Second order Sigma-Delta ($\Sigma\Delta$) quantization of finite frame expansions[⋆]

John J. Benedetto[*]

Department of Mathematics, University of Maryland, College Park, MD 20742

Alexander M. Powell

Program in Applied and Computational Mathematics, Princeton University, Princeton, NJ 08544

Özgür Yılmaz

Department of Mathematics, The University of British Columbia, Vancouver, B.C. Canada V6T 1Z2

Abstract

The second order Sigma-Delta ($\Sigma\Delta$) scheme with linear quantization rule is analyzed for quantizing finite unit-norm tight frame expansions for $\mathbb{R}^d$. Approximation error estimates are derived, and it is shown that for certain choices of frames the quantization error is of order $1/N^2$, where N is the frame size. However, in contrast to the setting of bandlimited functions there are many situations where the second order scheme only gives approximation error of order $1/N$. For example, this is the case when quantizing harmonic frames of odd length in even dimensions. An important component of the error analysis involves extending existing stability results to yield smaller invariant sets for the linear second order $\Sigma\Delta$ scheme.

Key words: Quantization, Second order Sigma-Delta schemes, Finite frames, Stability.

[⋆] This work was supported in part by NSF DMS Grant 0139759, ONR Grant N000140210398, and NSF DMS Grant 0219233.

[*] Corresponding author.

1 Introduction

In many data driven applications it is important to find a digital signal representation which is well adapted to tasks such as storage, processing, transmission, and recovery. Given a signal x of interest, one often begins by expanding x over an at most countable dictionary $\{e_n\}_{n\in\Lambda}$ to obtain an *atomic decomposition*,

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

where the coefficients x_n are real or complex numbers. Such an expansion is *redundant* if the choice of x_n in (1) is not unique. Although (1) is a discrete representation, it is certainly not "digital" since the coefficient sequence $\{x_n\}_{n\in\Lambda}$ is real or complex valued. The intrinsically lossy process of reducing the continuous range of this sequence to a discrete, preferably finite, set $\mathcal{A}$, is called *quantization.* A scheme that maps the real or complex valued coefficients x_n of (1) to $q_n \in \mathcal{A}$ is said to be a *quantization scheme.* Equivalently, the map $Q : x \longrightarrow \tilde{x} = \sum_{n\in\Lambda} q_n e_n$ is called a *quantizer.* The pointwise performance of a quantizer is reflected by the approximation error $\|x - Qx\|$ where $\|\cdot\|$ is a suitable norm. One is usually constrained by the available "bit budget", which in turn restricts the cardinality of the *quantization alphabet* $\mathcal{A}$ as well as the redundancy of the atomic decomposition (1). It is a challenging problem to distribute the available bits appropriately so that $\mathcal{A}$ has a sufficient number of elements to ensure the precision of the approximation *and* to ensure that the expansion is redundant enough for a robust and numerically stable implementation. Furthermore, when the expansion is redundant, finding a good quantizer with a given alphabet $\mathcal{A}$ is also a non-trivial problem. These problems, which we shall refer to as the *quantization problem*, arise in many different applied settings.

A fundamental example of quantization in signal processing is the process of *analog to digital (A/D) conversion.* There the signal space of interest consists of bandlimited functions. When a bandlimited function, f, is uniformly sampled at rate λ at or above the Nyquist rate, it can be fully reconstructed from its samples in the form of a sampling expansion,

$$f(t) = \frac{1}{\lambda}\sum_{n\in\mathbb{Z}} f(\frac{n}{\lambda})\varphi(t - \frac{n}{\lambda}), \tag{2}$$

where φ is an appropriate sampling kernel. Sampling expansions are a type of atomic decomposition, where the dictionary elements are translates of the sampling kernel, and the coefficients are the sample values. The A/D conversion problem is to replace the analog samples $f(\frac{n}{\lambda})$ by quantized coefficients q_n so that the resulting quantized expansion is a good approximation to the original signal. Since one usually oversamples in practice, i.e., takes λ strictly greater than the Nyquist rate, the sampling expansions are generally redundant. More information on sampling theorems and analog to digital conversion can be found in [1–3].

Another important example of quantization arises in image processing in the setting of *digital halftoning*, [4,5]. There the signal space consists of all digital greyscale images at a fixed resolution such as 512×512. One may think of images in terms of non-redundant atomic decompositions, where the dictionary elements are pixels, and the pixel coefficients are the greyscale intensities at the pixels. The halftoning problem is to print the color or greyscale image using only black or white "dots". In this example, the original pixel coefficients are already discrete (e.g., 256 level greyscale), but the printer requires a representation using an even smaller set, namely black or white. The practical halftoning methods most commonly used in printers, such as dithering and error diffusion, achieve remarkably good image quality.

In certain settings it is equally natural to view the quantization problem as a coding problem. For example, Goyal, Kovačević, Kelner, and Vetterli [6,7], cf., [8], propose using *finite tight frames* for $\mathbb{R}^d$ to transmit data over *erasure channels*. Given a signal $x \in \mathbb{R}^d$ which one wishes to transmit, one computes its coefficients with respect to a finite tight frame for $\mathbb{R}^d$, and transmits the corresponding coefficients. It is not possible to send the coefficients with infinite precision, so one must decide on a robust way to code or quantize the coefficients for transmission. In an erasure channel, errors are modeled by the loss, i.e., erasure, of certain transmitted coefficients. Redundancy is especially important in such applications since it provides resistance to data loss. In particular, it has been shown that the redundancy of frames can be used to "mitigate the effect of the losses in packet-based communication systems"[9], cf., [10].

Note that one can consider some A/D conversion problems in the setting of frame theory. For example, when φ is the $\mathrm{sinc}(\cdot)$ sampling kernel, then the sampling expansion (2) is in fact a redundant tight frame expansion when $\lambda > 1$ because the set $\{\varphi(\cdot - \frac{n}{\lambda})\}$ is a tight frame for the space of square integrable bandlimited functions, and the sample values $f(\frac{n}{\lambda})$ are the corresponding frame coefficients. Thus, in this setting, an A/D conversion scheme quantizes the frame coefficients of the function f with respect to the frame $\{\varphi(\cdot - \frac{n}{\lambda})\}$.

2 Overview and main results

In this paper we shall focus on the quantization problem for certain finite atomic decompositions for $\mathbb{R}^d$, i.e., finite unit-norm tight frame expansions for $\mathbb{R}^d$. In particular, we shall analyze the performance of a second order Sigma-Delta $(\Sigma\Delta)$ scheme and show that it outperforms the standard quantization techniques in many settings. We begin by presenting the basic definitions and theorems on finite frames in Section 3.

In Section 4, we discuss the quantization problem for finite frame expansions. Section 4.1 presents the standard PCM technique of quantization and states the corresponding approximation error estimates. We emphasize that PCM is poorly suited to quantizing highly redundant frame expansions. First order Sigma-Delta ($\Sigma\Delta$) schemes offer a different approach which is better suited for quantizing redundant expansions than PCM. In Section 4.2, we review the basic first order $\Sigma\Delta$ scheme and error estimates derived in [11,12]. In Section 4.3 we discuss second order $\Sigma\Delta$ schemes with the aim of showing that they can be used to quantize effectively finite frame expansions. In fact, they will outperform both PCM and first order $\Sigma\Delta$ schemes in many settings.

Section 5 discusses the key property of stability for the second order $\Sigma\Delta$ quantizer defined in Section 4.3. Section 5.1 reviews the invariant set construction in [13], and Section 5.2 derives an improved version which will be needed for our subsequent error estimates. The improvement allows us to bound the invariant set inside $(-2, 2) \times \mathbb{R}$. This property, which is crucial for our error bounds, does not hold in [13].

In Section 6 we derive approximation error estimates for the 1-bit second order $\Sigma\Delta$ scheme with linear quantization rule which was introduced in Section 4.3. We introduce the notion of higher order frame variation in Section 6.1. This allows us to derive a general error bound in Section 6.2, and then to apply it to the quantization of specific classes of frame expansions, such as harmonic frames. For example, let $H_N^d = \{e_n\}_{n=0}^{N-1}$ be the harmonic frame for $\mathbb{R}^d$ with N elements, and assume that d is even. If N is even, then we prove that the second order $\Sigma\Delta$ scheme gives approximation error $||x - \widetilde{x}|| \lesssim 1/N^2$ for the Euclidean norm $\| \cdot \|$. However, if N is odd, we show that the approximation error satisfies $1/N \lesssim ||x - \widetilde{x}|| \lesssim 1/N$. The notation $A \lesssim B$ means that there exists an absolute constant C such that $A \leq CB$. This dichotomy between even and odd cases stands in unexpected contrast to the behavior of $\Sigma\Delta$ algorithms in other settings such as A/D conversion of bandlimited signals.

