SIGMA-DELTA QUANTIZATION AND FINITE FRAMES

John J. Benedetto, Özgür Yılmaz

University of Maryland
Department of Mathematics
College Park, MD 20742

Alexander M. Powell

Princeton University
Program in Applied and Computational Mathematics
Fine Hall, Washington Road
Princeton, NJ 08544

ABSTRACT

It is shown that Sigma-Delta ($\Sigma\Delta$) algorithms can be used effectively to quantize finite frame expansions for $\mathbf{R}^d$. Error estimates for various quantized frame expansions are derived, and in particular, it is shown that $\Sigma\Delta$ quantizers outperform the standard PCM schemes.

1. INTRODUCTION

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

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

where x_n are real or complex numbers. Such an expansion is said to be *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. Therefore, a second step is needed to reduce the continuous range of this sequence to a discrete, preferably finite, set. This second step is called *quantization*.

2. FRAME THEORETIC BACKGROUND

In various applications it is convenient to assume that the signals of interest are elements of a Hilbert space, e.g., band-limited functions, $L^2(\mathbf{R}^d)$, or $\mathbf{R}^d$. In this case, one can consider more structured dictionaries, such as frames.

Definition 1 *A collection $F = \{e_n\}_{n\in\Lambda}$ in a Hilbert space H is a* frame *if*

$$\forall x \in H, \quad A\|x\|^2 \leq \sum |\langle x, e_n\rangle|^2 \leq B\|x\|^2,$$

where the frame bounds $0 < A \leq B < \infty$ are fixed constants.

The frame is *tight* if $A = B$. An important remark is that the frame bound A of a *uniform* or *normalized* tight frame, i.e., a tight frame with $\|e_n\| = 1$ for all n, "measures" the redundancy of the system. If $A = 1$ then a uniform tight frame $\{e_n\}$ is an orthonormal basis and there is no redundancy. The larger the frame bound $A > 1$ is, the more redundant a uniform tight frame is.

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

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

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

$$AI \leq S \leq BI,$$

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

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

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

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

$$x = \sum_{n\in\Lambda} \langle x, e_n\rangle (S^{-1}e_n) \tag{2}$$

$$= \sum_{n\in\Lambda} \langle x, (S^{-1}e_n)\rangle e_n \tag{3}$$

The atomic decompositions in (2) and (3) are the first step towards a digital representation. If the frame is tight

with frame bound A, then both frame expansions are equivalent and we have

$$\forall x \in H, \quad x = A^{-1} \sum_{n \in \Lambda} \langle x, e_n \rangle e_n. \tag{4}$$

For the important case of *finite uniform frames* for $\mathbf{R}^d$ and $\mathbf{C}^d$, the frame constant A is N/d, where N is the cardinality of the frame [1], [2], [3], [4].

3. QUANTIZATION

In this section we shall discuss the quantization of tight frame expansions, (4). An intuitive quantization technique is the $2\lceil 1/\delta \rceil$-level PCM quantizer with step size δ, given by replacing $x_n = \langle x, e_n \rangle$ with $q_n = \delta(\lceil x_n/\delta \rceil - 1/2)$. One can show that if $|x_n| < 1$ for all n then $\sup |x_n - q_n| \leq \delta/2$ for all n.

If $\{e_n\}_{n=1}^N$ is a uniform tight frame for $\mathbf{R}^d$, and $\| \cdot \|$ denotes the Euclidean norm in $\mathbf{R}^d$, then the approximation error satisfies

$$\|x - A^{-1} \sum_{n=1}^{N} q_n e_n\| \leq \left(\frac{d}{2}\right) \delta, \tag{5}$$

where $A = N/d$. This error estimate does not utilize the redundancy of the frame. (5) can be improved by making the assumption that the quantization error sequence $\{\eta_n\} = \{x_n - q_n\}$ is a signal independent sequence of $i.i.d.$ random variables with mean 0 and variance $\delta^2/12$. This is Bennett's *white noise assumption* [5]. Here, the sequence $\{\eta_n\}$ is randomized by assuming that it is computed for a random signal $x \in H$ with a smooth probability distribution. In this case, one can show that the mean square (approximation) error (MSE) satisfies

$$MSE = E\|x - A^{-1} \sum_{n=1}^{N} q_n e_n\|^2 \leq \frac{d\delta^2}{12A}, \tag{6}$$

where $A = N/d$, and E is the expectation with respect to the associated probability distribution, cf., [3]. Note that (6) is unsatisfactory for the following reasons:

(a) The white noise assumption does not hold in some elementary settings. For example, consider the tight frame $\{e_n = (\cos(n(2\pi/N)), \sin(n(2\pi/N))\}_{n=1}^N$ for $\mathbf{R}^2$ with even N. Clearly, $e_n = -e_{n+N/2}$ for any n, and this violates Bennett's assumption. Thus, the predicted MSE will not be attained in this case.

(b) The MSE bound (6) only gives information about the average quantizer performance.

(c) As one increases the redundancy of the expansion, i.e., as the frame bound A increases, the MSE given in (6) decreases only as $1/A$, i.e., the redundancy of the expansion is not utilized very efficiently.

