^{1/2}} + \frac{k}{\delta}.$$

Set $\alpha = 3/4$ and $K = N^{1/4}$. By (30) we have that if N is sufficiently large compared to δ then

$$\begin{aligned} D_j^N &\leq \frac{1}{K} + \frac{N^\alpha \log(K)}{j} + \frac{K^{1/2} N^{1-\frac{\alpha}{2}}}{\delta^{1/2} j} + \frac{K}{\delta j} \\ &\lesssim \frac{1}{N^{1/4}} + \frac{N^{3/4} \log(N)}{j} + \frac{N^{3/4}}{\delta^{1/2} j} + \frac{N^{1/4}}{\delta j} \\ &\lesssim \frac{1}{N^{1/4}} + \frac{N^{3/4} \log(N)}{j}. \end{aligned}$$

Thus by (29) we have

$$|v_n^N| \leq \frac{\delta n}{N^{1/4}} + \delta N^{3/4} \log N \lesssim \delta N^{3/4} \log N,$$

and the proof is complete.

Combining Theorem V.2 and (26) gives the following improved error estimate. Although this estimate guarantees approximation on the order of $\frac{\log N}{N^{5/4}}$ for even N, it is important to emphasize that *the implicit constants depend on x*. For comparison, note that Corollary III.8 only bounds the error by the order of $\frac{1}{N}$, but has explicit constants independent of x.

Corollary V.3. *Let $\{F_N\}_{N=d}^\infty$ be a family of normalized tight frames for $\mathbb{R}^d$, for which each $F_N = \{e_n^N\}_{n=1}^N$ satisfies the zero sum condition (18). Let $x \in \mathbb{R}^d$ satisfy $||x|| \leq (K - 1/2)\delta$, let $\{x_n^N\}_{n=1}^N$ be the frame coefficients of x with respect to F_N, and suppose there exists $h \in \mathcal{M}_\Omega, \Omega > 0$, such that*

$$\forall N \text{ and } 1 \leq n \leq N, \quad x_n^N = h(n/N).$$

Additionally, suppose that $f_n^N = e_n^N - e_{n+1}^N$ satisfies

$$\forall N, \ n = 1, \cdots, N, \ ||f_n^N|| \lesssim \frac{1}{N} \text{ and } \ ||f_n^N - f_{n+1}^N|| \lesssim \frac{1}{N^2},$$

and set $u_0^N = 0$ in (14).

If N is even and sufficiently large we have

$$||x - \widetilde{x}_N|| \lesssim \frac{\delta \log N}{N^{5/4}}$$

If N is odd and sufficiently large we have

$$\frac{\delta}{N} \lesssim ||x - \widetilde{x}_N|| \leq \frac{\delta d}{2N} (\sigma(F_N, p_N) + 1).$$

The implicit constants are independent of δ and N, but do depend on x, and hence h.

Proof: By Theorem V.2,

$$\left|\left| \frac{2}{N} \left(\sum_{n=1}^{N-2} v_n^N (f_n^N - f_{n+1}^N) + v_{N-1}^N f_{N-1}^N \right) \right|\right| \lesssim \frac{\delta \log N}{N^{5/4}}. \tag{36}$$

Thus, by Theorem III.7, (36), and (26), N being even implies

$$||x - \widetilde{x}_N|| \lesssim \frac{\delta \log N}{N^{5/4}}.$$

If N is odd, then by Theorem III.7, (26), and (36) we have

$$\frac{\delta}{N} = \frac{2 \ |u_N^N| \ ||e_N^N||}{N} \lesssim ||x - \widetilde{x}_N|| + \frac{\delta \log N}{N^{5/4}}.$$

Combining this with (22) completes the proof.

Applying Corollary V.3 to the quantization of frame expansions given by the roots of unity explains the different error behavior for even and odd N seen in Figure 1.

