

SMOOTH FRAME-PATH TERMINATION FOR HIGHER ORDER SIGMA-DELTA QUANTIZATION

BERNHARD G. BODMANN, VERN I. PAULSEN, AND SOHA A. ABDULBAKI

ABSTRACT. We study the performance of finite frames for the encoding of vectors by applying standard higher-order sigma-delta quantization to the frame coefficients. Our results are valid for any quantizer with accuracy $\epsilon > 0$ operating in the no-overload regime. The frames under consideration are obtained from regular sampling of a path in a Hilbert space. In order to achieve error bounds that are comparable to results on higher-order sigma-delta for the quantization of oversampled bandlimited functions, we construct frame paths that terminate smoothly in the zero vector, that is, with an appropriate number of vanishing derivatives at the endpoint.

1. INTRODUCTION

For many common tasks of digital signal processing, signals are represented by a finite number of linear coefficients, and there is a norm topology on the space of signals that comes from an inner product. In short, a signal is a vector in a finite dimensional real or complex Hilbert space. In this work, we investigate how to approximate such vectors accurately with a round-off scheme that relies on the redundancy of frame expansions to suppress the quantization error. This means, the frame coefficients may only assume finitely many values, and the reconstruction is linear in these coefficients. Several quantization strategies have been investigated [GVT98, GKK01, BLO03, BYP04, BPY06], in analogy with results for oversampled signals that are bandlimited [NST97, Yil02, Gün04] or contained in shift-invariant spaces [BH01]. Following a strategy similar to a previous paper [BP], we study error bounds for standard second and higher-order sigma-delta modulation applied to the frame coefficients.

The second order portion of this work can be seen as complementary to the paper by Benedetto, Powell and Yilmaz [BPY06], where a whole family of quantization maps is used in conjunction with single-bit quantizers. We allow the use of arbitrary (scalar) quantizers of a given accuracy $\epsilon > 0$ in the no-overload regime, but for reasons of simplicity, we restrict ourselves to a standard-version of the sigma-delta algorithm based on tracking the iterated cumulative quantization errors.

The results presented here explain how to obtain error bounds comparable to those of higher order sigma-delta encoding for oversampled bandlimited functions. In order to stay in the no-overload regime, raising the order requires us to increase the number of quantization levels exponentially in K . Adapting the careful treatment of one-bit sigma-delta quantization in [DD03], one may expect similarly robust results for quantizing frame coefficients without the need to invoke more than two levels.

2000 *Mathematics Subject Classification.* Primary 46L05; Secondary 46A22, 46H25, 46M10, 47A20.
Key words and phrases. frames, sigma-delta quantization.

This research was partially supported by the National Science Foundation grant DMS-0300128.

The families of frames considered here are obtained from regular sampling of a path $f : [a, b] \rightarrow \mathcal{H}$ in a finite-dimensional Hilbert space \mathcal{H} . A frame path allows the redundancy to be increased to any desired level, that is, there is a constant $C > 0$ such that regular sampling of f produces NC -tight frames $F_N = \{f_1, f_2, \dots, f_N\}$ for most $N \in \mathbb{N}$. Our main result is an upper bound for the Euclidean norm of the error induced by the higher-order sigma-delta encoding of frame coefficients. This bound requires that the frame path has a sufficient number of vanishing derivatives at the endpoint in order to suppress boundary terms. We call a frame path *Kth-order zero terminated* if it has $K - 1$ absolutely continuous derivatives and satisfies $f(b) = f'(b) = \dots = f^{(K-1)}(b) = 0$. Given a scalar quantizer Q of accuracy $\epsilon > 0$ and an NC -tight frame F_N obtained from regular sampling of such a path, then for $x \in \text{span } F_N$, the error of the standard K th-order sigma-delta encoding of frame coefficients is bounded by

$$\|x - Q_{F_N}^{(K)}(x)\| \leq \frac{\epsilon(b-a)^{K-1}}{CN^K} \max_{\|v\|=1} \int_a^b |\langle f^{(K)}(t), v \rangle| dt$$

where it is assumed that x does not lead to overloading of the quantizer. In analogy with the results from our a previous paper [BP], we can further estimate the so-called *total directional variation* of $f^{(K)}$, $\max_{\|v\|=1} \int_a^b |\langle f^{(K)}(t), v \rangle| dt$, and compare our bounds for the performance of different frame paths.

This paper is organized as follows: In Section 2, we review the background of the second-order sigma-delta algorithm, and show how this algorithm is used for the encoding of vectors by applying it to the sequence of frame coefficients. Then we derive error bounds of this encoding scheme for certain tight frames. The key element for the construction of these frames is the suppression of boundary terms with the help of projections. Section 3 introduces the analogue of this strategy for sigma-delta algorithms of any order. The central topics of Section 4 are the construction of smoothly terminated frame paths and bounds on the quantization error for frames obtained from regular sampling.

2. SECOND-ORDER SIGMA-DELTA ENCODING

We begin by recalling the basics of the sigma-delta algorithm. We follow the notation of our first-order results.

Definition 2.1. *A function Q on \mathbb{R} is called a quantizer with accuracy $\epsilon > 0$ on the interval $[-L, +L]$ if it has a finite range $\mathbb{A} \subset [-L, +L]$ and for any $x \in [-L, +L]$, $Q(x)$ satisfies $|x - Q(x)| \leq \epsilon$. The range \mathbb{A} of the quantizer Q is also called the alphabet.*

Remark 2.2. This alphabet could consist of all integer multiples of a fixed step-size δ contained in the interval $[-L, +L]$. The quantizer that assigns to $x \in [-L, +L]$ the unique value $m\delta$, $m \in \mathbb{Z}$, satisfying $(m - \frac{1}{2})\delta < x \leq (m + \frac{1}{2})\delta$ is commonly referred to as the *uniform mid-tread quantizer with step-size δ* [OS89, PM96]. Alternatively, one may choose the alphabet to be $\mathbb{A} = (\mathbb{Z} + \frac{1}{2})\delta \cap [-L, +L]$ and assign to $x \in [-L, +L]$ the value $(m + \frac{1}{2})\delta$ such that $m\delta < x \leq (m + 1)\delta$, which would result in the so-called *uniform mid-riser quantizer with step-size δ* . Note that all of these are quantizers with accuracy $\delta/2$. All of the results presented here are valid for quantizers with accuracy $\epsilon > 0$, and do not require them to be uniform. This may be beneficial for robustness against hardware imperfections such as inaccurate quantizer calibration or drifting of the individual quantization levels during the encoding process [DD03]. In fact, our results

do not require that the quantized value $Q(x)$ is equal to a fixed constant each time x is quantized, all that is needed is that this value remains within ϵ of x .

2.1. Encoding of sequences. In this work, we exclusively consider the behavior of the so-called standard higher-order sigma-delta quantizer. Elsewhere, a whole family of such algorithms has been investigated [Yil02]. The standard second-order sigma-delta algorithm for bounded sequences is defined as follows.

Definition 2.3. *Let Q be a quantizer with accuracy $\epsilon > 0$ on the interval $[-L, +L]$. Given an input sequence $\{x_j\}_{j=1}^{\infty}$, then the **standard second-order sigma-delta quantized sequence** $\{q_j\}_{j=1}^{\infty}$ associated with the initial values $u_0^{(1)} \in \mathbb{R}$ and $u_0^{(2)} \in \mathbb{R}$ is obtained by inductively defining q_j , $u_j^{(1)}$ and $u_j^{(2)}$ for $j \in \mathbb{N}$ with the quantization prescription*

$$q_j := Q(x_j + u_{j-1}^{(1)} + u_{j-1}^{(2)}),$$

the cumulative quantization error

$$u_j^{(1)} := x_j - q_j + u_{j-1}^{(1)},$$

and the iterated sum

$$u_j^{(2)} := u_{j-1}^{(2)} + u_j^{(1)} = u_{j-1}^{(2)} + u_{j-1}^{(1)} + x_j - q_j.$$

The above iteration corresponds to the standard second order double-loop sigma-delta circuit [HKB92]. In more general algorithms, the argument of the quantizer Q can consist of other linear combinations of x , $u_{j-1}^{(1)}$ and $u_{j-1}^{(2)}$ or it can even be nonlinear in these quantities [Tha02, DD03, GT04]. Any such type of second-order algorithm is called stable when bounded input sequences and appropriate initial values for $u_0^{(1)}$, $u_0^{(2)}$ guarantee that the sequence $u^{(2)}$ is bounded.

Proposition 2.4. *If Q is a quantizer with accuracy $\epsilon > 0$ on $[-L, L]$, the initial values are chosen such that $|u_0^{(1)}| \leq 2\epsilon$ and $|u_0^{(2)}| \leq \epsilon$, and the input sequence $\{x_j\}$ is bounded by $\|x\|_{\infty} < L - 3\epsilon$, then $\|u^{(2)}\|_{\infty} \leq \epsilon$ and $\|u^{(1)}\|_{\infty} \leq 2\epsilon$. Moreover, the pairs $(u_j^{(1)}, u_j^{(2)})$ satisfy $u_j^{(2)} \in [-\epsilon, \epsilon]$ and $|u_j^{(1)} - u_j^{(2)}| \leq \epsilon$.*

Proof. We proceed by induction. We note that if $|u_{j-1}^{(1)}| \leq 2\epsilon$ and $|u_{j-1}^{(2)}| \leq \epsilon$, then $x_j + u_{j-1}^{(1)} + u_{j-1}^{(2)}$ is in the domain of accuracy for the quantizer, and

$$|u_j^{(2)}| = |u_{j-1}^{(2)} + u_{j-1}^{(1)} + x_j - q_j| \leq \epsilon.$$

Now the bound for $u_j^{(1)}$ follows from $|u_j^{(1)}| = |u_j^{(2)} - u_{j-1}^{(2)}| \leq 2\epsilon$. Finally, the set containing all pairs $(u_j^{(1)}, u_j^{(2)})$ comes from the bound $|u_j^{(2)}| \leq \epsilon$ and $|u_j^{(1)} - u_j^{(2)}| = |u_{j-1}^{(2)}| \leq \epsilon$. \square

Unfortunately, for general bounded sequences, the error of the second-order sigma-delta quantizer is worse than that of the first-order algorithm.