Sigma-Delta ($\Sigma\Delta$) quantizers are widely implemented to quantize oversampled bandlimited functions [6, 7]. When used to quantize oversampled bandlimited functions, first-order 1-bit $\Sigma\Delta$ quantizers yield approximations where the pointwise approximation error is bounded by $C_1 A^{-1}$ [7] or better [8], and the MSE behaves like A^{-3} [9], where A is the frame bound of the corresponding tight frame for the space of bandlimited functions.

4. FIRST-ORDER $\Sigma\Delta$ QUANTIZERS

In this section, we introduce the standard first order $\Sigma\Delta$ scheme with the aim of using it to quantize finite frame expansions in $\mathbf{R}^d$.

Given the *midrise* quantization alphabet $\mathcal{A}_K^\delta := \{(-K+1/2)\delta, (-K+3/2)\delta, \cdots, (-1/2)\delta, (1/2)\delta, \cdots, (K-1/2)\delta\}$, we define

$$Q(u) := \arg\min_{q \in \mathcal{A}_K^\delta} |u - q| \tag{7}$$

For simplicity, we only consider midrise quantizers, although our results are also valid more generally, e.g., for *midtread* quantization alphabets.

Definition 4 *Given a sequence of frame coefficients $\{x_n\}_{n=1}^N$, a first-order $\Sigma\Delta$ quantizer produces the quantized sequence $\{q_n\}$ by running the iteration*

$$u_n = u_{n-1} + x_n - q_n, \quad q_n = Q(u_{n-1} + x_n), \tag{8}$$

where $\{u_n\}$ is an auxiliary sequence of state variables, and Q is the $2K$-level midrise uniform scalar quantizer defined by (7).

We say that a first-order $\Sigma\Delta$ quantizer is a *$2K$-level first-order $\Sigma\Delta$ quantizer with step size δ* if it is defined by means of (8), where Q is as in (7).

The following proposition asserts that the first-order $\Sigma\Delta$ quantizer is *stable*.

Proposition 5 *Let K be a positive integer, let $\delta > 0$, and consider the $\Sigma\Delta$ system defined by (8) and (7). If $|x_n| \leq (K - 1/2)\delta$ for all n and $|u_0| \leq \delta/2$, then $|u_n| \leq \delta/2$ for all n.*

5. $\Sigma\Delta$ QUANTIZATION OF FRAMES FOR $\mathbf{R}^2$

Theorem 6 *Let K be a positive integer, let $\delta > 0$, and consider $\{e_n = (e_n^1, e_n^2)\}_{n=1}^N$ a uniform tight frame for $\mathbf{R}^2$, ordered so that $\arctan(e_k^2/e_k^1) \leq \arctan(e_l^2/e_l^1)$ if $k \leq l$. Let $x \in \mathbf{R}^2$ satisfy $\|x\| \leq (K - 1/2)\delta$, and suppose $\{q_n\}_{n=1}^N$ is produced by a $2K$-level first-order $\Sigma\Delta$ quantizer with step size δ using the frame coefficients $\{x_n = \langle x, e_n \rangle\}$ as the input. Then, the approximation $\widetilde{x} := 2N^{-1} \sum_{n=1}^N q_n e_n$ satisfies*

$$\|x - \widetilde{x}\| \leq N^{-1}(2\pi + 2)\delta. \tag{9}$$

Corollary 7 *Given the hypotheses of Theorem 6, the estimate in (9) can be replaced by*

(i) $\|x-\widetilde{x}\| \leq N^{-1}(2\pi+1)\delta$ *if we choose* $u_0 = 0$ *in (8), and by*

(ii) $\|x-\widetilde{x}\| \leq 2\pi N^{-1}\delta$ *when* N *is even if we have a midrise quantizer, if we choose* $u_0 = 0$*, and if* $\sum_{n=1}^{N} e_n = 0$*.*

The above results can be generalized to the case of uniform non-tight frames. Let $\{e_n\}_{n=1}^N$ be a uniform frame for $\mathbf{R}^2$ with frame bounds A and B, and with frame operator S. Let $x_n := \langle x, e_n\rangle$ and suppose q_n is obtained by quantizing x_n using a $2K$-level first-order $\Sigma\Delta$ quantizer with step size δ. Define $\widetilde{x} := \sum_{n=1}^{N} q_n(S^{-1}e_n)$. Note that $B^{-1} \leq \|S^{-1}\| \leq A^{-1}$.

Proposition 8 *With the setup of the previous paragraph, the pointwise approximation error* $\|x-\widetilde{x}\|$ *satisfies the inequality*

$$\|x-\widetilde{x}\| \leq \|S^{-1}\|(2\pi+2)\frac{\delta}{2}. \tag{10}$$

Moreover, we can replace $(2\pi+2)$ *in (10) with* $(2\pi+1)$ *and* 2π *if the conditions listed in Corollary 7 (i) and (ii), respectively, are satisfied.*

A simple comparison of the error bounds obtained in this section with the MSE error bounds for PCM quantizers, which were discussed in Section 3, shows that the MSE corresponding to first-order $\Sigma\Delta$ quantizers is smaller than the MSE corresponding to PCM quantizers for uniform tight frames of $\mathbf{R}^2$ with redundancy A if

- $A > 1.5(2\pi+1)^2 \approx 80$ for any uniform tight frame for $\mathbf{R}^2$, as long as the frame elements are ordered as described in Theorem 6 with the additional condition that u_0 is chosen to be 0, or
- $A > 1.5(2\pi)^2 \approx 59$ if the uniform tight frame for $\mathbf{R}^2$ is as described in Corollary 7 (ii).