Example V.4 (Refined estimates for R_N). *Let $R_N = \{e_n^N\}_{n=1}^N$ be as in Example IV.1, i.e., R_N is the normalized tight frame for $\mathbb{R}^2$ given by the Nth roots of unity. Suppose $x \in \mathbb{R}^2$, $0 < ||x|| \leq (K - 1/2)\delta$, and that N is sufficiently large with respect to δ. The frame coefficients of $x = (a, b) \in \mathbb{R}^2$ with respect to R_N are given by $\{x_n^N\}_{n=1}^N = \{h(n/N)\}_{n=1}^N$, where $h(t) = a\cos(2\pi t) + b\sin(2\pi t)$.*

It is straighforward to show that $f_n^N = e_n^N - e_{n+1}^N$ satisfies

$$||f_n^N|| \leq \frac{2\pi}{N} \quad \text{and} \quad ||f_n^N - f_{n+1}^N|| \leq \frac{(2\pi)^2}{N^2},$$

and that $h \in \mathcal{M}_1$. Therefore, by Corollary V.3 and (23), if N is even then

$$||x - \widetilde{x}|| \lesssim \frac{\delta \log N}{N^{5/4}},$$

and if N is odd then

$$\frac{\delta}{N} \lesssim ||x - \widetilde{x}|| \leq \frac{\delta(2\pi + 1)}{N}.$$

The implicit constants are independent of δ and N, but do depend on x.

It is sometimes also possible to apply Corollary V.3 to harmonic frames.

Example V.5 (Refined estimates for H_N^d). *Let the dimension d be even, and let $H_N^d = \{e_n^N\}_{n=1}^N$ be as in Example IV.2, i.e., H_N^d is an harmonic frame for $\mathbb{R}^d$. Suppose $x \in \mathbb{R}^d$, $0 < ||x|| \leq (K - 1/2)\delta$, and that N is sufficiently large with respect to δ. The frame coefficients of $x = (a_1, b_1, \cdots, a_{d/2}, b_{d/2}) \in \mathbb{R}^d$ with respect to H_N^d are given by $\{x_n^N\}_{n=1}^N = \{h(n/N)\}_{n=1}^N$, where*

$$h(t) = \sqrt{\frac{2}{d}} \left(\sum_{j=1}^{d/2} a_j \cos(2\pi jt) + \sum_{j=1}^{d/2} b_j \sin(2\pi jt) \right).$$

Figure 2 in Example IV.2 shows the approximation error when the point $x = (1/\pi, 1/50, \sqrt{3/17}, e^{-1}) \in \mathbb{R}^4$ is represented with the harmonic frames $\{H_N^4\}_{N=4}^\infty$ and quantized using the $\Sigma\Delta$ scheme. For this choice of x it is straightforward to verify that $h \in M_{d/2}$. A direct estimate also shows that $f_n^N = e_n^N - e_{n+1}^N$ satisfies

$$||f_n^N|| \lesssim \frac{1}{N} \quad \text{and} \quad ||f_n^N - f_{n+1}^N|| \lesssim \frac{1}{N^2}.$$

Therefore, by Corollary V.3 and (23), if N is even then

$$||x - \widetilde{x}|| \lesssim \frac{\delta \log N}{N^{5/4}},$$

and if N is odd then

$$\frac{\delta}{N} \lesssim ||x - \widetilde{x}|| \leq \frac{\delta d}{2N} \left(\frac{10\pi}{\sqrt{3}} + 1 \right).$$

The implicit constants are independent of δ and N, but do depend on x.

VI. Comparison of $\Sigma\Delta$ with PCM

In this section we shall compare the mean squared error (MSE) given by $\Sigma\Delta$ quantization of finite frame expansions with that given by PCM schemes. We shall show that the $\Sigma\Delta$ scheme gives better MSE estimates than PCM quantization when dealing with sufficiently redundant frames. Throughout this section, let $F_N = \{e_n^N\}_{n=1}^N$ be a family of normalized tight frames for $\mathbb{R}^d$, and let

$$x = \frac{d}{N} \sum_{n=1}^N x_n^N e_n^N \quad \text{and} \quad \widetilde{x}_N = \frac{d}{N} \sum_{n=1}^N q_n^N e_n^N$$

be corresponding frame expansions and quantized frame expansions, where $x_n^N = \langle x, e_n^N \rangle$ are the frame coefficients of $x \in \mathbb{R}^d$ with respect to F_N, and where q_n^N are quantized versions of x_n^N.