Section 7 is devoted to a hybrid PCM/$\Sigma\Delta$ scheme for multibit quantization of finite frame expansions. The hybrid scheme achieves the same asymptotic utilization of frame redundancy as the 1-bit second order scheme in the previous section, with the added benefit of multibit resolution.

3 Finite frames for $\mathbb{R}^d$

Finite frames are a natural model and tool for many applications. In addition to applications related to erasure channels, [6,7,9,8,10], the use of finite frames has also been proposed for generalized multiple description coding [14,15], for multiple-antenna code design [16], for formulating results on Welch bound inequality sequences [17], and for solving modified quantum detection problems [18].

Definition 3.1 *A set $\{e_n\}_{n=1}^N \subseteq \mathbb{R}^d$ of vectors is a finite frame for $\mathbb{R}^d$ if there exists $0 < A \leq B < \infty$ such that*

$$\forall x \in \mathbb{R}^d, \quad A||x||^2 \leq \sum_{n=1}^{N} |\langle x, e_n \rangle|^2 \leq B||x||^2, \tag{3}$$

where $||\cdot||$ denotes the Euclidean norm.

The constants A and B are called *frame bounds*, and the coefficients $\langle x, e_n \rangle$, $n = 1, \ldots, N$, are called the *frame coefficients* of x with respect to $\{e_n\}_{n=1}^N$. If the frame bounds are equal, $A = B$, then the frame is said to be *tight*. If all the frame elements satisfy $||e_n|| = 1$ then the frame is said to be *uniform* or *unit-norm*. There are several natural operators associated to a frame.

Definition 3.2 *Given a finite frame $\{e_n\}_{n=1}^N$ for $\mathbb{R}^d$ with frame bounds A and B. The* analysis operator *$F : \mathbb{R}^d \to l^2(\{1, \cdots, N\})$ is defined by $(Fx)_k = \langle x, e_k \rangle$, and the* synthesis operator *$F^* : l^2(\{1, \cdots, N\}) \to H$ is its adjoint defined by $F^*(\{c_n\}_{n=1}^N) = \sum_{n=1}^N c_n e_n$. The operator $S = F^*F$ is called the* frame operator*, and it satisfies*

$$AI \leq S \leq BI, \tag{4}$$

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

$$B^{-1}I \leq S^{-1} \leq A^{-1}I. \tag{5}$$

The following theorem shows how frames are used to give atomic decompositions.

Theorem 3.3 *Let $\{e_n\}_{n=1}^N$ be a frame for $\mathbb{R}^d$ with frame bounds A and B, and let S be the corresponding frame operator. Then $\{S^{-1}e_n\}_{n=1}^N$ is a frame for $\mathbb{R}^d$ with frame bounds B^{-1} and A^{-1}, and*

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

If the frame is tight with frame bound A, then both frame expansions in Theorem 3.3 reduce to

$$\forall x \in \mathbb{R}^d, \quad x = A^{-1} \sum_{n=1}^{N} \langle x, e_n \rangle e_n. \tag{6}$$

It is important to be able to construct useful classes of frames. Given a set $\{v_n\}_{n=1}^N$ of N vectors in $\mathbb{R}^d$ or $\mathbb{C}^d$, define the associated matrix M to be the $d \times N$ matrix whose columns are the v_n. The following lemma can be found in [19].

Lemma 3.4 *A set $\{v_n\}_{n=1}^N$ of vectors in $H = \mathbb{R}^d$ or $\mathbb{C}^d$ is a tight frame with frame bound A if and only if its matrix M satisfies $MM^* = AI_d$, where M^* is the conjugate transpose of M, and I_d is the $d \times d$ identity matrix.*

For the important case of *finite unit-norm tight frames* for $\mathbb{R}^d$ and $\mathbb{C}^d$, the frame bound A is N/d, where N is the cardinality of the frame [6,20,19]. A physical characterization of finite unit-norm tight frames is given in [21]. In contrast to the above lemma, note that general unstructured finite frames are easy to construct - one simply needs to span the whole space.

The simplest examples of unit-norm tight frames for $\mathbb{R}^2$ are given by the Nth roots of unity viewed as elements of $\mathbb{R}^2$. Namely, if

$$e_n^N = (\cos(2\pi n/N), \sin(2\pi n/N)),$$

then $E_N = \{e_n^N\}_{n=1}^N$ is a unit-norm tight frame for $\mathbb{R}^2$ with frame bound $N/2$.

The most natural examples of unit-norm tight frames for $\mathbb{R}^d, d > 2$, are the harmonic frames, e.g., see [6,19,20]. These frames are constructed using columns of the Fourier matrix. We follow the notation of [19], 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 is even let

$$\begin{aligned} e_j = \sqrt{\frac{2}{d}} \Big[&\cos\frac{2\pi j}{N}, \sin\frac{2\pi j}{N}, \cos\frac{2\pi 2j}{N}, \sin\frac{2\pi 2j}{N}, \cos\frac{2\pi 3j}{N}, \\ &\sin\frac{2\pi 3j}{N}, \cdots, \cos\frac{2\pi \frac{d}{2} j\pi}{N}, \sin\frac{2\pi \frac{d}{2} j\pi}{N} \Big] \end{aligned}$$

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

If d is odd let

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

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

It is shown in [19] that H_N^d, as defined above, is a unit-norm tight frame for $\mathbb{R}^d$. If d is even then H_N^d satisfies the zero sum condition $S = 0$, where $S = \sum_{n=0}^{N-1} e_n$, [12]. If d is odd then the frame is not zero sum, and $S = \frac{N}{\sqrt{d}}(1, 0, \cdots, 0)$.

4 Quantization of finite frame expansions

Let $F = \{e_n\}_{n=1}^N$ be a finite unit-norm tight frame for $\mathbb{R}^d$ and let x be in $\mathbb{R}^d$. We shall study how to quantize the frame expansion

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

In fact, we wish to replace the sequence $\{x_n\}_{n=1}^N$ of frame coefficients by a quantized sequence $\{q_n\}_{n=1}^N$, where q_n are chosen from a given quantization alphabet $\mathcal{A}$, such that

$$\widetilde{x} = \frac{d}{N} \sum_{n=1}^{N} q_n e_n \tag{7}$$

is close to x in the Euclidean norm, $|| \cdot ||$.

In any practical setting, the quantization alphabet will be a finite set. This, in turn, imposes a uniform bound M on the norm of the vectors x to be quantized. Note that in this case, the frame coefficients x_n of x satisfy $|x_n| \leq M$. Below, we shall specify the value of M whenever appropriate.

Let us also mention that while we shall only consider *linear reconstruction* as in (7), there do exist other more general, but more computationally complex, nonlinear reconstruction techniques, e.g., [22,23,20].

4.1 PCM quantization

Pulse Code Modulation (PCM) schemes are probably the simplest approach to quantizing finite frame expansions. Consider $x \in \mathbb{R}^d$, $\|x\| \leq 1$. The $2K$-level

PCM scheme with step size δ replaces each x_n with

$$q_n = Q(x_n) = \text{argmin}_{q \in \mathcal{A}_K^\delta} |x_n - q|, \tag{8}$$

where $A_K^\delta = \{(-K+1/2)\delta, (-K+3/2)\delta, \cdots, (K-1/2)\delta\}$. The function Q, defined as above, is called the K-level midrise uniform scalar quantizer with step size δ.

One can show that

$$\forall n, \quad |x_n| < K\delta \implies \sup_n |x_n - q_n| \leq \delta/2,$$

and that one has the basic error estimate

$$||x - \widetilde{x}|| = \frac{d}{N} || \sum_{n=1}^N (x_n - q_n) e_n || \leq \frac{\delta d}{2N} \sum_{n=1}^N ||e_n|| = \frac{\delta d}{2}.$$

While it is possible to decrease the error $||x - \widetilde{x}||$ by decreasing the step size δ, this error estimate does not utilize the redundancy of the frame. The following example highlights the inability of PCM to make good use of redundancy.

Example 4.1 *Let $E_N = \{e_n\}_{n=1}^N$ be the unit-norm tight frame for $\mathbb{R}^2$ given by the Nth roots of unity. The frame coefficients of $x = (0, b) \in \mathbb{R}^2, 0 < b < 1$ with respect to E_N are given by $x_n^N = \langle x, e_n^N \rangle = b \sin(2\pi n/N)$. If we use the 2-level PCM scheme with step size $\delta = 1$ to quantize the frame coefficients x_n^N then we obtain*

$$q_n^N = \begin{cases} 1/2 & \text{if } 1 \leq n < N/2, \\ -1/2 & \text{if } N/2 \leq n \leq N. \end{cases}$$

Using this, one can verify that, for $N \geq 2$,

$$\frac{1}{2\pi} \leq || \frac{2}{N} \sum_{n=1}^N q_n^N e_n^N ||.$$

This shows that the quantized expansion has its norm bounded away from zero independent of $0 < b < 1$ and N. Thus, if b is sufficiently small then, regardless of how redundant the frame is, one can not achieve arbitrarily small approximation error when this 1-bit PCM scheme is used to quantize the frame expansion for (0,b). In the next section we shall present a quantization scheme which does not suffer from this type of limitation, and, in fact, it is able to utilize frame redundancy much more efficiently than PCM.