Corollary 2.5. *For any given $j \in \mathbb{N}$, we have a bound for the difference of the moving M -term averages,*

$$\left| \frac{1}{M} \sum_{k=0}^{M-1} x_{j+k} - \frac{1}{M} \sum_{k=0}^{M-1} q_{j+k} \right| = \frac{1}{M} |u_{j+M-1}^{(1)} - u_{j-1}^{(1)}| \leq \frac{4\epsilon}{M}.$$

Choosing $u_0^{(1)} = 0$ gives an improved bound for the difference of the initial M -term averages,

$$\left| \frac{1}{M} \sum_{k=1}^M x_k - \frac{1}{M} \sum_{k=1}^M q_k \right| = |u_M^{(1)}|/M \leq \frac{2\epsilon}{M}.$$

Remark 2.6. If we take a constant input, $x_j = x \in \mathbb{R}$ for all $j \in \mathbb{N}$, and $u_0^{(1)} = u_0^{(2)} = 0$, then we see that given any quantizer with accuracy $\epsilon > 0$, the second-order sigma-delta algorithm gives a sequence of quantized values $\{q_k\}$ satisfying $|x - \frac{1}{M} \sum_{k=1}^M q_k| \leq \frac{2\epsilon}{M}$ for any choice of $M \in \mathbb{N}$.

At the cost of reducing the domain of the input values, we can improve the convergence for constant input by using a *modulated version* of the sigma-delta algorithm.

Definition 2.7. Given an input sequence $\{x_j\}_{j=1}^N$ and coefficients $\{c_j\}_{j=1}^N \subset \mathbb{R}$, we define the **$\{c_j\}$ -modulated second-order sigma-delta quantization** associated with the initial values $u_0^{(1)}, u_0^{(2)} \in \mathbb{R}$ to be the quantized coefficients obtained from the standard second-order sigma-delta algorithm applied to the input sequence $\{x_j c_j\}_{j \in \mathbb{N}}$.

Proposition 2.8. Let $c_j = (6/(N-1)(2N-1))^{1/2}(N-j)$, then $\sum_{j=1}^N |c_j|^2 = N$. If we take a constant input value $x \in [-L/\sqrt{3} + \sqrt{3}\epsilon, L/\sqrt{3} - \sqrt{3}\epsilon]$, and consider reconstructing x from the output of the $\{c_j\}$ -modulated second-order sigma-delta quantization by averaging $\frac{1}{N} \sum_j c_j q_j$, then the error is

$$\left| x - \frac{1}{N} \sum_{j=1}^N c_j q_j \right| \leq \frac{2\sqrt{6}\epsilon}{N\sqrt{(N-1)(2N-1)}}.$$

Proof. One may check that the modulated sequence satisfies $\|\{c_j x_j\}\|_\infty \leq L - 3\epsilon$ and by the preceding analysis, the $\{c_j\}$ -modulated quantizer is stable.

Since $c_N = 0$, and the differences $c_j - c_{j-1}$ are constant, we simplify the reconstruction error

$$\begin{aligned} \frac{1}{N} \sum_{j=1}^N (c_j x - q_j) c_j &= \frac{1}{N} \sum_{j=1}^{N-1} (u_j^{(1)} - u_{j-1}^{(1)}) c_j \\ &= \frac{1}{N} \sum_{j=1}^{N-2} u_j^{(1)} (c_j - c_{j+1}) + \frac{u_{N-1}^{(1)} c_{N-1}}{N} = \frac{1}{N} \sum_{j=1}^{N-2} (u_j^{(2)} - u_{j-1}^{(2)}) (c_j - c_{j+1}) + \frac{u_{N-1}^{(1)} c_{N-1}}{N} \\ &= \frac{1}{N} \sum_{j=1}^{N-2} u_j^{(2)} (c_j - 2c_{j+1} + c_{j+2}) + \frac{u_{N-1}^{(1)} c_{N-1}}{N} = \frac{u_{N-1}^{(1)} c_{N-1}}{N} \end{aligned}$$

and by the explicit value of c_{N-1} as well as $|u_{N-1}^{(1)}| \leq 2\epsilon$

$$\left| \frac{1}{N} \sum_{j=1}^{N-2} (c_j x - q_j) c_j \right| \leq \frac{2\sqrt{6}\epsilon}{N\sqrt{(N-1)(2N-1)}}.$$

□

2.2. Encoding of Vectors. Given a vector $x = (x_1, \dots, x_d) \in \mathbb{R}^d$ with $x_i \in [-L/\sqrt{3} + \sqrt{3}\epsilon, +L/\sqrt{3} - \sqrt{3}\epsilon]$ for all $i \in \{1, 2, \dots, d\}$, then one could apply the modulated second-order sigma-delta quantization to each coordinate. Let $N = Md$, $M \in \mathbb{N}$. Let $c_j = (6/(M-1)(2M-1))^{1/2}(M-j)$, then $\sum_{j=1}^M |c_j|^2 = M$, for $j \in \{1, 2, \dots, M\}$. For each coordinate x_i , we obtain the first M quantized values $\{q_{i,k}\}_{k=1}^M$. Averaging as above would yield $y_i = \frac{1}{M} \sum_{k=1}^M c_k q_{i,k}$ and $|x_i - y_i| \leq \frac{2\sqrt{6}\epsilon}{M\sqrt{(M-1)(2M-1)}}$. This way, the vector x is encoded with $N = Md$ quantities and the Euclidean distance between $y = (y_1, \dots, y_d)$ and x satisfies

$$\|x - y\| \leq \frac{2\sqrt{6}\epsilon\sqrt{d}}{M\sqrt{(M-1)(2M-1)}} \leq \frac{2\sqrt{3}\epsilon d^{5/2}}{N(N - \frac{1}{d})}.$$

In the following, we investigate the performance of the second-order sigma-delta algorithm applied to the frame coefficients of a vector. Our goal is to find a bound for the quantization error with the same asymptotic behavior as the modulated coordinate-wise quantization. To this end, we adapt the modulation strategy in Section 2.4 by constructing frames that terminate in the zero-vector.

Definition 2.9. Let \mathcal{H} be a real or complex Hilbert space with dimension $\dim \mathcal{H} = d$. A set of vectors $F = \{f_1, \dots, f_N\} \subset \mathcal{H}$ is called an **A-tight frame**, provided that for every $x \in \mathcal{H}$, we have the norm equality

$$A\|x\|^2 = \sum_{j=1}^N |\langle x, f_j \rangle|^2.$$

If, in addition, we have $\|f_j\|^2 = Ad/N$ for all $j \in \{1, 2, \dots, N\}$, we say that $\{f_j\}_{j=1}^N$ is a **uniform A-tight frame**. Finally, if we want to distinguish the order of the vectors, we speak of an **ordered frame**.

Remark 2.10. When F is a uniform N/d -tight frame for a d -dimensional Hilbert space, we have $\|f_j\| = 1$ for all $j \in \{1, 2, \dots, N\}$ according to the above definition. These are the frames that we are primarily interested in, but we shall see later that the additional flexibility is important for our constructions.

The norm equality in Definition 2.9 is equivalent to the requirement that every $x \in \mathcal{H}$ is reconstructed perfectly from its **frame coefficients** $\{\langle x, f_j \rangle\}_{j=1}^N$ according to

$$x = \frac{1}{A} \sum_{j=1}^N \langle x, f_j \rangle f_j.$$

Definition 2.11. Let Q be a quantizer with accuracy $\epsilon > 0$ on $[-L, +L]$, let $F = \{f_j\}_{j=1}^N$ be an A -tight (ordered) frame for a real Hilbert space \mathcal{H} and let $x \in \mathcal{H}$. Then the **second-order sigma-delta quantized vector** $Q_F^{(2)}(x)$ is defined by

$$Q_F^{(2)}(x) := \frac{1}{A} \sum_{j=1}^N q_j f_j,$$

where the quantized frame coefficients have been obtained from the standard second-order sigma-delta algorithm applied to the input sequence $\{\langle x, f_j \rangle\}_{j=1}^N$.

We henceforth call the map $Q_F^{(2)}$ a **second-order sigma-delta quantizer on \mathcal{H}** . When referring to $Q_F^{(2)}$, it is always implicit that we have chosen an associated (scalar) quantizer Q with a certain accuracy $\epsilon > 0$ on an interval $[-L, +L]$, the initial values $u_0^{(1)} = 0$ and $u_0^{(2)} = 0$, and an ordered frame F for \mathcal{H} .