In Example II.5, we showed that if one uses the PCM scheme (8) to produce the quantized frame expansion $\widetilde{x}_N$, then under Bennett's white noise assumption the PCM scheme has mean squared error

$$MSE_{PCM} = \frac{d^2 \delta^2}{12N}. \tag{37}$$

However, as illustrated in Example II.6, this estimate is not rigorous since Bennett's white noise assumption is not mathematically justified and may in fact fail dramatically.

If one uses $\Sigma\Delta$ quantization to produce the quantized frame expansion $\widetilde{x}_N$, then one has the error estimate

$$||x - \widetilde{x}_N|| \leq \frac{\delta d}{2N} (\sigma(F, p) + 2) \tag{38}$$

given by Corollary III.5. Here p is a permutation of $\{1, \cdots, N\}$ which denotes the order in which the $\Sigma\Delta$ scheme is run. This immediately yields the following MSE estimate for the $\Sigma\Delta$ scheme.

Theorem VI.1. *Given the $\Sigma\Delta$ scheme of Definition III.1, let $F = \{e_n\}_{n=1}^N$ be a normalized tight frame for $\mathbb{R}^d$, and let p be a permutation of $\{1, 2, \cdots, N\}$. For each $x \in \mathbb{R}^d$ satisfying $||x|| \leq (K - 1/2)\delta$, $\widetilde{x}$ shall denote the corresponding quantized output of the $\Sigma\Delta$ scheme. Let $B \subseteq \{x \in \mathbb{R}^d : ||x|| \leq (K - 1/2)\delta\}$ and define the mean square error of the $\Sigma\Delta$ scheme over B by*

$$MSE_{\Sigma\Delta} = \int_B ||x - \widetilde{x}||^2 \, d\mu(x),$$

where μ is any probability measure on B. Then

$$MSE_{\Sigma\Delta} \leq \frac{\delta^2 d^2}{4N^2} (\sigma(F, p) + 2)^2.$$

Proof: Square (38) and integrate.

One may analogously derive MSE bounds from any of the error estimates in Section III-B; we shall examine this in the subsequent example. The above estimate is completely deterministic; namely, it does not depend on statistical assumptions such as the analysis for PCM using Bennett's white noise assumption.

In Section IV, we saw that it is possible to choose families of frames, $F_N = \{e_n^N\}_{n=1}^N$, for $\mathbb{R}^d$, and permutations $p = p_N$, such that the resulting frame variation $\sigma(F_N, p_N)$ is uniformly bounded. Whenever this is the case, Theorem VI.1 yields the MSE bound $MSE_{\Sigma\Delta} \lesssim 1/N^2$, which is better than the PCM bound (37) by one order of approximation. For example, if one quantizes harmonic frame expansions in their natural order, then, by (25), Theorem VI.1 gives $MSE_{\Sigma\Delta} \lesssim 1/N^2$. Thus, for the quantization of harmonic frame expansions one may summarize the difference between $\Sigma\Delta$ and PCM as

$$MSE_{\Sigma\Delta} \lesssim 1/N^2 \quad \text{and} \quad 1/N \lesssim MSE_{PCM} \lesssim 1/N.$$

This says that $\Sigma\Delta$ schemes utilize redundancy better than PCM.

Let us remark that for the class of *consistent reconstruction* schemes considered in [2], Goyal, Vetterli, and Nguyen bound the MSE from below by b/A^2, where b is some constant and $A = N/d$ is the redundancy of the frame. Thus, the MSE estimate derived in Theorem VI.1 for the $\Sigma\Delta$ scheme achieves this same optimal MSE order.

Returning to classical PCM (with linear reconstruction), it is important to note that although $MSE_{\Sigma\Delta} \lesssim 1/N^2$ is much better than $MSE_{PCM} \lesssim 1/N$ for large N, it is still possible to have $MSE_{PCM} \leq MSE_{\Sigma\Delta}$ if N is small, i.e., if the frame has low redundancy. For example, if the frame being quantized is an orthonormal basis, then PCM schemes certainly offer better MSE than $\Sigma\Delta$ since in this case there is an isometry between the frame coefficients and the signal they represent. Nonetheless, for sufficiently redundant frames $\Sigma\Delta$ schemes provide better MSE than PCM.