Although the above example shows that PCM fails to take advantage of frame redundancy, at least for certain x in the unit ball of $\mathbb{R}^d$, there have been attempts to show that PCM utilizes redundancy on average. This approach considers the average approximation error corresponding to an ensemble of vectors. To that end one makes the hypothesis that the quantization error sequence $\{x_n - q_n\}_{n=1}^{N}$ can be modeled as a signal independent sequence of *i.i.d.* random variables with mean 0 and variance $\delta^2/12$. This is called Bennett's white noise assumption [24,20], and it yields a mean square error

$$MSE = E||x - \frac{d}{N}\sum_{n=1}^{N} q_n e_n||^2 = \frac{d^2\delta^2}{12N}.$$

The problem with this estimate is that the white noise assumption is not rigorously justified and is actually false in many simple circumstances, e.g., see [12]. Moreover, the MSE bound only decreases on the order of $1/N$; one can do better than this, e.g., [12].

Some existing approaches to finite frame quantization improve the PCM quantization error by using more advanced reconstruction strategies. For example, consistent reconstruction is one especially important class of nonlinear reconstruction, [22,23,20]. Other differently motivated approaches include [25,26]. The noise shaping approach of [26] was shown to be effective for subband coding applications and is related to the $\Sigma\Delta$ techniques which we describe in the following sections.

4.2 First order $\Sigma\Delta$ quantization

Sigma Delta ($\Sigma\Delta$) schemes were introduced in the engineering community for the purpose of coarsely quantizing sampling expansions in the setting of analog to digital conversion, e.g., see [27,28]. In particular, $\Sigma\Delta$ schemes were originally used to quantize a sampling expansion

$$f(t) = \frac{1}{\lambda}\sum_{n\in\mathbb{Z}} f(\frac{n}{\lambda})\varphi(t - \frac{n}{\lambda})$$

by

$$\widetilde{f}(t) = \frac{1}{\lambda}\sum_{n\in\mathbb{Z}} q_n\varphi(t - \frac{n}{\lambda}),$$

where each $q_n \in \{-1, 1\}$. Although $\Sigma\Delta$ schemes have been successfully implemented in practice for quite a while, and exhibit excellent empirical approximation error and robustness properties, they have only recently attracted the attention of the mathematical community, [1]. Work on $\Sigma\Delta$ quantization now ranges from practical circuit implementation and design [29], to mathematical analysis relying on harmonic analysis, dynamical systems, analytic number theory, and issues in tiling, [30–33].

In [12], first order $\Sigma\Delta$ schemes were used to quantize finite frame expansions, and it was shown that they offer excellent approximation error, and outperform the standard techniques for quantizing finite frame expansions. In particular, let $F = \{e_n\}_{n=1}^N$ be a finite unit-norm tight frame for $\mathbb{R}^d$, and let p be a permutation of $\{1, 2, \cdots, N\}$. Given $x \in \mathbb{R}^d$ with frame coefficients $\{x_n\}_{n=1}^N$, the first order $\Sigma\Delta$ scheme produces the quantized sequence $\{q_n\}_{n=1}^N$ by iterating

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

where $\{u_n\}_{n=0}^N$ is the *state sequence* with $u_0 = 0$, and Q is a K-level midrise uniform quantizer with step size δ, i.e., $Q(w) = \operatorname{argmin}_{q \in \mathcal{A}_K^\delta} |w - q|$, where $A_K^\delta = \{(-K + 1/2)\delta, (-K + 3/2)\delta, \cdots, (K - 1/2)\delta\}$. We shall refer to the permutation p as the *quantization ordering* since it denotes the order in which frame coefficients are entered into the $\Sigma\Delta$ algorithm.

The quantized sequence $\{q_n\}_{n=1}^N$ is used to reconstruct an approximation $\widetilde{x}$ as in (7). It was shown in [12] that the first order $\Sigma\Delta$ scheme gives the approximation error bound

$$||x - \widetilde{x}|| \leq \frac{\delta d}{2N} \left(\frac{\sigma(F, p)}{2} + 1 \right), \tag{9}$$

where $\sigma(F, p)$ is a quantity called the frame variation which depends on the frame and the quantization ordering. For certain infinite families of frames, such as harmonic frames, it is possible to find a uniform upper bound on the frame variation independent of the size N of the frame, and depending only on the dimension d. In view of this, (9) shows that the $\Sigma\Delta$ scheme is able to utilize the frame redundancy to achieve improved approximation error. In particular, (9) implies that the first order $\Sigma\Delta$ scheme satisfies the MSE bound,

$$MSE \leq \frac{\delta^2 d^2}{4N^2} \left(\frac{\sigma(F, p)}{2} + 1 \right)^2,$$

which is better than the bound for PCM achieved using the non-rigorous Bennett white noise assumption, provided the frame redundancy is sufficiently large.

Comparing the first order $\Sigma\Delta$ scheme with PCM, we see that, unlike the situation in Example 4.1, the error estimate (9) shows that the approximation error in $\Sigma\Delta$ quantization decreases as the frame redundancy increases.

4.3 Second order $\Sigma\Delta$ schemes

Although (9) gives an error estimate whose utilization of redundancy is of order $1/N$, it is natural to seek even better utilization of the frame redundancy. Higher order schemes make this possible.

Given a sequence $\{x_n\}_{n=1}^N$ of frame coefficients and a permutation p of $\{1, 2, \cdots, N\}$, the general form of a second order $\Sigma\Delta$ scheme is

$$\begin{aligned} u_n &= u_{n-1} + x_{p(n)} - q_n, \\ v_n &= u_{n-1} + v_{n-1} + x_{p(n)} - q_n, \\ q_n &= \mathrm{Q}(F(u_{n-1}, v_{n-1}, x_{p(n)})), \end{aligned} \tag{10}$$

where $u_0 = v_0 = 0$, $Q : \mathbb{R} \to \mathbb{R}$ is an appropriate quantizer, and $F : \mathbb{R}^3 \to \mathbb{R}$ is a specified quantization rule. Some choices of F from the literature (e.g., [29,1,34]) are

- $F(u, v, x) = u + \gamma v$ with $\gamma > 0$;
- $F(u, v, x) = u + x + M\mathrm{sign}(v)$ with $M > 0$;
- $F(u, v, x) = (6x - 7)/3 + (u + (x + 3)/2)^2 + 2(1 - x)v$.

In this paper we only consider the linear rule $F(u, v, x) = F(u, v) = u + \gamma v$, where $\gamma > 0$ is fixed. This scheme is important because it has a simple form that is well suited for implementation and because it has desirable stability properties, e.g., [13]. Until Section 7 we shall also restrict ourselves to the 1-bit case,

$$Q(w) = \frac{\delta}{2}\mathrm{sign}(w) = \begin{cases} \frac{\delta}{2}, & \text{if } w \geq 0, \\ -\frac{\delta}{2}, & \text{if } w < 0. \end{cases}$$

Thus, we shall consider the scheme

$$\begin{aligned} u_n &= u_{n-1} + x_{p(n)} - q_n, \\ v_n &= u_{n-1} + v_{n-1} + x_{p(n)} - q_n, \\ q_n &= \frac{\delta}{2}\mathrm{sign}(\ u_{n-1} + \gamma v_{n-1}\ ), \end{aligned} \tag{11}$$

for $n = 1, \cdots, N$, with initial states $u_0 = v_0 = 0$. We consider the 1-bit case because it is the simplest case to analyze, and because it allows us to build on the results in [13]. However, we also examine a multibit hybrid PCM/$\Sigma\Delta$ scheme in Section 7.