2.3. Estimates for the maximal quantization error.

Proposition 2.12. *Let $F = \{f_1, f_2, \dots, f_N\}$ be an A -tight frame for a real Hilbert space \mathcal{H} , then for any $x \in \mathcal{H}$, the reconstruction error of the standard second-order sigma-delta quantization is*

$$x - Q_F^{(2)}(x) = \frac{1}{A} \left(\sum_{j=1}^{N-2} u_j^{(2)}(f_j - 2f_{j+1} + f_{j+2}) + u_{N-1}^{(2)}(f_{N-1} - f_N) + u_N^{(1)}f_N \right).$$

Proof. We use $\langle x, f_j \rangle - q_j = u_j^{(1)} - u_{j-1}^{(1)}$ and $u_j^{(1)} = u_j^{(2)} - u_{j-1}^{(2)}$ to obtain

$$\begin{aligned} x - Q_F^{(2)}(x) &= \frac{1}{A} \sum_{j=1}^N (\langle x, f_j \rangle - q_j) f_j = \frac{1}{A} \sum_{j=1}^N (u_j^{(1)} - u_{j-1}^{(1)}) f_j \\ &= \frac{1}{A} \left(\sum_{j=1}^{N-1} u_j^{(1)}(f_j - f_{j+1}) + u_N^{(1)}f_N \right) \\ &= \frac{1}{A} \left(\sum_{j=1}^{N-1} (u_j^{(2)} - u_{j-1}^{(2)})(f_j - f_{j+1}) + u_N^{(1)}f_N \right) \\ &= \frac{1}{A} \left(\sum_{j=1}^{N-2} u_j^{(2)}(f_j - 2f_{j+1} + f_{j+2}) + u_{N-1}^{(2)}(f_{N-1} - f_N) + u_N^{(1)}f_N \right). \end{aligned}$$

□

Definition 2.13. *Given a frame $F = \{f_1, f_2, \dots, f_N\}$ for a Hilbert space \mathcal{H} , we denote its norm as $\|F\|_\infty = \max_{1 \leq j \leq N} \|f_j\|$. If \mathcal{H} is a real Hilbert space, we define the **maximal error** $\mathcal{E}(Q_F^{(2)})$ of the sigma-delta quantizer $Q_F^{(2)}$ to be*

$$\mathcal{E}(Q_F^{(2)}) := \sup\{\|x - Q_F^{(2)}(x)\| : x \in \mathcal{H}, \|x\| \leq (L - 3\epsilon)/\|F\|_\infty\},$$

where ϵ is the accuracy of the (scalar) quantizer Q on $[-L, +L]$, the initial values for the iteration have been chosen as $u_0^{(1)} = 0$ and $u_0^{(2)} = 0$, and F is the tight, ordered frame that is used to define the second-order sigma-delta quantizer $Q_F^{(2)}$ on \mathcal{H} .

Definition 2.14. *Given a frame $F = \{f_1, \dots, f_N\}$, for a normed vector space, we set*

$$T_0^{(2)}(F) = \max\{\|\pm(f_1 - 2f_2 + f_3) \pm (f_2 - 2f_3 + f_4) \dots \pm (f_{N-2} - 2f_{N-1} + f_N)\|\},$$

and

$$T^{(2)}(F) = \max\{\|\pm(f_1 - 2f_2 + f_3) \dots \pm (f_{N-2} - 2f_{N-1} + f_N) \pm (f_{N-1} - f_N) \pm 2f_N\|\}$$

where the maxima are each taken over all changes of sign.

Proposition 2.15. *Given an A -tight frame F for a real Hilbert space \mathcal{H} and a quantizer Q with accuracy $\epsilon > 0$ on $[-L, +L]$, then*

$$\mathcal{E}(Q_F^{(2)}) \leq \frac{\epsilon}{A} T^{(2)}(F).$$

Proof. We recall that by definition of $\mathcal{E}(Q_F^{(2)})$, the error is maximized over $x \in \mathcal{H} : \|x\| \leq (L - 3\epsilon)/\|F\|_\infty$. For any such x , the sequence of frame coefficients is bounded by $\|\langle x, f_j \rangle\|_\infty \leq L - 3\epsilon$, so the quantization process is stable. Therefore

$$\|x - Q_F^{(2)}(x)\| \leq \max_{|s_j| \leq 1} \frac{\epsilon}{A} \left\| \sum_{j=1}^{N-2} s_j (f_j - 2f_{j+1} + f_{j+2}) + s_{N-1} (f_{N-1} - f_N) + 2s_N f_N \right\|.$$

The right-hand side of this inequality can be regarded as the norm of a linear map that takes \mathbb{R}^N equipped with the max-norm to the Hilbert space \mathcal{H} . A convexity argument now shows the norm of such a linear map must be attained at an extreme point of the unit ball of the domain and these are the vectors all of whose entries are $s_j = \pm 1$. \square

Since the second-order sigma-delta algorithm applied to general sequences does not guarantee $u_N = 0$, we cannot derive a bound for the reconstruction error better than $2\epsilon\|f_N\|/A$. Consequently, for uniform N/d -tight frames in \mathbb{R}^d , the error bound cannot decay faster than $2\epsilon d/N$. This is in strong contrast to the result on sigma-delta modulation for oversampling of band-limited functions [DD03]. Matching the asymptotics of the error bound for second-order sigma-delta is the motivation for introducing non-uniform tight frames that terminate at the zero vector. This strategy is in the same spirit as the modulated version of the standard second-order sigma-delta algorithm for sequences.

2.4. Zero-terminating projections. Constructing frames that terminate in $f_N = 0$ is fairly straightforward using the so-called projection method. In fact, we know that any tight frame is obtained from projecting an orthonormal system of vectors together with an overall scalar multiplication, see e.g. [Nai43], [AG93, Vol. 2, Appendix I] or [Cas00, Theorem 4.10], so the following construction gives rise to all zero-terminated frames.

Lemma 2.16. *If $F = \{f_1, \dots, f_N\}$ is an A -tight frame for \mathcal{H} and $P : \mathcal{H} \rightarrow \mathcal{K}$ is the orthogonal projection onto a subspace \mathcal{K} of \mathcal{H} , then $P(F) = \{P(f_1), \dots, P(f_N)\}$ is an A -tight frame for \mathcal{K} . So for any $x \in \mathcal{K}$, we can reconstruct*

$$x = \frac{1}{A} \sum_{j=1}^N \langle x, P f_j \rangle P f_j$$

from the frame coefficients of x with respect to $P(F)$.

Proof. Since $x = Px$ and F is an A -tight frame, we have

$$x = P(Px) = \frac{1}{A} \sum_{j=1}^N \langle Px, f_j \rangle P f_j.$$

Now by the self-adjointness of P , $\langle Px, f_j \rangle = \langle x, P f_j \rangle$ for all $j \in \{1, 2, \dots, N\}$, which gives the desired reconstruction identity. \square

Proposition 2.17. *Let $F = \{f_1, f_2, \dots, f_N\}$ be an A -tight frame for \mathcal{H} and let P be the orthogonal projection onto the orthogonal complement of $\{f_N\}$ in \mathcal{H} . Then for any $x \in \{f_N\}^\perp$, the reconstruction error of the standard second-order sigma-delta quantization using the projected frame $P(F) = \{Pf_1, Pf_2, \dots, Pf_N\}$ satisfies*

$$x - Q_{P(F)}^{(2)}(x) = \frac{1}{A} \left(\sum_{j=1}^{N-2} u_j^{(2)} P(f_j - 2f_{j+1} + f_{j+2}) + u_{N-1}^{(2)} Pf_{N-1} \right).$$

Proof. Follows directly from Proposition 2.12 together with $Pf_N = 0$. \square

Theorem 2.18. *Consider an A -tight frame F for a real Hilbert space \mathcal{H} and a quantizer Q with accuracy $\epsilon > 0$ on $[-L, +L]$. If P projects orthogonally onto $\{f_N\}^\perp$, then*

$$\mathcal{E}(Q_{P(F)}^{(2)}) \leq \frac{\epsilon}{A} (T_0^{(2)}(F) + \|Pf_{N-1}\|).$$

Proof. Following along the lines of the proof of Proposition 2.15, together with Minkowski's inequality and with the fact that $Pf_N = 0$, we have

$$\mathcal{E}(Q_{P(F)}^{(2)}) \leq \frac{\epsilon}{A} (T_0^{(2)}(P(F)) + \|Pf_{N-1}\|).$$

The right-hand side of this inequality increases further when replacing $T_0^{(2)}(P(F))$ with $T_0^{(2)}(F)$, because P is a projection, so for each choice of $\{s_j\}$,

$$\left\| \sum_{j=1}^{N-2} s_j P(f_j - 2f_{j+1} + f_{j+2}) \right\| \leq \left\| \sum_{j=1}^{N-2} s_j (f_j - 2f_{j+1} + f_{j+2}) \right\|.$$

\square

3. HIGHER ORDER SIGMA-DELTA QUANTIZATION

This section introduces a standard higher-order sigma-delta quantization algorithm. Our goal for the remainder of the paper is to generalize the use of zero-terminating projections to any order of the sigma-delta algorithm.