Example VI.2 (Normalized frames for $\mathbb{R}^2$). *In view of Theorem IV.4 it is easy to obtain uniform bounds for the frame variation of frames F for $\mathbb{R}^2$. In particular, one can always find a permutation p such that $\sigma(F, p) \leq 2\pi$.*

A simple comparison of the MSE error bounds for PCM and $\Sigma\Delta$ discussed above shows that the MSE corresponding to first-order $\Sigma\Delta$ quantizers is less than the MSE corresponding to PCM algorithms for normalized tight frames for $\mathbb{R}^2$ in the following cases when the redundancy A satisfies the specified inequalities:

- $A > 1.5(2\pi)^2 \approx 59$ *if the normalized tight frame for $\mathbb{R}^2$ has even length, is zero sum, is ordered as in Theorem IV.4, and we set $u_0 = 0$, see Corollary III.8,*
- $A > 1.5(2\pi + 1)^2 \approx 80$ *for any normalized tight frame for $\mathbb{R}^2$, as long as the frame elements are ordered as described in Theorem IV.4, and u_0 is chosen to be 0, see Corollary III.5,*
- $A > 1.5(2\pi + 2)^2 \approx 103$ *for any normalized tight frame for $\mathbb{R}^2$, as long as the frame elements are ordered as described in Theorem IV.4, see Corollary III.6;*

Figure 5 shows the MSE achieved by $2K$-level PCM algorithms and $2K$-level first-order $\Sigma\Delta$ quantizers with step size $\delta = 1/K$ for several values of K for normalized tight frames for $\mathbb{R}^2$ obtained by the Nth roots of unity. The plots suggest that if the frame bound is larger than approximately 10, the first-order $\Sigma\Delta$ quantizer outperforms PCM.

Example VI.3 (7th Roots of Unity). *Let $x = (1/3, 1/2) \in \mathbb{R}^2$, and let $R_7 = \{e_N\}_{N=1}^7$ be the normalized tight frame for $\mathbb{R}^2$ given by*

$$e_n = (\cos(2\pi n/7), \sin(2\pi n/7)), \quad n = 1, \cdots, 7.$$

The point x has the frame expansion

$$x = \frac{2}{7}\sum_{n=1}^{7} x_n e_n, \qquad x_n = \langle x, e_n\rangle.$$

One may compute that

$$(x_1, x_2, x_3, x_4, x_5, x_6, x_7) \approx (0.5987, 0.4133, -0.0834, \\ -0.5173, -0.5616, -0.1831, 0.3333).$$

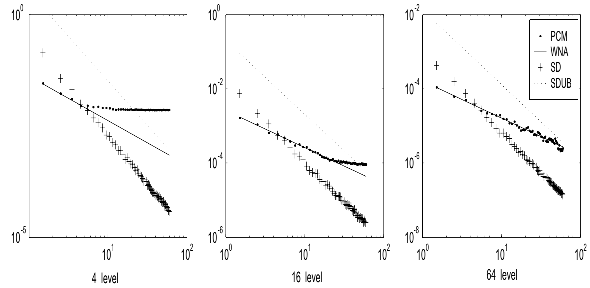

Fig. 5. Comparison of the MSE for $2K$-level PCM algorithms and $2K$-level first-order $\Sigma\Delta$ quantizers with step size $\delta = 1/K$. Frame expansions of 100 randomly selected points in $\mathbb{R}^2$ for frames obtained by the Nth roots of unity were quantized. In the figure legend PCM and SD correspond to the MSE for PCM and the MSE for first-order $\Sigma\Delta$ obtained experimentally, respectively. In the legend, the bound on the MSE for PCM, computed with white noise assumption, is denoted by WNA. Finally, SDWN in the legend stands for the MSE bound for $\Sigma\Delta$ that we would obtain if the approximation error was uniformly distributed between 0 and the upper bound in the odd case of (24).