One surprising point of this paper is that $\Sigma\Delta$ schemes behave quite differently when used to quantize finite frame expansions than they do for their original purpose of quantizing sampling expansions for bandlimited functions. In particular, when a stable second order $\Sigma\Delta$ scheme is used to quantize A/D sampling expansions one has the approximation error estimate

$$||f - \widetilde{f}||_{L^\infty(\mathbb{R})} \lesssim \frac{1}{\lambda^2},$$

where λ is the sampling rate. By analogy, one might expect that when a second order $\Sigma\Delta$ scheme is used to quantize a finite frame expansion that one will have

$$||x - \widetilde{x}|| \lesssim \frac{1}{N^2}, \tag{12}$$

where N is the frame size, which is analogous to the sampling rate in that it determines the redundancy of the atomic decomposition. Here $\widetilde{x}$ is defined as in (7). We shall see that (12) is not true in general. We shall show that there are many circumstances where one is only able to achieve an approximation where the approximation error is of order $1/N$ as N tends to infinity. We shall also specify certain conditions under which we can ensure that the approximation error behaves asymptotically like $1/N^2$. A key issue will be that the finite nature of the problem for finite frame expansions gives rise to non-zero boundary terms in certain situations, and that these boundary terms may negatively affect error estimates.

5 Stability for the second order linear scheme

For any $\Sigma\Delta$ scheme to have potential use in practice it is crucial for it to be *stable*. In other words, given a bounded input sequence x_n, the state variables (u_n and v_n in the second order case) of the scheme should remain bounded. It is relatively simple to verify that the first order $\Sigma\Delta$ scheme is stable, but this is more complicated for second order $\Sigma\Delta$ schemes.

The approach to proving stability taken in [13] is to view the problem in terms of finding an invariant set for a certain mapping of $\mathbb{R}^2$ to $\mathbb{R}^2$. We say that a set $S \subseteq \mathbb{R}^d$ is an *invariant set* for a map $T : \mathbb{R}^d \to \mathbb{R}^d$ if $T(S) \subseteq S$. For simplicity we shall only present proofs for $Q(w) = \text{sign}(w)$; the extension to the case $Q(w) = \frac{\delta}{2}\text{sign}(w)$ is straightforward.

Following the presentation in [13], $0 \leq \alpha < 1$ will denote an upper bound on the absolute value of the input sequence of frame coefficients, i.e., $|x_n| \leq \alpha < 1$. Suppose $\gamma > 0$ is given so that the quantization rule $q_n = \text{sign}(u_{n-1} + \gamma v_{n-1})$ is defined. Let $\delta_n = |x_n - q_n|$, $\delta_- = 1 - \alpha$, and $\delta_+ = 1 + \alpha$, and note that $\delta_- \leq \delta_n \leq \delta_+$. We may now rewrite (11) as

$$(u_n, v_n) = \begin{cases} S_l^\delta(u_{n-1}, v_{n-1}) = (u_{n-1} - \delta_n, u_{n-1} + v_{n-1} - \delta_n), & \text{if } q_n = 1, \\ S_r^\delta(u_{n-1}, v_{n-1}) = (u_{n-1} + \delta_n, u_{n-1} + v_{n-1} + \delta_n), & \text{if } q_n = -1. \end{cases} \tag{13}$$

When convenient we simply write

$$(u_n, v_n) = S_\gamma(u_{n-1}, v_{n-1}, \delta_n).$$

It is important to keep in mind that the map S_γ is determined by the choice of parameter γ.

With this setup, given $0 \leq \alpha < 1$ and $\gamma > 0$, the stability problem is to find a set $R \subset \mathbb{R}^2$ such that if $\delta \in [\delta_-, \delta_+] = [1-\alpha, 1+\alpha]$ then

$$(u,v) \in R \implies S_\gamma(u,v,\delta) \in R. \tag{14}$$

5.1 *An invariant set for the linear scheme*

In this section we recall the invariant set construction of [13].

Given the parameters $0 \leq \alpha < 1$ (so that δ_- and δ_+ are defined) and $0 < C$. We define

$$B_1(u) = \begin{cases} -\frac{1}{2\delta_-}\left(u - \frac{\delta_-}{2}\right)^2 + \frac{\delta_-}{8} + C, & \text{if } u \geq 0, \\ -\frac{1}{2\delta_+}\left(u - \frac{\delta_+}{2}\right)^2 + \frac{\delta_+}{8} + C, & \text{if } u < 0, \end{cases} \tag{15}$$

and

$$B_2(u) = v \begin{cases} \frac{1}{2\delta_+}\left(u + \frac{\delta_+}{2}\right)^2 - \frac{\delta_+}{8} - C, & \text{if } u \geq 0. \\ \frac{1}{2\delta_-}\left(u + \frac{\delta_-}{2}\right)^2 - \frac{\delta_-}{8} - C, & \text{if } u < 0. \end{cases} \tag{16}$$

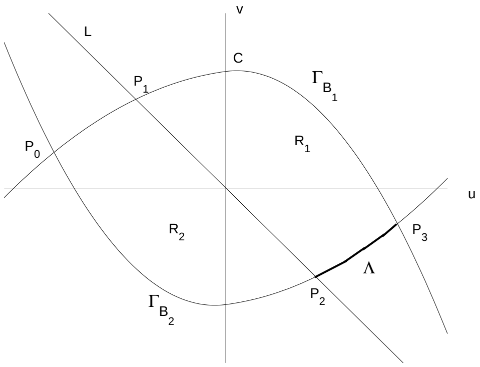

Fig. 1. The invariant set of [13]. Γ_{B_1} and Γ_{B_2} are the graphs of the functions B_1 and B_2 respectively. L is the line on which $F(u,v) = u + \gamma v = 0$.

Let $l(u) = -\frac{1}{\gamma}u$ be the equation of the line corresponding to $F(u,v) = u + \gamma v = 0$. Define

$$\begin{aligned} R_1 &= \{(u,v) : v \leq B_1(u), v \geq B_2(v), v \geq l(u)\}, \\ R_2 &= \{(u,v) : v \leq B_1(u), v \geq B_2(v), v < l(u)\}, \\ R &= R_1 \cup R_2. \end{aligned} \tag{17}$$

Thus, R is the region bounded between the graphs of B_1 and B_2. Note that R is fully determined by the two parameters α and C. [13] showed that for certain choices of the parameters α, γ, C, (14) holds. Figure 1 shows the graphs Γ_{B_1}

and Γ_{B_2}, of B_1 and B_2, respectively. Let $P_0 = (u_0, v_0)$ be the left intersection point of Γ_{B_1} and Γ_{B_2}, let $P_1 = (u_1, v_1)$ be the left intersection point of L and Γ_{B_1}, let $P_2 = (u_2, v_2)$ be the right intersection point of L and Γ_{B_2}, and let $P_3 = (u_3, v_3) = (-u_0, -v_0)$ be the right intersection point of Γ_{B_1} and Γ_{B_2}. Additionally, let Λ be the part of Γ_{B_2} which lies between P_2 and the right intersection point P_3 of Γ_{B_1} and Γ_{B_2}. One has

$$\begin{aligned} u_0 &= -[2C(1-\alpha^2)]^{\frac{1}{2}}, \\ v_0 &= B_1(u_0), \end{aligned}$$

and $(u_2, v_2) = (-u_1, -v_1)$.

The following lemma [13] shows that the region below Γ_{B_1} is invariant under S_l^δ, and that the region above Γ_{B_2} is invariant under S_r^δ.

Lemma 5.1 *If $\delta \in [\delta_-, \delta_+]$ then the region T_1 below the graph Γ_{B_1} of B_1 is invariant under the mapping*

$$S_l^\delta : (u, v) \longmapsto (u - \delta, u + v - \delta).$$

Likewise, the region T_2 above the graph Γ_{B_2} of B_2 is invariant under the mapping

$$S_r^\delta : (u, v) \longmapsto (u + \delta, u + v + \delta).$$

This invariance means that $S_l^\delta(T_1) \subseteq T_1$ and $S_r^\delta(T_2) \subseteq T_2$.

The next result [13] shows that the image of R_1 under S_l^δ stays above Γ_{B_2}, and analogously for R_2.

Theorem 5.2 *Let $P_1 = (u_1, v_1)$ be the intersection point of the line L defined by $F(u, v) = u + \gamma v = 0$ and let Γ_{B_1} be as shown in Figure 1. Suppose*

$$u_0 + \delta_+ \leq u_1 \leq -\delta_+. \tag{18}$$

Then $S_l^\delta(R_1) \subseteq R$ and $S_r^\delta(R_2) \subseteq R$ for any $\delta \in [\delta_-, \delta_+]$.

Combining the previous two results gives the following stability result [13].

Theorem 5.3 *If the parameters $0 \leq \alpha < 1$ and $0 < C, \gamma$ are chosen so that*

$$\delta \in [\delta_-, \delta_+] \text{ and } u_0 + \delta_+ \leq u_1 \leq -\delta_+,$$

then

$$(u, v) \in R \implies S_\gamma(u, v, \delta) \in R.$$

In particular, if $|x_n| \leq \alpha$ then the state variables of the second order $\Sigma\Delta$ scheme satisfy $(u_n, v_n) \in R$ for all n.