3.1. Encoding of sequences.

Definition 3.1. *Let Q be a quantizer with accuracy $\epsilon > 0$ on the interval $[-L, +L]$. Given an input sequence $\{x_j\}_{j=1}^\infty$, then the **standard K th-order sigma-delta quantized sequence** $\{q_j\}_{j=1}^\infty$ associated with the initialization vector $u_0 \in \mathbb{R}^K$ is defined as follows: For each $j \in \mathbb{N}$, we denote the vector of the iterated cumulative error variables as $u_j = (u_j^{(1)}, u_j^{(2)}, \dots, u_j^{(K)})$. The quantized output values are obtained by inductively defining q_j and u_j for $j \in \mathbb{N}$ with the prescription*

$$q_j := Q\left(x_j + \sum_{l=1}^K u_{j-1}^{(l)}\right),$$

and the map for updating the internal variables

$$u_j^{(0)} := x_j - q_j$$

and

$$u_j^{(l)} := u_{j-1}^{(l)} + u_j^{(l-1)}, 1 \leq l \leq K.$$

Proposition 3.2. *Given a quantizer Q with accuracy $\epsilon > 0$ on $[-L, L]$. If the input sequence $\{x_j\}$ is bounded by*

$$\|x\|_\infty \leq L - (2^K - 1)\epsilon$$

and initially, the internal variables satisfy

$$|u_0^{(l)}| \leq 2^{K-l}\epsilon,$$

then the above iteration implies that each sequence $\{u_j^{(l)}\}_{j \in \mathbb{N}}$ is bounded by

$$\|u^{(l)}\|_\infty \leq 2^{K-l}\epsilon.$$

Proof. We proceed by induction. If for a given $j \in \mathbb{N}$,

$$|u_{j-1}^{(l)}| \leq 2^{K-l}\epsilon,$$

then we can bound the sum

$$\left| \sum_{l=1}^K u_{j-1}^{(l)} \right| \leq (2^K - 1)\epsilon.$$

and with the bound on the input sequence, the argument of the quantizer is within $[-L, +L]$. The iterative assignment of $u_j^{(l)}$ implies that for each $1 \leq l \leq K$,

$$u_j^{(l)} = \sum_{k=1}^l u_{j-1}^{(k)} + x_j - q_j.$$

Because of the accuracy of the quantizer on $[-L, L]$,

$$|u_j^{(K)}| = |x_j + \sum_{l=1}^K u_{j-1}^{(l)} - q_j| \leq \epsilon.$$

Now we can deduce from $u_j^{(l)} = u_j^{(l+1)} - u_{j-1}^{(l+1)}$ in conjunction with the assumed bounds on u_{j-1} that

$$|u_j^{(K-l)}| \leq 2^l \epsilon.$$

□

3.2. Encoding of vectors.

Definition 3.3. *Let Q be a quantizer with accuracy $\epsilon > 0$ on $[-L, +L]$, let $F = \{f_j\}_{j=1}^N$ be an A -tight (ordered) frame for a real Hilbert space \mathcal{H} and let $x \in \mathcal{H}$. Then the K th-order sigma-delta quantized vector $Q_F^{(K)}(x)$ is defined by*

$$Q_F^{(K)}(x) := \frac{1}{A} \sum_{j=1}^N q_j f_j,$$

where the frame coefficients have been quantized by selecting $\mathbf{u}_0^{(l)} = \mathbf{0}$ for $1 \leq l \leq K$ and the quantized frame coefficients are

$$q_j := Q(\langle x, f_j \rangle) + \sum_{l=1}^K u_{j-1}^{(l)}$$

while the internal variables are updated according to

$$u_j^{(0)} := \langle x, f_j \rangle - q_j$$

and

$$u_j^{(l)} := u_{j-1}^{(l)} + u_j^{(l-1)}, 1 \leq l \leq K.$$

We henceforth call the map $Q_F^{(K)}$ a ***K***th-order ***sigma-delta*** quantizer on \mathcal{H} .

Definition 3.4. We denote the unilateral shift operator as S . It acts on sequences by $(Sx)_j = x_{j-1}$ for $j > 1$. By convention, we set $x_0 = 0$, so $(Sx)_1 = 0$. The shift operator also applies to frames, that is, vector-valued sequences, by $(Sf)_j = f_{j-1}$ and $(Sf)_1 = 0$. The adjoint S^* gives $(S^*x)_j = x_{j+1}$, $j \in \mathbb{N}$. Together with the identity operator I , we abbreviate the ***K***th ***finite backward difference operator*** acting on a sequence x as $(I - S)^K x$.

In order to simplify many of the formulas that we encounter in this section, it is convenient to set various quantities equal to zero when they fall outside the range of their definition. This is often referred to as *zero-padding*. Thus, for example, instead of given a tight frame, $F = \{f_1, \dots, f_N\}$, we often consider the zero-padded set $\{f_1, f_2, \dots, f_N, f_{N+1}, \dots\}$ with $f_j = 0$, for all $j \geq N + 1$. Alternatively, we then speak of a tight frame $F = \{f_j\}_{j \in \mathbb{N}}$ with support $\{1, 2, \dots, N\}$.

The following quantities will play a key role in our error estimates.

Definition 3.5. Given a frame $F = \{f_j\}_{j \in \mathbb{N}}$ for a normed vector space over \mathbb{R} , with F supported on the index set $\{1, 2, \dots, N\}$, we define

$$T_0^{(K)}(F) := \max\{\|\pm (I - S^*)^K f_1 \pm (I - S^*)^K f_2 + \dots \pm (I - S^*)^K f_{N-K}\|\},$$

and

$$T^{(K)}(F) := \max\{\|\pm (I - S^*)^K f_1 \pm (I - S^*)^K f_2 \dots \pm (I - S^*)^K f_N\|\}$$

where the maxima are each taken over all changes of sign.

The convenience of the notation is at the cost of obscuring that, for example, for a frame supported on $\{1, 2, \dots, N\}$, the last term in $T^{(1)}(F)$ is $\pm(f_N - f_{N+1}) = \pm f_N$.

Proposition 3.6. Let $F = \{f_j\}_{j \in \mathbb{N}}$ be a finitely supported A -tight frame for a real Hilbert space \mathcal{H} , then for any $x \in \mathcal{H}$, the reconstruction error of the standard K th-order sigma-delta quantization is

$$x - Q_F^{(K)}(x) = \frac{1}{A} \sum_{j=1}^{\infty} u_j^{(K)} ((I - S^*)^K f)_j.$$

If F has support $\{1, 2, \dots, N\}$, making the boundary terms in this identity explicit gives

$$x - Q_F^{(K)}(x) = \frac{1}{A} \left(\sum_{j=1}^{N-K} u_j^{(K)} ((I - S^*)^K f)_j + \sum_{j=N-K+1}^N u_j^{(N+1-j)} ((I - S^*)^{N-j} f)_j \right).$$

Proof. Analogous to the proof of Proposition 2.12. \square

Definition 3.7. Given a tight frame $F = \{f_1, f_2, \dots, f_N\}$ for a real Hilbert space \mathcal{H} , we define the **maximal error** $\mathcal{E}(Q_F^{(K)})$ of the standard K th-order sigma-delta quantizer $Q_F^{(K)}$ to be

$$\mathcal{E}(Q_F^{(K)}) := \sup\{\|x - Q_F^{(K)}(x)\| : x \in \mathcal{H}, \|x\| \leq (L - (2^K - 1)\epsilon)/\|F\|_\infty\},$$

where ϵ is the accuracy of the (scalar) quantizer Q on $[-L, L]$, and the initial values for the iteration have been chosen as $u_0^{(l)} = 0$ for all $1 \leq l \leq K$.

Proposition 3.8. Given a finitely supported, A -tight frame F for a real Hilbert space \mathcal{H} and a quantizer Q with accuracy $\epsilon > 0$ on $[-L, +L]$, then

$$\mathcal{E}(Q_F^{(K)}) \leq \frac{\epsilon}{A} T^{(K)}(F).$$

Proof. We recall that by definition of $\mathcal{E}(Q_F^{(K)})$, the error is maximized over $x \in \mathcal{H} : \|x\| \leq (L - (2^K - 1)\epsilon)/\|F\|_\infty$. For any such x , the sequence of frame coefficients is bounded by $\|\{\langle x, f_j \rangle\}\|_\infty \leq L - (2^K - 1)\epsilon$, so the quantization process is stable and the estimates for the internal variables $u_j^{(l)}$ apply. Let $\{1, 2, \dots, N\}$ be the support of F , then

$$\|x - Q_F^{(K)}(x)\| \leq \max_{|s_j| \leq 1} \frac{\epsilon}{A} \left\| \sum_{j=1}^N s_j ((I - S^*)^K f)_j \right\|.$$

Again, by convexity the maximum must be attained at an extreme point of the domain where the entries of s are $s_j = \pm 1$. \square

4. SMOOTHLY TERMINATED FRAME PATHS

As demonstrated in a previous paper [BP], many families of frames come from regular sampling along a smooth path in \mathbb{R}^d and for these types of frames it is fairly easy to see that the quantity $T_0^{(K)}(F)$ is uniformly bounded independent of N . Moreover, if the path and sufficiently many of its derivatives terminate in the zero vector, then we can use zero-padding to derive analogous estimates for $T(F)$, and hence, for the magnitude of the error caused by higher order sigma-delta quantization. We make this precise in the following definition. Since the notion of frame paths is equally interesting in the complex case, we include complex Hilbert spaces.