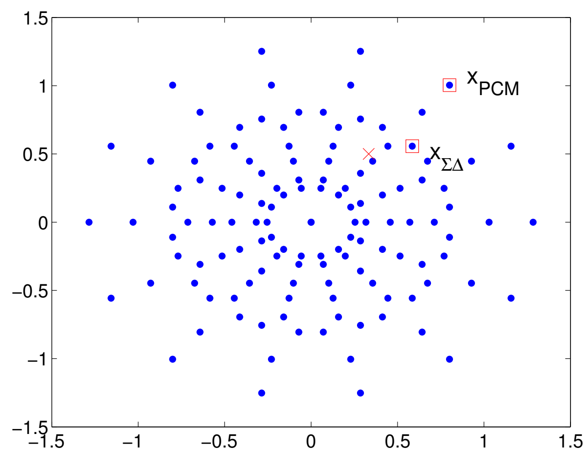

Fig. 6. The elements of Γ from Example VI.3 are denoted by solid dots, and the point $x = (1/3, 1/2)$ is denoted by '$\times$'. Note that $x \notin \Gamma$. '$X_{\Sigma\Delta}$' is the quantized point in Γ obtained using 1st order $\Sigma\Delta$ quantization, and 'X_{PCM}' is the quantized point in Γ obtained by PCM quantization.

If we consider the 1-bit alphabet $\mathcal{A}_1^2 = \{-1, 1\}$ then the quantization problem is to replace x by an element of

$$\Gamma = \{\frac{2}{7}\sum_{n=1}^{7} q_n e_n : q_n \in \mathcal{A}_1^2\}.$$

Figure 6 shows the elements of Γ denoted by solid dots, and shows the point x denoted by an '$\times$'. Note that $x \notin \Gamma$.

The 1st order $\Sigma\Delta$ scheme with 2-level alphabet $\mathcal{A}_1^2$ and natural ordering p quantizes x by $x_{\Sigma\Delta} \approx (.5854, 5571) \in \Gamma$. This corresponds to replacing $(x_1, x_2, x_3, x_4, x_5, x_6, x_7)$ by

$$(q_1, q_2, q_3, q_4, q_5, q_6, q_7) = (1, 1, -1, -1, -1, 1, -1).$$

The 2-level PCM scheme quantizes x by $x_{PCM} \approx (.8006, 1.0039) \in \Gamma$. This corresponds to replacing $(x_1, x_2, x_3, x_4, x_5, x_6, x_7)$ by

$$(q_1, q_2, q_3, q_4, q_5, q_6, q_7) = (1, 1, -1, -1, -1, -1, 1).$$

The points $x_{\Sigma\Delta}$ and x_{PCM} are shown in Figure 6 and it is visually clear that $||x - x_{\Sigma\Delta}|| < ||x - x_{PCM}||$.

The sets Γ corresponding to more general frames and alphabets than in Example VI.3 possess many interesting properties. This is a direction of ongoing work of the authors together with Yang Wang.

VII. Conclusion

We have introduced the K-level $\Sigma\Delta$ scheme with stepsize δ as a technique for quantizing finite frame expansions for $\mathbb{R}^d$. In Section III, we have proven that if F is a normalized tight frame for $\mathbb{R}^d$ of cardinality N, and $x \in \mathbb{R}^d$, then the K-level $\Sigma\Delta$ scheme with stepsize δ has approximation error

$$||x - \widetilde{x}|| \leq \frac{\delta d}{2N}(\sigma(F,p) + 2),$$

where the frame variation $\sigma(F,p)$ depends only on the frame F and the order p in which frame coefficients are quantized. As a corollary, for harmonic frames $H_N^d = \{e_n\}_{n=1}^N$ for $\mathbb{R}^d$ this gives the approximation error estimate

$$||x - \widetilde{x}|| \leq \frac{\delta d}{2N}(2\pi(d+1) + 1).$$

In Section V we showed that there are certain cases where the above error bounds can be improved to

$$||x - \widetilde{x}|| \lesssim 1/N^{\frac{5}{4}},$$

where the implicit constant depends on x. Section VI compares mean square error (MSE) for $\Sigma\Delta$ schemes and PCM schemes. A main consequence of our approximation error estimates is that

$$MSE_{\Sigma\Delta} \lesssim 1/N^2 \quad \text{whereas} \quad 1/N \lesssim MSE_{PCM},$$

when linear reconstruction is used, see (38) and (37). This shows that first order $\Sigma\Delta$ schemes outperform the standard PCM scheme if the frame being quantized is sufficiently redundant. We have also shown that $\Sigma\Delta$ quantization with linear reconstruction achieves the same order $1/N^2$ MSE as PCM with consistent reconstruction.