The error estimates which we derive in Section 6 will depend critically on being able to bound the invariant set R inside of $(-2,2)\times\mathbb{R}$. Unfortunately, the condition (18) prevents this from being the case. In particular, it was shown in [13] that the condition (18) only makes sense if $C \geq 2\frac{1+\alpha}{1-\alpha}$. This, in turn, implies that $u_0 = -[2C(1-\alpha^2)]^{\frac{1}{2}} \leq -2(1+\alpha^2) < -2$, and that $u_3 > 2$. Thus, the hypotheses of Theorem 5.2 make it impossible to bound R inside $(-2,2)\times\mathbb{R}$.

5.2 An improved stability theorem

To ensure that the invariant set stays inside $(-2,2)\times\mathbb{R}$ we must introduce weaker hypotheses than (18).

Theorem 5.4 *Let $P_1 = (u_1, v_1)$ be the intersection point of the line L defined by $F(u,v) = u + \gamma v = 0$ and let Γ_{B_1} be as shown in Figure 1. Suppose*

$$u_0 + \delta_+ \leq u_1 \tag{19}$$

and

$$u_2 + v_2 - \delta \geq B_2(u_2 - \delta), \quad \textit{for} \quad \delta = \delta_- \textit{ and } \delta = \delta_+. \tag{20}$$

Then $S_l^\delta(R_1) \subseteq R$ for any $\delta \in [\delta_-, \delta_+]$.

PROOF. By Lemma 5.1 it suffices to show that $S_l^\delta(R_1)$ lies above Γ_{B_2}. Using the simplifications and convexity arguments in the proof of Theorem 4 in [13], it suffices to show that $S_l^\delta(P_1)$, $S_l^\delta(P_2)$, and $S_l^\delta(\Lambda)$ lie above Γ_{B_2} for $\delta = \delta_-$ and $\delta = \delta_+$.

The conditions (19) and (20) respectively ensure that $S_l^\delta(P_1)$ and $S_l^\delta(P_2)$ lie above Γ_{B_2} for $\delta \in \{\delta_-, \delta_+\}$. Therefore it remains to show that $S_l^\delta(\Lambda)$ lies above Γ_{B_2} for $\delta \in \{\delta_-, \delta_+\}$. Since P_2 is the left endpoint of Λ, and since $S_l^\delta(P_2)$ lies above Γ_{B_2} for $\delta = \delta_-$ and $\delta = \delta_+$, it will suffice to show that the graph of $S_l^\delta(\Lambda)$ has a larger derivative than the corresponding portion of B_2, for $\delta \in \{\delta_-, \delta_+\}$.

Let $(u, B_2(u)) \in \Lambda$, where $u_2 \leq u \leq -u_0 = u_3$. By definition $S_l^\delta(u, B_2(u)) = (u - \delta, u + B_2(u) - \delta)$. So the image of Λ under S_l^δ is given by the graph of

$$f(u) = u + B_2(u+\delta), \quad \text{for } u_2 - \delta \leq u \leq -u_0 - \delta.$$

We want to show that

$$f'(u) = 1 + B_2'(u+\delta) \geq B_2'(u) \quad \text{on} \quad [u_2 - \delta, -u_0 - \delta]$$

for $\delta = \delta_-$ and $\delta = \delta_+$. From the definition of B_2 we have

$$B_2'(u) = \begin{cases} \frac{1}{\delta_+}(u + \frac{\delta_+}{2}), & \text{if } u \geq 0, \\ \frac{1}{\delta_-}(u + \frac{\delta_-}{2}), & \text{if } u < 0. \end{cases}$$

Since $0 \leq u_2$ we have $u + \delta \geq 0$ and $f'(u) = 1 + \frac{1}{\delta_+}(u + \delta + \frac{\delta_+}{2})$.

1. Let us first consider the case $\delta = \delta_+ = 1 + \alpha$. We have

$$f'(u) = 1 + \frac{1}{1+\alpha}(u + \frac{3}{2}(1+\alpha)) = \frac{5}{2} + \frac{u}{1+\alpha}, \quad u \in [u_2 - \delta_+, -u_0 - \delta_+].$$

If $u < 0$ then

$$B_2'(u) = \frac{1}{2} + \frac{u}{1-\alpha} < \frac{5}{2} + \frac{u}{1+\alpha} = f'(u).$$

If $0 \leq u$ then

$$B_2'(u) = \frac{1}{2} + \frac{u}{1+\alpha} < \frac{5}{2} + \frac{u}{1+\alpha} = f'(u).$$

Thus $f'(u) \geq B_2'(u)$, and we have that $S_l^{\delta_+}(\Lambda)$ lies above Γ_{B_2}.

2. Let us now consider the case $\delta = \delta_-$. We have

$$f'(u) = 1 + \frac{1}{1+\alpha}(u + 1 - \alpha + \frac{1+\alpha}{2}) = \frac{3}{2} + \frac{u}{1+\alpha} + \frac{1-\alpha}{1+\alpha}.$$

If $u < 0$ then

$$B_2'(u) = \frac{1}{2} + \frac{u}{1-\alpha} \leq \frac{1}{2} + \frac{u}{1+\alpha} < \frac{3}{2} + \frac{u}{1+\alpha} + \frac{1-\alpha}{1+\alpha} = f'(u).$$

If $u \geq 0$ then

$$B_2'(u) = \frac{1}{2} + \frac{u}{1+\alpha} \leq \frac{3}{2} + \frac{u}{1+\alpha} + \frac{1-\alpha}{1+\alpha} = f'(u).$$

Thus $f'(u) \geq B_2'(u)$, and we have that $S_l^{\delta_-}(\Lambda)$ lies above Γ_{B_2}. $\square$

A similar proof as above gives the analogous result for $S_r^\delta(R_2)$.

Theorem 5.5 *Let $P_1 = (u_1, v_1)$ be the intersection point of the line L defined by $F(u,v) = u + \gamma v = 0$ and let Γ_{B_1} be as shown in Figure 1. Suppose*

$$u_0 + \delta_+ \leq u_1 \tag{21}$$

and

$$u_1 + v_1 + \delta \leq B_1(u_1 + \delta) \quad , \textit{for} \quad \delta = \delta_-, \textit{ and } \delta = \delta_+. \tag{22}$$

Then $S_r^\delta(R_2) \subseteq R$ for any $\delta \in [\delta_-, \delta_+]$.

Combining Theorems 5.4 and 5.5, we have the following stability theorem.

Theorem 5.6 *Suppose $|x_n| \leq \alpha$ for all n, and that u_n, v_n are the state variables of the second order linear $\Sigma\Delta$ scheme. If the parameters γ, α, C satisfy the hypotheses of Theorems 5.4 and 5.5, then*

$$\forall n, \quad (u_n, v_n) \in R.$$

Our main motivation for deriving the stability result, Theorem 5.6, under the weaker hypotheses of Theorems 5.4 and 5.5 was to obtain invariant sets which can be bounded inside $(-2,2)\times\mathbb{R}$. Let us now check that this is indeed possible under the weaker hypotheses. We need to show that there exist combinations of the parameters $0 < \gamma, C$ and $0 < \alpha < 1$ such that

$$\text{Condition } (A): \qquad (19), (20), (21), (22) \text{ hold, and } -2 < u_0.$$

One can check that this is possible by inspection. First note that the items in Condition (A) can all be written exclusively in terms of γ, C, and α. In fact, one has $\delta_- = 1-\alpha$, $\delta_+ = 1+\alpha$, $u_0 = -[2C(1-\alpha^2)]^{\frac{1}{2}}$, and $v_0 = B_1(u_0)$. It is also straightforward to derive that

$$u_1 = \frac{(1+\alpha)(1+\frac{2}{\gamma}) - \sqrt{(1+\alpha)^2(1+\frac{2}{\gamma})^2 + 8(1+\alpha)C}}{2}$$

and

$$v_1 = -\frac{1}{\gamma}u_1.$$

Figure 2 plots a range of the parameters $0 < \gamma$ and $0 < \alpha < 1$, with C fixed at $C = 1.99$, for which Condition (A) holds.

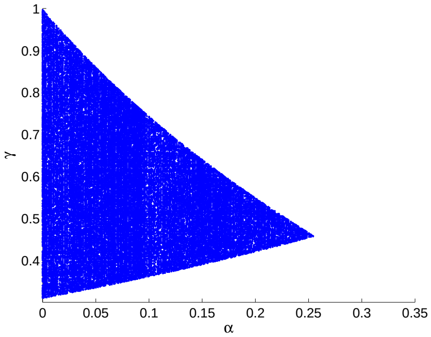

Fig. 2. With $C = 1.99$ fixed, the figure shows a range of the parameters γ and α for which Condition (A) holds. In particular, for these choices of parameters the invariant set R is bounded inside $(-2,2)\times\mathbb{R}$.