Definition 4.1. Let \mathcal{H} be a finite dimensional Hilbert space, real or complex. A continuous map $f : [a, b] \rightarrow \mathcal{H}$ is called a **frame path for regular sampling** provided that there is a constant $C > 0$ and infinitely many choices of N such that the set $F_N = \{f(a + \frac{b-a}{N}), f(a + \frac{2(b-a)}{N}), \dots, f(b)\}$ is an NC -tight frame for \mathcal{H} . If $f(b) = 0$, we call f a **zero-terminated frame path**. If f is $K - 1$ -times continuously differentiable with $f^{(K-1)}$ absolutely continuous and $f(b) = \dots = f^{(K-1)}(b) = 0$, then we call f a **K th-order zero-terminated frame path**.

Note that if f is frame path for regular sampling that is zero-terminated of order K , then we may extend the domain of definition of f for $t > b$ by setting $f(t) = 0, t \geq b$, and this extended function will be $K - 1$ -times continuously differentiable with $f^{(K-1)}$ absolutely continuous.

4.1. Examples of smoothly terminated frame paths.

Example 4.2 (The Modulated Basis Repetition). Let $\{e_i\}_{i=1}^d$ be the canonical basis of \mathbb{R}^d and let $f : [0, d] \rightarrow \mathbb{R}^d$ be piecewise defined by $f(t) = (1 - \cos(2\pi t))e_j$ for $j - 1 \leq t \leq j$. Then for any $N = Md$, with $3 \leq M \in \mathbb{N}$, regular sampling gives a $3N/2d$ -tight frame F_N with $\|F\|_\infty \leq 2$. This frame path terminates at $f(d) = 0$, and $f'(d) = 0$, so it is a second-order zero-terminated frame path.

In order to convert a sufficiently smooth frame path into a zero-terminated frame path of order K , we use the orthogonal projection $P : \mathcal{H} \rightarrow \{f(b), \dots, f^{(K-1)}(b)\}^\perp$. More precisely, given a frame path $f : [a, b] \rightarrow \mathcal{H}$ for regular sampling, with constant $C > 0$, then the projection $g(t) = Pf(t)$ is a K th-order zero-terminated frame path for regular sampling with constant $C > 0$ on the subspace $\{f(b), \dots, f^{(K-1)}(b)\}^\perp$ of codimension at least $\dim(\mathcal{H}) - K$.

Definition 4.3. *Given a frame path $f : [a, b] \rightarrow \mathcal{H}$ for regular sampling, with constant $C > 0$, that has $K-1$ absolutely continuous derivatives, we call the frame path $g : [a, b] \rightarrow \{f(b), \dots, f^{(K-1)}(b)\}^\perp$ described above the **K th-order zero-terminating projection of the frame path** and we let $G_N = \{g(a + \frac{b-a}{N}), \dots, g(b)\}$ denote the corresponding NC -tight frames.*

Example 4.4 (Complex Fourier Frame). Let $f(t) = \frac{1}{\sqrt{d}}(e^{2\pi it}, e^{4\pi it}, \dots, e^{2d\pi it})$ for $t \in [0, 1]$. It is fairly easily checked that the $N \times d$ matrix whose rows are the vectors $\{f(j/N)\}_{j=1}^N$ in F_N has orthonormal columns and hence these vectors (and their complex conjugates) define an isometry from \mathbb{C}^d to \mathbb{C}^N . Thus, F_N is a uniform N/d -tight frame for every $N \geq d$, and $f(t)$ is a uniform frame path for regular sampling. The subspace $\{f(1)\}^\perp$ is the subspace consisting of vectors whose coordinates sum to zero. Thus, the zero-terminating projection of f will yield N/d -tight frames G_N , for this $(d-1)$ -dimensional subspace.

Example 4.5 (The Harmonic Frames). When $d = 2k$ is even these are defined by regularly sampling the path $f(t) = \sqrt{\frac{2}{d}}(\cos(2\pi t), \sin(2\pi t), \cos(4\pi t), \sin(4\pi t), \dots, \cos(2\pi kt), \sin(2\pi kt))$ in the interval $[0, 1]$.

When $d = 2k + 1$, the harmonic frames are defined by regularly sampling $f(t) = \sqrt{\frac{2}{d}}(\frac{1}{\sqrt{2}}, \cos(2\pi t), \sin(2\pi t), \dots, \cos(2\pi kt), \sin(2\pi kt))$ in the interval $[0, 1]$. When $N > d$, the vectors in F_N are a uniform N/d -tight frame.

The set of vectors F_N that one obtains this way was one of the earliest examples of a uniform N/d -tight frame as introduced in [GVT98]. The fact that the map f is a uniform frame path is verified fairly easily by taking real and imaginary parts of the complex Fourier frame.

The zero-terminated frame path obtained by projecting the harmonic frame path onto $\{f(1)\}^\perp$ is explicitly computed as

$$\begin{aligned}
 g(t) &= f(t) - \langle f(t), f(1) \rangle f(1) \\
 &= \sqrt{\frac{2}{d}} \left((\cos(2\pi t), \sin(2\pi t), \cos(4\pi t), \sin(4\pi t), \dots, \cos(2\pi kt), \sin(2\pi kt)) \right. \\
 &\quad \left. - \left(\frac{\sin(\pi(d+1)t)}{d \sin(\pi t)} - \frac{1}{d} \right) (1, 0, 1, 0, \dots, 1, 0) \right)
 \end{aligned}$$

when $d = 2k$ is even, and

$$\begin{aligned}
 g(t) &= \sqrt{\frac{2}{d}} \left(\left(\frac{1}{\sqrt{2}}, \cos(2\pi t), \sin(2\pi t), \cos(4\pi t), \sin(4\pi t), \dots, \cos(2\pi kt), \sin(2\pi kt) \right) \right. \\
 &\quad \left. - \frac{\sin(\pi dt)}{d \sin(\pi t)} \left(\frac{1}{\sqrt{2}}, 1, 0, 1, 0, \dots, 1, 0 \right) \right).
 \end{aligned}$$

when $d = 2k + 1$ is odd. Regular sampling of g yields G_N , a frame spanning the $(d-1)$ -dimensional subspace of \mathbb{R}^d orthogonal to $(1, 0, \dots, 1, 0)$ or $(\frac{1}{\sqrt{2}}, 1, 0, \dots, 1, 0)$ for even or odd d , respectively.

Since the harmonic frame path satisfies $\|f(t)\| = 1$, we know $f(1) \perp f'(1)$. Therefore the second-order zero-terminated path h is obtained from the above path by an additional projection,

$$h(t) = f(t) - \langle f(t), f(1) \rangle f(1) - \frac{\langle f(t), f'(1) \rangle}{\|f'(1)\|^2} f'(1).$$

We compute $\|f'(t)\|^2 = \frac{8\pi^2}{d} \frac{k(k+1)(2k+1)}{6}$ where $d = 2k$ if d is even and $d = 2k + 1$ if it is odd.

For even d , we arrive at the second-order zero-terminated frame path

$$h(t) = g(t) - \sqrt{\frac{2}{d}} \frac{3(d \sin(\pi t) \cos(\pi(d+1)t) - \sin(\pi dt))}{d(d+1)(d+2) \sin^2(\pi t)} (0, 2, 0, 3, \dots, 0, d)$$

and for odd d , we get

$$h(t) = g(t) - \sqrt{\frac{2}{d}} \frac{3((d-1) \sin(\pi t) \cos(\pi dt) + \sin(\pi(1-d)t))}{(d-1)d(d+1) \sin^2(\pi t)} (0, 0, 2, 0, 3, \dots, 0, d-1).$$

To illustrate the projected frame paths, we have included a plot of the zero-terminating projection of the harmonic frame in $d = 3$.

We also include the second-order zero-terminating projection of the harmonic frame path in $d = 4$, together with sample points for $N = 40$ vectors.

We call a frame path $f : [a, b] \rightarrow \mathbb{R}^d$ **reversible** if for all $t \in [a, b]$, $f(t) = f(a+b-t)$. In a previous paper, we raised the question of a semicircle-version for harmonic frames. We may now give a construction of this type of frame, albeit a non-uniform one.

Example 4.6 (The Reversible Harmonic Frame and Harmonic Semicircle Frames). The reversible harmonic frames are defined by regularly sampling the path $f(t) = \sqrt{\frac{2}{d}} (\frac{1}{\sqrt{2}}, \cos(\pi t), \cos(2\pi t), \dots, \cos(\pi(d-1)t))$ in the interval $[0, 2]$. Since only cosines appear in its entries, indeed $f(t) = f(2-t)$. When $N > d$, the vectors in F_N are a non-uniform $N/2d$ -tight frame. This can be seen from the fact that they are obtained from

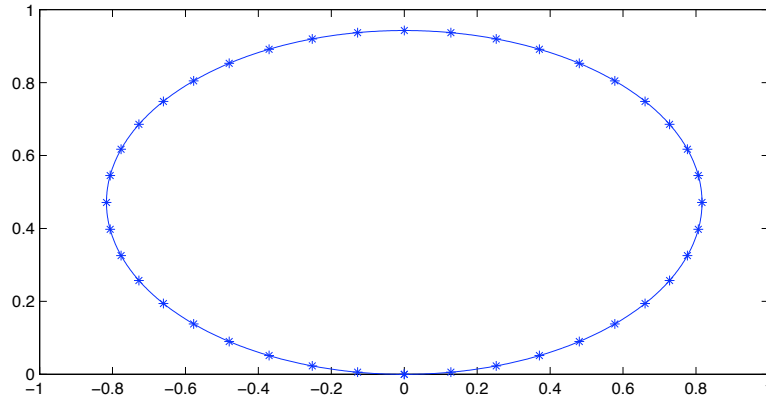


FIGURE 1. Zero-terminating projection of the harmonic frame path for $d = 3$, together with sample points for a frame with $N = 40$ vectors. The coordinate system is given by the two basis vectors $e_1 = (0, 0, 1)$ and $e_2 = \sqrt{2/3}(1, 0, -\frac{1}{2})$.