Our error estimates for first order $\Sigma\Delta$ schemes make it reasonable to hope that second order $\Sigma\Delta$ schemes can perform even better. This is, in fact, the case, but the analysis of second order schemes becomes much more complicated and is considered separately in [30].

Appendix I
Discrepancy and uniform distribution

Let $\{u_n\}_{n=1}^N \subseteq [-1/2, 1/2)$, where $[-1/2, 1/2)$ is identified with the torus $\mathbb{T}$. The *discrepancy* of $\{u_n\}_{n=1}^N$ is defined by

$$\mathrm{Disc}(\{u_n\}_{n=1}^N) = \sup_{I \subset \mathbb{T}} \left| \frac{\#(\{u_n\}_{n=1}^N \cap I)}{N} - |I| \right|,$$

where the $\sup$ is taken over all subarcs I of $\mathbb{T}$.

The *Erdös-Turan inequality* allows one to estimate discrepancy in terms of exponential sums:

$$\forall K, \quad \mathrm{Disc}(\{u_n\}_{n=1}^j) \leq \frac{1}{K} + \frac{1}{j}\sum_{k=1}^K \frac{1}{k}\left|\sum_{n=1}^j e^{2\pi i k u_n}\right|.$$

Koksma's inequality states that for any function $f : [-1/2, 1/2) \to \mathbb{R}$ of bounded variation,

$$\left|\frac{1}{N}\sum_{n=1}^N f(u_n) - \int_{-1/2}^{1/2} f(t)dt\right| \leq \mathrm{Var}(f)\ \mathrm{Disc}(\{u_n\}_{n=1}^N).$$

Acknowledgment

The authors would like to thank Ingrid Daubechies, Sinan Güntürk, and Nguyen Thao for valuable discussions on the material. The authors also thank Götz Pfander for sharing insightful observations on finite frames.