Finally, let us mention that the previous stability result extends to the general 1-bit case with $Q(w) = \frac{\delta}{2}\text{sign}(w)$.

Corollary 5.7 *Suppose $|x_n| \leq \frac{\delta}{2}\alpha$ for all n, and that u_n, v_n are the state variables of the second order linear $\Sigma\Delta$ scheme with $Q(w) = \frac{\delta}{2}\mathrm{sign}(w)$. If the parameters γ, α, C satisfy the hypotheses of Theorems 5.4 and 5.5, then*

$$\forall n, \quad (u_n, v_n) \in \frac{\delta}{2}R.$$

6 Approximation error

We are now ready to derive approximation error estimates for the second order $\Sigma\Delta$ scheme (11). More precisely, given a unit-norm tight frame $F = \{e_n\}_{=1}^{N}$ for $\mathbb{R}^d$, a permutation p of $\{1, 2, \cdots, N\}$, and $x \in \mathbb{R}^d$, we let $x_{p(n)} = \langle x, e_{p(n)} \rangle$ be the frame coefficients. The $\Sigma\Delta$ scheme (11) produces a quantized sequence $\{q_n\}_{n=1}^{N}$ where each $q_n \in \{-1, 1\}$. We shall derive estimates for the approximation error

$$||x - \widetilde{x}|| = ||\frac{d}{N} \sum_{n=1}^{N} (x_{p(n)} - q_n) e_{p(n)}||, \tag{23}$$

where $\widetilde{x}$ is as in (7).

$\Sigma\Delta$ schemes are iterative in nature, and it was observed in [12] that the approximation error for $\Sigma\Delta$ quantization of finite frame expansions depends closely on the order in which frame coefficients are quantized, i.e., it depends on the choice of p. The intuition behind this is that $\Sigma\Delta$ schemes are able to take advantage of "interdependencies" between the frame elements in a redundant frame expansion. In fact, it is advantageous to order the frame so that adjacent frame elements are closely correlated in order to obtain optimally small approximation error. To make this more precise we introduce the notion of frame variation.

6.1 Frame variation

Given a finite frame $F = \{e_n\}_{n=1}^{N}$ for $\mathbb{R}^d$ and a permutation p of $\{1, 2, \cdots, N\}$. The *jth order frame variation $\sigma_j(F, p)$* of F with respect to p is defined by

$$\sigma_j(F, p) = \sum_{n=1}^{N-j} ||\Delta^j e_{p(n)}||,$$

where Δ^j is the jth order difference operator defined by

$$\Delta^1 e_n = \Delta e_n = e_n - e_{n+1} \quad \text{and} \quad \Delta^j e_n = \Delta^{j-1} \Delta^1 e_n.$$

The first order variation, $\sigma(F,p) = \sigma_1(F,P) = \sum_{n=1}^{N-1} ||e_{p(n)} - e_{p(n+1)}||$, is simply an overall measure how well adjacent frame elements are correlated in the permutation p. The first order frame variation was used in [12] to analyze first order $\Sigma\Delta$ schemes. Since this paper deals with a specific second order scheme, our error estimates will involve computations with the second order frame variation. The following result shows that harmonic frames in their natural ordering have uniformly bounded 2nd order frame variation.

Lemma 6.1 *Let $H_N^d = \{e_n\}_{n=0}^{N-1}$ be an harmonic frame for $\mathbb{R}^d$ as defined in Section 3, and let p be the identity permutation of $\{0,1,\cdots,N-1\}$. Then*

$$\sigma_2(H_N^d,p) \leq \frac{2\pi^2 d^2}{N}.$$

PROOF. First suppose d is even. Using the definitions of the second order variation and harmonic frame, and using the mean value theorem to obtain the second inequality, we have

$$\begin{aligned}\sqrt{\frac{d}{2}}\sigma_2(H_N^d,p) &= \sqrt{\frac{d}{2}}\sum_{j=0}^{N-3} ||e_j - 2e_{j+1} + e_{j+2}|| \\ &\leq \sum_{j=0}^{N-3}\left[\sum_{k=1}^{d/2}\left(\cos\frac{2\pi kj}{N} - 2\cos\frac{2\pi k(j+1)}{N} + \cos\frac{2\pi k(j+2)}{N}\right)^2\right. \\ &\left. + \sum_{k=1}^{d/2}\left(\sin\frac{2\pi kj}{N} - 2\sin\frac{2\pi k(j+1)}{N} + \sin\frac{2\pi k(j+2)}{N}\right)^2\right]^{\frac{1}{2}} \\ &\leq \sum_{j=0}^{N-3}\left[2\sum_{k=1}^{d/2}\left(\frac{2\pi k}{N}\right)^4\right]^{\frac{1}{2}} \leq \frac{4\pi^2}{N}\sqrt{2}\left[\sum_{k=1}^{d/2} k^4\right]^{\frac{1}{2}} \leq \frac{2\pi^2 d^{5/2}}{N\sqrt{2}}.\end{aligned}$$

If d is odd we likewise have

$$\sqrt{\frac{d}{2}}\sigma_2(H_N^d,p) \leq \frac{4\pi^2}{N}\sqrt{2}\left[\sum_{k=1}^{(d-1)/2} k^4\right]^{\frac{1}{2}} \leq \frac{2\pi^2 d^{5/2}}{N\sqrt{2}}.$$

Thus,

$$\sigma_2(H_N^d,p) \leq \frac{2\pi^2 d^2}{N}.$$

□

6.2 Error estimates

Using the second order frame variation, we are now ready to derive error estimates for the scheme (11).

Theorem 6.2 *Let $F = \{e_n\}_{n=1}^N$ be a finite unit-norm tight frame for $\mathbb{R}^d$ and let p be a permutation of $\{1, 2, \cdots, N\}$. Suppose that $x \in \mathbb{R}^d$. Let $\{q_n\}_{n=1}^N$ be the quantized bits produced by (10) for any function Q, and let the input be given by the frame coefficients $\{x_{p(n)}\}_{n=1}^N$ of x. Then*

$$||x - \widetilde{x}|| \leq \frac{d}{N}\left(||v||_\infty \sigma_2(F, p) + |v_{N-1}|\ ||\Delta e_{p(N-1)}|| + |u_N|\right),$$

where $\widetilde{x}$ is as in (7) and $||\cdot||_\infty$ denotes the l^∞ norm of a sequence.

PROOF. Let $f_n = e_{p(n)} - e_{p(n+1)}$, and also recall that $u_0 = v_0 = 0$ and $\Delta v_n = u_n$ by (11). Then

$$\begin{aligned} x - \widetilde{x} &= \frac{d}{N}\sum_{n=1}^{N}(x_{p(n)} - q_n)e_{p(n)} \\ &= \frac{d}{N}\left(\sum_{n=1}^{N} u_n e_{p(n)} - \sum_{n=1}^{N} u_{n-1}e_{p(n)}\right) \\ &= \frac{d}{N}\left(\sum_{n=1}^{N-1} u_n(e_{p(n)} - e_{p(n+1)}) - u_0 e_{p(1)} + u_N e_{p(N)}\right) \\ &= \frac{d}{N}\left(\sum_{n=1}^{N-1} \Delta v_n f_n + u_N e_{p(N)}\right) \\ &= \frac{d}{N}\left(\sum_{n=1}^{N-2} v_n(f_n - f_{n+1}) + v_{N-1}f_{N-1} - v_0 f_1 + u_N e_{p(N)}\right) \\ &= \frac{d}{N}\left(\sum_{n=1}^{N-2} v_n(f_n - f_{n+1}) + v_{N-1}f_{N-1} + u_N e_{p(N)}\right). \end{aligned} \tag{24}$$

Thus,

$$||x - \widetilde{x}|| \leq \frac{d}{N}\left(||v||_\infty \sigma_2(F, p) + |v_{N-1}|\ ||\Delta e_{p(N-1)}|| + |u_N|\right). \tag{25}$$

□

For our subsequent approximation error estimates, it will be especially important to determine the value of $|u_N|$.