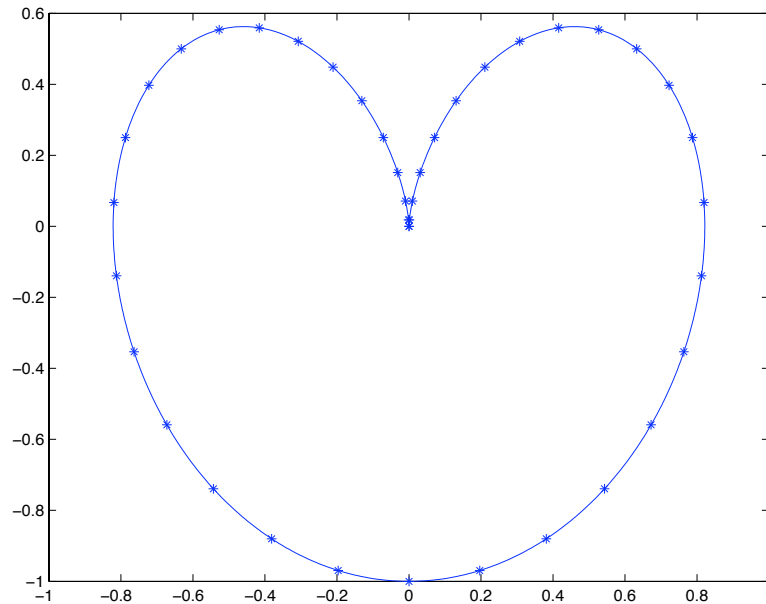


FIGURE 2. Second-order zero-terminating projection of the harmonic frame path for $d = 4$, with sample points for $N = 40$ vectors. The axes correspond to the basis vectors $e_1 = \sqrt{1/2}(1, 0, -1, 0)$ and $e_2 = \sqrt{1/5}(0, 2, 0, -1)$.

projecting the harmonic frame path in \mathbb{R}^{2d-1} parametrized by $0 \leq t \leq 2$ onto the d -dimensional subspace spanned by the odd-numbered canonical basis vectors, followed by an appropriate rescaling. Zero-termination is obtained by further projecting this frame

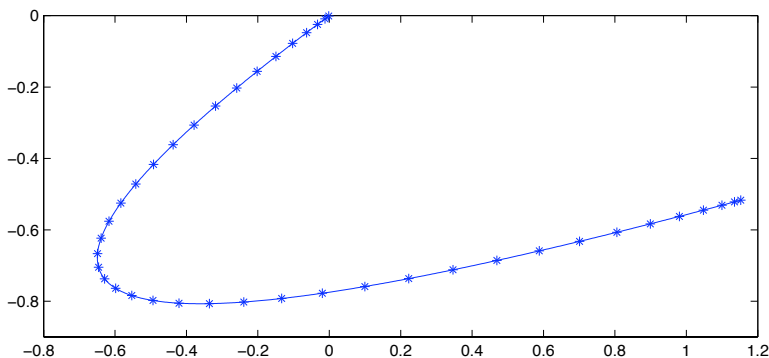


FIGURE 3. Second-order zero-terminating projection of the harmonic semicircle frame path for $d = 3$, with sample points for $N = 40$ vectors. The axes correspond to the basis vectors $e_1 = \sqrt{1/2}(0, 1, 1)$ and $e_2 = \sqrt{1/10}(-2\sqrt{2}, -1, 1)$.

path onto $\{f(2)\}^\perp$. The projected reversible harmonic frame path is given by

$$g(t) = \sqrt{\frac{2}{d}} \left(\left(\frac{1}{\sqrt{2}}, \cos(\pi t), \cos(2\pi t), \dots, \cos(\pi(d-1)t) \right) - \frac{\sin(\pi(d-\frac{1}{2})t)}{(d-\frac{1}{2})\sin(\pi t)} \left(\frac{1}{\sqrt{2}}, 1, 1, \dots, 1 \right) \right)$$

Because of the reversal-symmetry of the reversible harmonic path, it is enough to sample on the semicircle parametrized by $0 \leq t \leq 1$, which results in the so-called harmonic semicircle frames. More precisely, if the number N of frame vectors is even and the regular sampling starts with an offset, $F_N = \{f(1/2N), f(3/2N), \dots, f((1-1/2N))\}$, then $f_{N/2-j} = f_{N/2+j+1}$ for any $j \in \{1, 2, \dots, N/2-1\}$, and half of the frame vectors and coefficients can be eliminated by symmetry.

To obtain a zero-terminated path from the harmonic semicircle path, we project onto $\{f(1)\}^\perp$, which results in

$$g(t) = \sqrt{\frac{2}{d}} \left(\left(\frac{1}{\sqrt{2}}, \cos(\pi t), \cos(2\pi t), \dots, \cos(\pi(d-1)t) \right) - \frac{\sin(\pi(d-\frac{1}{2})(t+1))}{(d-\frac{1}{2})\sin(\pi(t+1))} \left(\frac{1}{\sqrt{2}}, -1, 1, -1, \dots, (-1)^{d-1} \right) \right)$$

The frames $\{G_N\}$ resulting from sampling are seen to be N/d -tight and satisfy $\|G_N\|_\infty \leq 2 - \frac{1}{d}$.

We have included an illustration of the zero-terminating projection of the harmonic semicircle path. Since only cosines were used for the frame path, the projection also has vanishing derivatives, $g'(0) = g'(1) = 0$. For this reason, g is already a second-order zero-terminating frame path.

4.2. Estimates for the maximal error of frames obtained by sampling. The following observations help to obtain bounds on $T_0^{(K)}(F_N)$ that are independent of N for these frames.

Definition 4.7. Given a frame path $f : [a, b] \rightarrow \mathbb{R}^d$, we extend it by zero outside of its original domain and define for any displacement $\Delta t \in \mathbb{R}$ the path shifted by Δt to be

$$S_{\Delta t} f(t) := f(t - \Delta t).$$

For consistency of notation, we also identify $S_{\Delta t}^* = S_{-\Delta t}$. With the identity operator I , the **finite K th-order forward difference of f at t_j** is written as $(I - S_{\Delta t}^*)^K f(t_j)$.

Definition 4.8. Given a path $f : [a, b] \rightarrow \mathcal{H}$ in a real or complex Hilbert space \mathcal{H} , with $K - 1$ absolutely continuous derivatives, we define the **K th-order total directional variation of f** to be

$$V^{(K)}(f) = \sup \left\{ \int_a^b |\langle f^{(K)}(t), v \rangle| dt : \|v\| = 1 \right\}.$$

The following two lemmas are useful for estimating $T^{(K)}$ and $T_0^{(K)}$.

Lemma 4.9. Let $F = \{f_j\}_{j \in \mathbb{N}}$ be a frame for a finite dimensional Hilbert space over \mathbb{R} , with the support of F contained in $\{1, 2, \dots, N\}$, then

$$T_0^{(K)}(F) = \max \left\{ \sum_{j=1}^{N-K} |\langle (I - S^*)^K f_j, v \rangle| : \|v\| = 1 \right\},$$

and

$$T^{(K)}(F) = \max \left\{ \sum_{j=1}^N |\langle (I - S^*)^K f_j, v \rangle| : \|v\| = 1 \right\}.$$

Proof. The norm of any vector is attained by taking the inner product with a unit vector. For any vector v , there exists a choice of signs to make all the inner products non-negative. \square

Lemma 4.10. Let f be a frame path with $K - 1$ absolutely continuous derivatives on the interval $[t_1, t_N]$, which is split into $N - 1$ subintervals $\{[t_j, t_{j+1}]\}_{j=1}^{N-1}$ of equal length Δt . Then

$$\sum_{j=1}^{N-K} |\langle (I - S_{\Delta t}^*)^K f(t_j), v \rangle| \leq (\Delta t)^{K-1} \int_{t_1}^{t_N} |\langle f^{(K)}(s), v \rangle| ds.$$

Proof. Let $\{\beta_{j,K}\}_{j \in \mathbb{Z}}$ be the K th-order B-spline functions of width Δt . That is, $\beta_{j,1}(t) = \chi_{[t_j, t_{j+1})}$ and $\beta_{j,l} = \beta_{j,l-1} * \beta_{0,1}$ for any $l > 1$. We recall that the support of $\beta_{j,K}$ is $[t_j, t_j + K\Delta t]$, that all $\beta_{j,K} \geq 0$ and that $\sum_{j=-\infty}^{\infty} \beta_{j,K}(t) = 1$. Repeatedly integrating by parts yields that

$$(I - S_{\Delta t}^*)^K f(t_j) = (\Delta t)^{K-1} \int_{t_j}^{t_j + K\Delta t} f^{(K)}(s) \beta_{j,K}(t) ds.$$

Using the non-negativity of $\beta_{j,K}$, we estimate

$$|\langle (I - S_{\Delta t}^*)^K f(t_j), v \rangle| = (\Delta t)^{K-1} \int_{t_j}^{t_j + K\Delta t} |\langle f^{(K)}(s), v \rangle| \beta_{j,K}(s) ds$$

and when summing, $\sum_{j=1}^{N-K} \beta_{K,j}(s) \leq \chi_{[t_1, t_N]}(s)$ helps further estimate

$$\sum_{j=1}^{N-K} |\langle (I - S_{\Delta t}^*)^K f(t_j), v \rangle| \leq (\Delta t)^{K-1} \int_{t_1}^{t_N} |\langle f^{(K)}(s), v \rangle| ds.$$

□

Theorem 4.11. *Let $f : [a, b] \rightarrow \mathbb{R}^d$ be a frame path with $K - 1$ absolutely continuous derivatives and let $\{F_N\}_{N \in \mathbb{U}}$ be the frames obtained by regular sampling, then*

$$T_0^{(K)}(F_N) \leq \left(\frac{b-a}{N}\right)^{K-1} V^{(K)}(f).$$

If f is a K -th order zero terminated frame path, then

$$T^{(K)}(F_N) \leq \left(\frac{b-a}{N}\right)^{K-1} V^{(K)}(f).$$

Proof. The first inequality is a direct consequence of Lemma 4.9 and Lemma 4.10, with $\Delta t = \frac{b-a}{N}$, $t_1 = a$, $t_N = b$.