References

[1] N. Thao and M. Vetterli, "Deterministic analysis of oversampled A/D conversion and decoding improvement based on consistent estimates," *IEEE Transactions on Signal Processing*, vol. 42, no. 3, pp. 519–531, March 1994.
[2] V. Goyal, M. Vetterli, and N. Thao, "Quantized overcomplete expansions in $\mathbb{R}^n$: Analysis, synthesis, and algorithms," *IEEE Transactions on Information Theory*, vol. 44, no. 1, pp. 16–31, January 1998.
[3] I. Daubechies, *Ten Lectures on Wavelets*. Philadelphia, PA: SIAM, 1992.
[4] J. Benedetto and O. Treiber, "Wavelet frames: Multiresolution analysis and extension principles," in *Wavelet Transforms and Time-Frequency Signal Analysis*, L. Debnath, Ed. Birkhäuser, 2001.
[5] J. Munch, "Noise reduction in tight Weyl-Heisenberg frames," *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 608–616, March 1992.
[6] I. Daubechies and R. DeVore, "Reconstructing a bandlimited function from very coarsely quantized data: A family of stable sigma-delta modulators of arbitrary order," *Annals of Mathematics*, vol. 158, no. 2, pp. 679–710, 2003.
[7] C. Güntürk, J. Lagarias, and V. Vaishampayan, "On the robustness of single loop sigma-delta modulation," *IEEE Transactions on Information Theory*, vol. 12, no. 1, pp. 63–79, January 2001.
[8] Ö. Yılmaz, "Stability analysis for several sigma-delta methods of coarse quantization of bandlimited functions," *Constructive Approximation*, vol. 18, pp. 599–623, 2002.
[9] ——, "Coarse quantization of highly redundant time-frequency representations of square-integrable functions," *Appl. Comput. Harmon. Anal.*, vol. 14, pp. 107–132, 2003.
[10] V. Goyal, J. Kovačević, and J. Kelner, "Quantized frame expansions with erasures," *Appl. Comput. Harmon. Anal.*, vol. 10, pp. 203–233, 2001.
[11] T. Strohmer and R. W. Heath Jr., "Grassmannian frames with applications to coding and communcations," *Appl. Comput. Harmon. Anal.*, vol. 14, no. 3, pp. 257–275, 2003.
[12] C. Güntürk, "Approximating a bandlimited function using very coarsely quantized data: Improved error estimates in sigma-delta modulation," *J. Amer. Math. Soc.*, to appear.
[13] W. Chen and B. Han, "Improving the accuracy estimate for the first order sigma-delta modulator," *J. Amer. Math. Soc.*, submitted in 2003.
[14] V. Goyal, J. Kovačević, and M. Vetterli, "Quantized frame expansions as source-channel codes for erasure channels," in *Proc. IEEE Data Compression Conference*, 1999, pp. 326–335.
[15] G. Rath and C. Guillemot, "Syndrome decoding and performance analysis of DFT codes with bursty erasures," in *Proc. Data Compression Conference (DCC)*, 2002, pp. 282–291.
[16] P. Casazza and J. Kovačević, "Equal-norm tight frames with erasures," *Advances in Computational Mathematics*, vol. 18, no. 2/4, pp. 387–430, February 2003.
[17] G. Rath and C. Guillemot, "Recent advances in DFT codes based on quantized finite frames expansions for ereasure channels," *Preprint*, 2003.
[18] V. Goyal, J. Kovačević, and M. Vetterli, "Multiple description transform coding: Robustness to erasures using tight frame expansions," in *Proc. International Symposium on Information Theory (ISIT)*, 1998, pp. 326–335.
[19] B. Hochwald, T. Marzetta, T. Richardson, W. Sweldens, and R. Urbanke, "Systematic design of unitary space-time constellations," *IEEE Transactions on Information Theory*, vol. 46, no. 6, pp. 1962–1973, 2000.

[20] Y. Eldar and G. Forney, "Optimal tight frames and quantum measurement," *IEEE Transactions on Information Theory*, vol. 48, no. 3, pp. 599–610, March 2002.

[21] W. Bennett, "Spectra of quantized signals," *Bell Syst.Tech.J.*, vol. 27, pp. 446–472, July 1948.

[22] R. Duffin and A. Schaeffer, "A class of nonharmonic Fourier series," *Trans. Amer. Math. Soc.*, vol. 72, pp. 341–366, 1952.

[23] J. J. Benedetto and M. W. Frazier, Eds., *Wavelets: Mathematics and Applications*. Boca Raton, FL: CRC Press, 1994.

[24] O. Christensen, *An Introduction to Frames and Riesz Bases*. Boston, MA: Birkhäuser, 2003.

[25] G. Zimmermann, "Normalized tight frames in finite dimensions," in *Recent Progress in Multivariate Approximation*, K. Jetter, W. Haussmann, and M. Reimer, Eds. Birkhäuser, 2001.

[26] J. Benedetto and M. Fickus, "Finite normalized tight frames," *Advances in Computational Mathematics*, vol. 18, no. 2/4, pp. 357–385, February 2003.

[27] J. Candy and G. Temes, Eds., *Oversampling Delta-Sigma Data Converters*. IEEE Press, 1992.

[28] R. Gray, "Quantization noise spectra," *IEEE Transactions on Information Theory*, vol. 36, no. 6, pp. 1220–1244, November 1990.

[29] S. Norsworthy, R.Schreier, and G. Temes, Eds., *Delta-Sigma Data Converters*. IEEE Press, 1997.

[30] J. Benedetto, A. Powell, and Ö. Yılmaz, "Second order sigma-delta ($\Sigma\Delta$) quantization of finite frame expansions," *Preprint*, 2004.

[31] E. Janssen and D. Reefman, "Super-Audio CD: an introduction," *IEEE Signal Processing Magazine*, vol. 20, no. 4, pp. 83–90, July 2003.

[32] L. Kuipers and H. Niederreiter, *Uniform Distribution of Sequences*. New York: Wiley-Interscience, 1974.