Lemma 6.3 *Let $F = \{e_n\}_{n=1}^N$ be a finite unit-norm tight frame for $\mathbb{R}^d$ and suppose that $S = \sum_{j=1}^N e_n$. If $x \in \mathbb{R}^d$ is the signal being quantized, then*

$$u_N \in \begin{cases} \langle x, S\rangle + \delta\mathbb{Z}, & \textit{if } N \textit{ is even}, \\ \langle x, S\rangle + \delta(\mathbb{Z} + \frac{1}{2}), & \textit{if } N \textit{ is odd}. \end{cases} \tag{26}$$

PROOF. First note that by definition of the $\Sigma\Delta$ scheme and our hypotheses

$$u_N = u_0 + \sum_{j=1}^{N} \langle x, e_n \rangle - \sum_{j=1}^{N} q_N = \langle x, S \rangle - \sum_{j=1}^{N} q_N.$$

Since $q_n \in \{-\frac{\delta}{2}, \frac{\delta}{2}\}$,

$$\sum_{j=1}^{N} q_N \in \begin{cases} \delta\mathbb{Z}, & \text{if } N \text{ is even,} \\ \delta(\mathbb{Z} + \frac{1}{2}), & \text{if } N \text{ is odd,} \end{cases}$$

and the result follows. □

Corollary 6.4 (Harmonic frames in even dimensions) *Let $F = H_N^d = \{e_n\}_{n=1}^N$ be an harmonic frame for $\mathbb{R}^d$, where d is even, and let p be the identity permutation of $\{1, 2, \cdots, N\}$. Suppose that $x \in \mathbb{R}^d$, $||x|| \le \frac{\delta}{2}\alpha$, $\alpha < 1$, and that the parameters α, C, γ satisfy Condition (A). Let $\{q_n\}_{n=1}^N$ be the quantized bits produced by (11) with $Q(w) = \frac{\delta}{2}\mathrm{sign}(w)$, and suppose the input is given by the frame coefficients $\{x_{p(n)}\}_{n=1}^N$ of x. Then if N is even, we have*

$$\|x - \widetilde{x}\| \le \frac{d\delta}{N^2}\left(C\pi^2 d^2 + C\pi d\right),$$

and if N is odd, we have

$$\frac{d\delta}{N}\left(\frac{1}{2} - \frac{C\pi^2 d^2 + C\pi d}{N}\right) \le \|x - \widetilde{x}\| \le \frac{d\delta}{N}\left(\frac{C\pi^2 d^2 + C\pi d}{N} + \frac{1}{2}\right).$$

In particular, if N is odd then $\frac{\delta}{N} \lesssim ||x - \widetilde{x}|| \lesssim \frac{\delta}{N}$.

PROOF. First note that by Corollary 5.7, the state variables u_n, v_n stay bounded in the set $\frac{\delta}{2}R$ defined by (17). Moreover, Condition (A) ensures that $R \subseteq \frac{\delta}{2}\left((-2, 2) \times [-C, C]\right)$. Thus, for all n, the state variables satisfy

$$|u_n| < \delta \text{ and } |v_n| \le C\frac{\delta}{2}. \tag{27}$$

Further since d is even, $S = \sum_{n=1}^N e_n = 0$. Also note that for harmonic frames H_N^d, $||\Delta e_n|| \le \frac{2\pi d}{N}$. This follows by calculating as in Lemma 6.1, or as in [12].

If N is even, then by Lemma 6.3, $u_N \in \delta\mathbb{Z}$, and thus (27) implies that $u_N = 0$. By Lemma 6.1 and Theorem 6.2,

$$||x - \widetilde{x}|| \le \frac{d\delta}{N^2}\left(C\pi^2 d^2 + C\pi d\right).$$

If N is odd, then proceeding as above we have that $|u_N| = \delta/2$. In this case, Lemma 6.1 and Theorem 6.2 imply

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

Likewise, working directly with (24) gives

$$\frac{d\delta}{2N} \leq ||x - \widetilde{x}|| + \frac{d\delta}{N^2}\left(C\pi^2 d^2 + C\pi d\right).$$

□

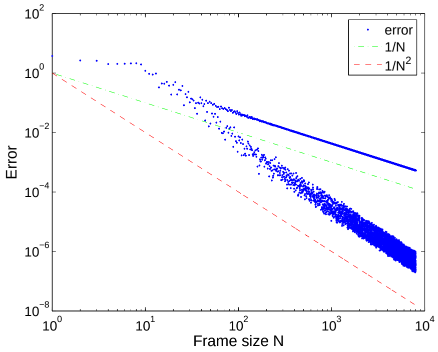

Fig. 3. The frame expansions of $(0.37/\pi, 0.0017, e^{-7}, 0.001) \in \mathbb{R}^4$ with respect to the harmonic frames H_N^4 are quantized using the second order $\Sigma\Delta$ scheme (11) with $\gamma = 1/2$. The figure shows a log-log plot of the approximation error $||x - \widetilde{x}||$ against the frame size N. The figure also shows the graphs of $1/N$ and $1/N^2$ for comparison.

Figure 3 shows a log-log plot of the approximation error when the second order $\Sigma\Delta$ scheme is used to quantize harmonic frame expansions in $\mathbb{R}^4$ of $x = (\frac{.37}{\pi}, 0.0017, e^{-7}, 0.001) \in \mathbb{R}^4$. The figure plots the approximation error, $||x - \widetilde{x}||$, against the cardinality, N, of the harmonic frame H_N^d. The quantization ordering is taken to be the natural ordering of the harmonic frame. The figure also plots the functions $1/N$ and $1/N^2$ for comparison. Observe that the approximation error behaves quite differently for N even and N odd. This observation agrees with the theoretical error estimates given by Corollary 6.4.

One has a similar result for odd dimensions. We omit the proof since it is similar to the proof of Corollary 6.4.

Corollary 6.5 (Harmonic frames in odd dimensions) *Let $F = H_N^d = \{e_n\}_{n=1}^N$ be an harmonic frame for $\mathbb{R}^d$, where d is odd, and let p be a permutation of $\{1, 2, \cdots, N\}$. Suppose that $x \in \mathbb{R}^d$, $||x|| \leq \frac{\delta}{2}\alpha$, $\alpha < 1$, and that the parameters α, C, γ satisfy Condition (A). Let $\{q_n\}_{n=1}^N$ be the quantized bits produced by (11) with $Q(w) = \frac{\delta}{2}\mathrm{sign}(w)$, and suppose the input is given by the* *frame coefficients* $\{x_{p(n)}\}_{n=1}^{N}$ *of* x. *Let* $S = \frac{N}{\sqrt{d}}(1, 0, \cdots, 0)$. *Then*

$$\frac{d\delta}{N}\left(C_{S,N} - \frac{C\pi^2 d^2 + C\pi d}{N}\right) \leq ||x - \widetilde{x}|| \leq \frac{d\delta}{N}\left(\frac{C\pi^2 d^2 + C\pi d}{N} + C_{S,N}\right),$$

where $C_{S,N}$ *is the unique element of* S_N *contained in* $(-\delta, \delta)$, *and*

$$S_N = \begin{cases} \langle x, S\rangle + \delta\mathbb{Z}, & \text{if } N \text{ is even,} \\ \langle x, S\rangle + \delta(\mathbb{Z} + \frac{1}{2}), & \text{if } N \text{ is odd.} \end{cases}$$

Note that if $C_{S,N} = 0$ *then* $||x - \widetilde{x}|| \lesssim \frac{\delta}{N^2}$; *otherwise* $\frac{\delta}{N} \lesssim ||x - \widetilde{x}|| \lesssim \frac{\delta}{N}$. *A simple case where* $C_{S,N} = 0$ *occurs when* x *is in the* $d-1$ *dimensional subspace determined by* $\langle x, S\rangle = 0$.

7 A hybrid multibit scheme

So far we have only considered the 1-bit, 2nd order scheme (11), i.e., with $Q(w) = \frac{\delta}{2}\text{sign}(w)$. Although the error estimate (9) for first order $\Sigma\Delta$ quantizers was derived for multibit quantizer functions, it is not so simple to extend second order results to general Q. The main reason for this is that the invariant set results of Section 5 do not immediately extend to the multibit case, although it is likely that analogous, but “messier”, stability results do exist. Deriving invariant sets for general K level quantizers with step size δ represents ongoing work.

Similar to (9), for multibit second order schemes we would like to have the error estimate

$$||x - \widetilde{x}|| \lesssim \frac{d\delta}{N^2} \tag{28}$$

hold for a range of x which is independent of δ. By Corollary 6.4, the 1-bit scheme can give the estimate (28); however, there the estimate only holds if $||x|| \leq \frac{\delta}{2}\alpha$. In particular, it does not hold for a range of x which is independent of δ. The most direct way to avoid this is to use a multibit scheme, i.e., a multilevel quantizer function, but as mentioned above, an analysis of the multibit second order $\Sigma\Delta$ scheme requires a deeper investigation of stability results, and takes us beyond the intended scope of this paper. In view of this, our solution is to introduce a hybrid PCM/$\Sigma\Delta$ scheme.