For the second inequality, extend the domain of definition to $t > b$ by setting $f(t) = 0$, then f still has $K - 1$ absolutely continuous derivatives on $[a, +\infty)$. Applying Lemma 4.10, with $\Delta t = \frac{b-a}{N}$ and $M = N + K$, we have that,

$$\begin{aligned} \sum_{j=1}^N |\langle (I - S^*)^K f(t_j), v \rangle| &= \sum_{j=1}^{M-K} |\langle (I - S^*)^K f(t_j), v \rangle| \\ &\leq (\Delta t)^{K-1} \int_{t_1}^{t_M} |\langle f^{(K)}(s), v \rangle| ds = \left(\frac{b-a}{N}\right)^{K-1} V^{(K)}(f), \end{aligned}$$

since $f^{(K)}(s) = 0$, for $t_N \leq s \leq t_M$. □

The following is an immediate consequence.

Corollary 4.12. *Let $f : [a, b] \rightarrow \mathbb{R}^d$ be a K th-order zero terminated frame path with constant $C > 0$ and let $\{F_N\}_{N \in \mathbb{U}}$ be the NC -tight frames obtained by regular sampling. If Q is any quantizer of accuracy $\epsilon > 0$, then*

$$\mathcal{E}(Q_{F_N}^{(K)}) \leq \frac{\epsilon(b-a)^{K-1}}{CN^K} V^{(K)}(f).$$

Example 4.13. Let $f : [0, d] \rightarrow \mathbb{R}^d$ be the frame path for modulated basis repetition, piecewise defined by $f(t) = (1 - \cos(2\pi t))e_i$ for $i-1 \leq t \leq i$, where e_i is the i th canonical basis vector in \mathbb{R}^d . After fixing $v \in \mathbb{R}^d$, we estimate using the Cauchy-Schwarz inequality,

$$\int_0^d |\langle f''(t), v \rangle| dt = \sum_{i=1}^d (2\pi)^2 \int_0^1 |\cos(2\pi t)| |v_i| dt = \sum_{i=1}^d 8\pi |v_i| \leq 8\pi \sqrt{d} \|v\|.$$

Therefore,

$$V^{(2)}(f) \leq 8\pi \sqrt{d}.$$

The path f is a second-order zero-terminated frame path, with constant $C = 3/2d$. Using the preceding corollary, we have that the maximal second-order quantization error for the $3N/2d$ -tight frame $F_N = \{f_1, f_2, \dots, f_N\}$ obtained from regular sampling is

$$\mathcal{E}(Q_{G_N}^{(2)}) \leq \frac{16\pi \epsilon d^{5/2}}{3N^2}.$$

Corollary 4.14. *Let $f : [a, b] \rightarrow \mathbb{R}^d$ be a frame path with constant $C > 0$ and with $K - 1$ absolutely continuous derivatives. Let g be the K th-order zero-terminating projection of f , and denote $\{G_N\}_{N \in \mathbb{U}}$ the NC -tight frames obtained from sampling g . Then*

$$\mathcal{E}(Q_{G_N}^{(K)}) \leq \frac{\epsilon(b-a)^{K-1}}{CN^K} V^{(K)}(f).$$

Proof. This follows from the preceding corollary because for any orthogonal projection P , the total directional frame variation is reduced by projecting the path f ,

$$V^{(K)}(Pf) \leq V^{(K)}(f).$$

□

Theorem 4.15. *Let f be the harmonic frame path in d dimensions and g its K th-order zero-terminating projection. Then if d is even,*

$$\mathcal{E}(Q_{G_N}^{(K)}) \leq \frac{\epsilon\pi^K d^{K-1/2}}{N^K}.$$

If d is odd, we have

$$\mathcal{E}(Q_{G_N}^{(K)}) \leq \frac{\epsilon\pi^K (d-1)^K}{\sqrt{d}N^K}.$$

Proof. We first consider the case of even d and $K = 2$. Fix a unit vector $v = (v_1, v_2, \dots, v_d)$. For d even, we observe $f''(t) = -\sqrt{\frac{2}{d}}(4\pi^2 \cos(2\pi t), 4\pi^2 \sin(2\pi t), 16\pi^2 \cos(4\pi t), \dots, (d\pi)^2 \cos(d\pi t), (d\pi)^2 \sin(d\pi t))$. To obtain the bound for $V^{(2)}(g)$, we use $V^{(2)}(g) \leq V^{(2)}(f)$. Subsequently using the Cauchy-Schwarz and Hölder inequalities gives

$$\begin{aligned} V^{(2)}(g) &\leq \int_0^1 |\langle f''(t), v \rangle|^2 dt \\ &\leq (2\pi)^2 \sqrt{\frac{2}{d}} \int_0^1 | -v_1 \cos(2\pi t) - v_2 \sin(2\pi t) - \dots \\ &\quad - (d/2)^2 v_{d-1} \cos(\pi dt) - (d/2)^2 v_d \sin(\pi dt) |^2 dt \\ &\leq (2\pi)^2 \sqrt{\frac{2}{d}} \left(\frac{1}{2}(v_1^2 + v_2^2 + 2^4 v_3^3 + 2^4 v_4^2 + \dots + (d/2)^4 v_{d-1}^2 + (d/2)^4 v_d^2) \right)^{1/2} \\ &\leq (2\pi^2) \frac{1}{\sqrt{d}} (d/2)^2 = \pi^2 d^{3/2}. \end{aligned}$$

For the K th-order estimate, we get analogously

$$\begin{aligned} V^{(K)}(g) &\leq \int_0^1 |\langle f^{(K)}(t), v \rangle|^2 dt \\ &\leq (2\pi)^K \sqrt{\frac{2}{d}} \left(\frac{1}{2}(v_1^2 + v_2^2 + 2^{2K} v_3^3 + 2^{2K} v_4^2 + \dots \right. \\ &\quad \left. + (d/2)^{2K} v_{d-1}^2 + (d/2)^{2K} v_d^2) \right)^{1/2} \\ &\leq \pi^K d^{K-1/2}. \end{aligned}$$

When d is odd, we have

$$V^{(K)}(g) \leq \pi^K \frac{(d-1)^K}{\sqrt{d}}.$$

Now using the Corollary 4.14, we arrive at the desired error estimate. \square

When the frame path is not zero terminated of the appropriate order, we need to include boundary terms. We focus on the simplest case, where f is terminated to order $K-1$.

Lemma 4.16. *Let $f : [a, b] \rightarrow \mathbb{R}^d$ be a frame path, with $K-1$ derivatives that are absolutely continuous and $f(b) = f'(b) = \dots = f^{(K-2)}(b) = 0$, but $f^{(K-1)}(b) \neq 0$, then*

$$T^{(K)}(F_N) \leq \left(\frac{b-a}{N}\right)^{K-1} \sup_{\|v\|=1} \left(\int_a^b |\langle f^{(K)}(t), v \rangle| dt + |\langle f^{(K-1)}(b), v \rangle| \right).$$

Proof. We proceed by approximating f with a family of K -times continuously differentiable curves $\{g_\eta\}_{\eta>0}$ which satisfy $f(t) = g_\eta(t)$ except when $t \in (b-\eta, b]$, and $g_\eta^{(l)}(b) = 0$, $1 \leq l \leq K-1$. To this end, we multiply $f^{(K-1)}$ with a non-increasing C^∞ -function h_η , $h_\eta(t) = 1$ for $a \leq t \leq b-\eta$, $h_\eta(b) = 0$, and then integrate $h_\eta f^{(K-1)}$ ($K-1$)-times.

Then the preceding lemma implies that for all N with each $\eta < \frac{b-a}{N}$

$$T^{(K)}(F_N) \leq \left(\frac{b-a}{N}\right)^{K-1} V^{(K)}(g_\eta).$$

By construction,

$$V^{(K)}(g_\eta) = \sup_{\|v\|=1} \left(\int_a^{b-\eta} |\langle f^{(K)}(t), v \rangle| dt + \int_{b-\eta}^b |\langle (f^{(K-1)} h_\eta)'(t), v \rangle| dt \right).$$

We note that the expression we are maximizing is continuous in v .

After splitting the term $|\langle (f^{(K-1)} h_\eta)'(t), v \rangle|$ with the triangle inequality, we see that

$$\int_{b-\eta}^b |\langle f^{(K-1)} h_\eta'(t), v \rangle| dt = \int_{b-\eta}^b |h_\eta'(t)| |\langle f^{(K-1)}(t), v \rangle| dt.$$

By the continuity of $f^{(K-1)}$ and by the normalization $\int_{b-\eta}^b |h_\eta'(t)| = 1$ due to the monotonicity of h_η , in the limit $\eta \rightarrow 0$ this expression converges to $|\langle f^{(K-1)}(b), v \rangle|$.

The convergence is uniform in the compact set $\{v : \|v\| = 1\}$. If we choose the family $\{h_\eta\}$ to be monotonic in η , then the remaining terms also converge uniformly and we conclude

$$T^{(K)}(F_N) \leq \left(\frac{b-a}{N}\right)^{K-1} \sup_{\|v\|=1} \left(\int_a^b |\langle f^{(K)}(t), v \rangle| dt + |\langle f^{(K-1)}(b), v \rangle| \right).$$

\square

Theorem 4.17. *Given a frame path f with $K-1$ absolutely continuous derivatives, and its $(K-1)$ -order zero-terminating projection g . Let $\{G_N\}_{N \in \mathbb{U}}$ denote the NC-tight frames obtained from regular sampling of g . Then*

$$\mathcal{E}(Q_{G_N}^{(K)}) \leq \frac{\epsilon(b-a)^{K-1}}{CN^K} \left(V^{(K)}(f) + \|f^{(K-1)}(b)\| \right).$$

Proof. This follows from the preceding lemma by applying the Cauchy-Schwarz inequality to the boundary term, and by the inequalities $V^{(K)}(g) \leq V^{(K)}(f)$ as well as $\|g^{(K-1)}(b)\| \leq \|f^{(K-1)}(b)\|$. \square

Theorem 4.18. *Let $f : [0, 1] \rightarrow \mathbb{R}^d$ be the sample path for the harmonic frames in \mathbb{R}^d , and let $g = P(f)$ where P is the projection of \mathbb{R}^d onto $\{f(1)\}^\perp$. Then $V^{(2)}(g) \leq \pi^2 d^{3/2}$. For any frame G_N with $N > d$ and any quantizer Q of accuracy ϵ , we have*

$$\mathcal{E}(Q_{G_N}^{(2)}) \leq \frac{\pi^2 \epsilon d^{5/2}}{N^2} + \frac{\pi \sqrt{d}(d+1)^{3/2} \epsilon}{\sqrt{3} N^2},$$

where the error is by definition maximized over x orthogonal to $f(1)$, with $\|x\| \leq L - 3\epsilon$.

Proof. We begin by recalling the proof of Theorem 4.15. For $d = 2k$ even, we derived $V^{(2)}(g) \leq \pi^2 d^{3/2}$. If $d = 2k + 1$, then $V^{(2)}(g) \leq \pi^2 (d-1)^2 / \sqrt{d}$.

Instead of computing $\|f'(1)\|$ directly, we estimate the norm

$$\|f(t) - f(1)\| = \sqrt{2 - 2\langle f(t), f(1) \rangle}.$$

For even d , we have to insert

$$\langle f(t), f(1) \rangle = \frac{\sin(\pi(d+1)t)}{d \sin(\pi t)} - \frac{1}{d}$$

and for odd d

$$\langle f(t), f(1) \rangle = \frac{\sin(\pi dt)}{d \sin(\pi t)}$$

The elementary inequalities

$$\frac{\sin(\pi dt)}{d \sin(\pi t)} \geq \frac{\pi dt - \frac{1}{6} \pi^3 d^3 t^3}{\pi dt} = 1 - \frac{1}{6} \pi^2 d^2 t^2$$

and

$$\begin{aligned} \frac{\sin(\pi(d+1)t)}{d \sin(\pi t)} - \frac{1}{d} &\geq \frac{\pi(d+1)t - \frac{1}{6} \pi^3 (d+1)^3 t^3}{\pi dt} - \frac{1}{d} \\ &= 1 - \frac{1}{6} \pi^2 \frac{(d+1)^3}{d} t^2 \end{aligned}$$

are valid for $t < \frac{1}{d}$, because the remainder in denominator and numerator are alternating series with decaying magnitude of terms. These estimates give in either case of even or odd d that

$$\|f(t) - f(1)\| \leq \frac{\pi(d+1)^{3/2}}{\sqrt{3d}} (1-t)$$

and therefore

$$\|f'(1)\| \leq \frac{\pi(d+1)^{3/2}}{\sqrt{3d}}.$$

Now applying the preceding theorem gives the claimed upper bound. \square

REFERENCES

- [AG93] N. I. Akhiezer and I. M. Glazman, *Theory of linear operators in Hilbert space*, Dover Publications Inc., New York, 1993, Translated from the Russian and with a preface by Merlynd Nestell, Reprint of the 1961 and 1963 translations, Two volumes bound as one.
- [BH01] H. Bölcskei and F. Hlawatsch, *Noise reduction in oversampled filter banks using predictive quantization*, IEEE Trans. Inform. Theory **47** (2001), no. 1, 155–172.
- [BLO03] B. Beferull-Lozano and A. Ortega, *Efficient quantization for overcomplete expansions in R^N* , IEEE Trans. Inform. Theory **49** (2003), 129–150.
- [BP] B. G. Bodmann and V. I. Paulsen, *Frame paths and error bounds for sigma-delta quantization*, Appl. Comput. Harmon. Anal., to appear (2006).
- [BPY06] J. J. Benedetto, A. M. Powell, and Ö. Yilmaz, *Second-order sigma-delta ($\Sigma\Delta$) quantization of finite frame expansions*, Appl. Comput. Harmon. Anal. **20** (2006), 126–148.
- [BYP04] J.J. Benedetto, O. Yilmaz, and A.M. Powell, *Sigma-delta quantization and finite frames*, Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004. (ICASSP '04), vol. 3, 2004, pp. iii – 937–940.
- [Cas00] P. G. Casazza, *The art of frame theory*, Taiwanese J. Math. **4** (2000), no. 2, 129–201.
- [DD03] I. Daubechies and R. DeVore, *Approximating a bandlimited function using very coarsely quantized data: a family of stable sigma-delta modulators of arbitrary order*, Ann. Math. **158** (2003), 679–710.
- [GKK01] V. K. Goyal, J. Kovačević, and J. A. Kelner, *Quantized frame expansions with erasures*, Appl. Comp. Harm. Anal. **10** (2001), 203–233.
- [GT04] C. S. Güntürk and N. T. Thao, *Refined error analysis in second-order $\Sigma\Delta$ modulation with constant inputs*, IEEE Trans. Inform. Theory **50** (2004), no. 5, 839–860.
- [Gün04] C. S. Güntürk, *Approximating a bandlimited function using very coarsely quantized data: improved error estimates in sigma-delta modulation*, J. Amer. Math. Soc. **17** (2004), no. 1, 229–242 (electronic).
- [GVT98] V. K. Goyal, M. Vetterli, and N. T. Thao, *Quantized overcomplete expansions in R^N : Analysis, synthesis and algorithms*, IEEE Trans. Inf. Theory **44** (1998), 16–31.
- [HKB92] N. He, F. Kuhlmann, and A. Buzo, *Multiloop sigma-delta quantization.*, IEEE Transactions on Information Theory **38** (1992), no. 3, 1015–1028.
- [Nai43] M. A. Naimark, *On a representation of additive operator functions of sets*, Doklady Akad. Nauk SSSR **41** (1943), 373–375.
- [NST97] S. R. Norsworthy, R. Schreier, and G. C. Temes (eds.), *Delta-sigma data converters*, IEEE Press, 1997.
- [OS89] A. V. Oppenheim and R. W. Schaffer, *Discrete-time signal processing*, Prentice-Hall, Englewood Cliffs, NJ, 1989.
- [PM96] J. G. Proakis and D. G. Manolakis, *Digital signal processing*, Prentice-Hall, Englewood Cliffs, NJ, 1996.
- [Tha02] N. T. Thao, *MSE behavior and centroid function of m -th order asymptotic $\Sigma\Delta$ modulators*, IEEE Trans. Circuits Syst., II **49** (2002), 86–100.
- [Yil02] O. Yilmaz, *Stability analysis for several sigma-delta methods of coarse quantization of bandlimited functions*, Constructive Approximation **18** (2002), 599–623.

BERNHARD G. BODMANN: DEPARTMENT OF APPLIED MATHEMATICS, UNIVERSITY OF WATERLOO, WATERLOO, ONTARIO, N2L 3G1

E-mail address: bgb@math.uwaterloo.edu, URL: <http://www.math.uwaterloo.ca/~bgb/>

VERN I. PAULSEN: DEPARTMENT OF MATHEMATICS, UNIVERSITY OF HOUSTON, 4800 CALHOUN ROAD, HOUSTON, TX 77204-3008 U.S.A.

E-mail address: vern@math.uh.edu, URL: <http://www.math.uh.edu/~vern/>

SOHA A. ABDULBAKI: DEPARTMENT OF MATHEMATICS, UNIVERSITY OF HOUSTON, 4800 CALHOUN ROAD, HOUSTON, TX 77204-3008 U.S.A.

E-mail address: sohabaki@math.uh.edu