The hybrid PCM/$\Sigma\Delta$ scheme consists of the following three steps:

(1) Let $x \in \mathbb{R}^d$ satisfy $||x|| < K\delta$. First, quantize the frame expansion of x using $q_n^Q = Q(x_n)$, where $Q(\cdot)$ is the K level midrise quantizer with

stepsize δ, and call the resulting signal

$$x_Q = \frac{d}{N} \sum_{n=1}^{N} q_n^Q \ e_n.$$

This is simply PCM quantization. In particular, we have $x = x^Q + x^R$, where $x_R = \frac{d}{N} \sum_{n=1}^{N} (x_n - q_n^Q) \ e_n$, and $x_n^R = x_n - q_n^Q$ satisfy $|x_n^R| \leq \delta/2$.

(2) Next apply the second order $\Sigma\Delta$ scheme (11) to x_n^R, and obtain

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

(3) We define the quantized output of the hybrid scheme to be

$$\widetilde{x_H} = \frac{d}{N} \sum_{n=1}^{N} (q_n^Q + q_n^R) e_n.$$

Theorem 7.1 *Let $F = \{e_n\}_{n=1}^N$ be a unit-norm tight frame for $\mathbb{R}^d$ such that $\sum_{n=1}^N e_n = 0$, let p be a permutation of $\{1, 2, \cdots, N\}$, and let $x \in \mathbb{R}^d$ satisfy $||x|| \leq K\delta$. If the parameters γ, α, C satisfy Condition (A), and if $\sum_{n=1}^N q_n^Q = 0$ then*

$$||x - \widetilde{x_H}|| \leq \frac{dC\delta}{2N} (\sigma_2(F, p) + \|e_{p(N-1)} - e_{p(N)}\|).$$

PROOF. As in the proof of Corollary 6.4, Condition (A) implies that the state variables satisfy $|u_n| < \delta$ and $|v_n| < C\frac{\delta}{2}$, and that $u_N = 0$. The result now follows from Theorem 6.2. □

The condition $\sum_{n=1}^N q_n^Q = 0$ holds in many settings, depending on the frame F, and the element x being quantized. For example, one has the following corollary.

Corollary 7.2 *Let $E_N = \{e_n\}_{n=1}^N$ be the unit-norm tight frame for $\mathbb{R}^2$ given by the Nth roots of unity, and suppose the parameters γ, α, C satisfy Condition (A). If N is even then for almost every $x \in \mathbb{R}^2$ satisfying $||x|| \leq \alpha$,*

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

PROOF. This follows from Theorem 7.1 since the symmetry of the frame and alphabet implies that whenever the frame coefficients $\langle x, e_n^N \rangle$ are all nonzero (note that this happens for a.e. $x \in \mathbb{R}^2$), one has $\sum_{n=1}^N q_n^Q = 0$. We also used that $\sigma_2(E_N, p) \leq (2\pi)^2/N$ and $||\Delta e_N|| \leq 2\pi/N$. □

Acknowledgments

The authors thank Ingrid Daubechies, Sinan Güntürk, and Nguyen Thao for valuable discussions on the material.

References

[1] I. Daubechies, R. DeVore, Approximating a bandlimited function using very coarsely quantized data: A family of stable sigma-delta modulators of arbitrary order, Ann. of Math. 158 (2) (2003) 679–710.

[2] J. J. Benedetto, P. S. G. Ferreira (Eds.), Modern Sampling Theory. Mathematics and Applications., Birkhäuser, Boston, MA, 2001.

[3] R. Gray, Quantization noise spectra, IEEE Transactions on Information Theory 36 (6) (1990) 1220–1244.

[4] R. L. Adler, B. P. Kitchens, M. Martens, C. P. Tresser, C. W. Wu, The mathematics of halftoning, IBM J. Res. and Dev. 47 (1) (2003) 5–15.

[5] R. Ulichney, Digital Halftoning, MIT Press, Cambridge, MA, 1987.

[6] V. Goyal, J. Kovačević, J. Kelner, Quantized frame expansions with erasures, Appl. Comput. Harmon. Anal. 10 (2001) 203–233.

[7] V. Goyal, J. Kovačević, M. Vetterli, Quantized frame expansions as source-channel codes for erasure channels, in: Proc. IEEE Data Compression Conference, 1999, pp. 326–335.

[8] G. Rath, C. Guillemot, Syndrome decoding and performance analysis of dft codes with bursty erasures, in: Proc. Data Compression Conference (DCC), 2002, pp. 282–291.

[9] P. Casazza, J. Kovačević, Equal-norm tight frames with erasures, Advances in Computational Mathematics 18 (2/4) (2003) 387–430.

[10] G. Rath, C. Guillemot, Recent advances in DFT codes based quantized frame expansions for erasure channels, Digital Signal Processing 14 (4) (2004) 332–354.

[11] J. J. Benedetto, Ö. Yılmaz, A. M. Powell, Sigma-delta quantization and finite frames, in: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Vol. 3, Montreal, Canada, 2004, pp. 937–940.

[12] J. J. Benedetto, A. M. Powell, Ö. Yılmaz, Sigma-Delta ($\Sigma\Delta$) quantization and finite frames, IEEE Transactions on Information Theory, Submitted, August 2004.

[13] Ö. Yılmaz, Stability analysis for several sigma-delta methods of coarse quantization of bandlimited functions, Constructive Approximation 18 (2002) 599–623.

[14] V. Goyal, J. Kovačević, 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.

[15] T. Strohmer, R. Heath Jr., Grassmannian frames with applications to coding and communications, Appl. Comput. Harmon. Anal. 14 (3) (2003) 257–275.

[16] B. Hochwald, T. Marzetta, T. Richardson, W. Sweldens, R. Urbanke, Systematic design of unitary space-time constellations, IEEE Trans. Inform. Theory 46 (6) (2000) 1962–1973.

[17] S. Waldron, Generalized Welch bound equality sequences are tight frames, IEEE Transactions on Information Theory 49 (9) (2003) 2307–2309.

[18] Y. Eldar, G. Forney, Optimal tight frames and quantum measurement, IEEE Transactions on Information Theory 48 (3) (2002) 599–610.

[19] G. Zimmermann, Normalized tight frames in finite dimensions, in: K. Jetter, W. Haussmann, M. Reimer (Eds.), Recent Progress in Multivariate Approximation, Birkhäuser, 2001.

[20] V. Goyal, M. Vetterli, N. Thao, Quantized overcomplete expansions in $\mathbb{R}^n$: Analysis, synthesis, and algorithms, IEEE Transactions on Information Theory 44 (1) (1998) 16–31.

[21] J. J. Benedetto, M. Fickus, Finite normalized tight frames, Advances in Computational Mathematics 18 (2/4) (2003) 357–385.

[22] N. Thao, M. Vetterli, Deterministic analysis of oversampled A/D conversion and decoding improvement based on consistent estimates, IEEE Transactions on Information Theory 42 (3) (1994) 519–531.

[23] Z. Cvetković, Resilience properties of redundant expansions under additive noise quantization, IEEE Transactions on Information Theory 49 (3) (2003) 644–656.

[24] W. Bennett, Spectra of quantized signals, Bell Syst. Tech. J. 27 (1948) 446–472.

[25] B. Beferull-Lozano, A. Ortega, Efficient quantization for overcomplete expansions in $\mathbb{R}^d$, IEEE Transactions on Information Theory 49 (1) (2003) 129–150.

[26] H. Bölcskei, Noise reduction in oversampled filter banks using predictive quantization, IEEE Transactions on Information Theory 47 (1) (2001) 155–172.

[27] P. Aziz, H. Sorensen, J. V. D. Spiegel, An overview of sigma-delta converters, IEEE Signal Processing Magazine 13 (1) (1996) 61–84.

[28] J. Candy, G. Temes (Eds.), Oversampling Delta-Sigma Data Converters, IEEE Press, 1992.

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

[30] C. S. Güntürk, Approximating a bandlimited function using very coarsely quantized data: improved error estimates in sigma-delta modulation, J. Amer. Math. Soc. 17 (1) (2004) 229–242.

[31] C. S. Güntürk, T. Nguyen, Ergodic dynamics in $\Sigma\Delta$ quantization: tiling invariant sets and spectral analysis of error, Advances in Applied Mathematics, to appear.

[32] C. S. Güntürk, T. Nguyen, Refined error analysis in second-order $\Sigma\Delta$ modulation with constant inputs, IEEE Transactions on Information Theory 50 (5) (2004) 839–860.

[33] W. Chen, B. Han, Improving the accuracy estimate for the first order sigma-delta modulator, J. Amer. Math. Soc., Submitted in 2003.

[34] N. Thao, MSE behavior and centroid function of mth order aymptotic $\Sigma\Delta$modulators, IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing 49 (2) (2002) 86–100.