Numerical experiments indicate that smaller redundandcy than above may still be sufficient for first order $\Sigma\Delta$ quantization to outperform PCM. Figure 1 shows the MSE achieved by $2K$-level PCM quantizers and $2K$-level first-order $\Sigma\Delta$ quantizers with step size $\delta = 1/K$ for several values of K for uniform tight frames for $\mathbf{R}^2$ obtained by the Nth roots of unity. The plots suggest that if the frame bound is larger than approximately 10, the first-order $\Sigma\Delta$ quantizer outperforms PCM.

Finally, we want to note that the upper-bound on the MSE for first-order $\Sigma\Delta$ quantizers is the asymptotic lower bound for the MSE for PCM quantizers with step size δ, given in [3].

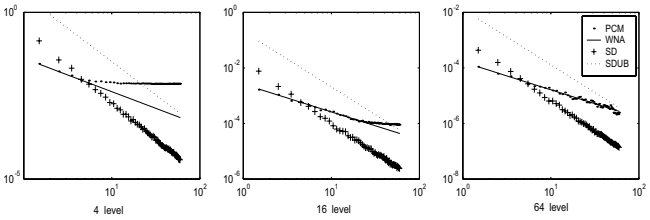
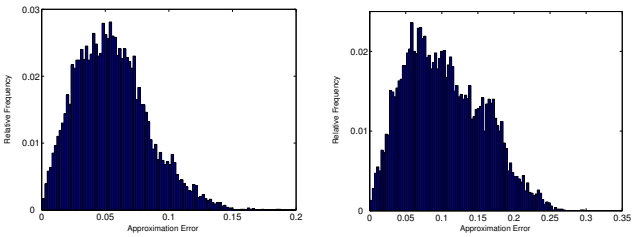


Fig. 1. Comparison of the MSE for $2K$-level PCM quantizers and $2K$-level first-order $\Sigma\Delta$ quantizers with step size $\delta = 1/K$. The figures plot MSE versus the frame bound, A. Frame expansions of 100 randomly selected points in $\mathbf{R}^2$ for frames obtained by the Nth roots of unity were quantized. In the figure legend PCM and SD correspond to the MSE for PCM and the MSE for first-order $\Sigma\Delta$ obtained experimentally, respectively. In the legend, the bound on the MSE for PCM, computed with white noise assumption, is denoted by WNA. Finally, SDWN in the legend stands for the MSE bound for $\Sigma\Delta$ that we would obtain if the approximation error was uniformly distributed between 0 and the upper bound of Corollary 7(i).



(a) natural order (b) reordered

Fig. 2. The histograms of Example 9.

6. QUANTIZATION OF UNIFORM FRAME EXPANSIONS IN HIGHER DIMENSIONS

In this section we show how to generalize the two dimensional results of the previous section to higher dimensions. We start by noting that $\Sigma\Delta$ schemes are defined in an iterative manner, see (8). Therefore, given a frame $\{e_n\}_{n=1}^N$ and $x \in \mathbf{R}^2$, the resulting quantization of the frame coefficients of x, as well as the approximation error bounds in Section 5 depend heavily on the order in which the frame coefficients are quantized. In Theorem 6, we imposed a natural order on the frame coefficients to obtain the estimate given by (9). Changing this order has a drastic affect on the approximation error.

Example 9 *Consider the uniform tight frame for* $\mathbf{R}^2$ *given by* $\{e_n\}_{n=1}^7$*, where* $e_n := (\cos(n2\pi/7), \sin(n2\pi/7))$*. We randomly choose 10,000 points in the unit ball of* $\mathbf{R}^2$*. First, we quantize the frame coefficients of each point using (8) in their natural order, by setting* $x_n = \langle x, e_n\rangle$ *in (8). Fig-*

ure 2 (a) shows the histogram of the corresponding approximation errors. Next, we quantize the frame coefficients of the same 10,000 points, only this time after reordering the frame coefficients as $x_1, x_4, x_7, x_3, x_6, x_2, x_5$. Figure 2 (b) shows the histogram of the corresponding approximation errors in this case. Clearly, the average approximation error for the new ordering is significantly larger than the average approximation error associated with the original ordering. This example and the earlier discussion motivate the following notation.

Definition 10 *Let $F = \{e_n\}_{n=1}^N$ be a finite frame for $\mathbf{R}^d$, and let p be a permutation of $\{1, 2, \ldots, N\}$. We define the (first-order)* variation *of the frame F* with respect to *p as*

$$\sigma(F, p) := \sum_{n=1}^{N-1} \|e_{p(n)} - e_{p(n+1)}\|. \tag{11}$$

Now, we can restate Theorem 6 in a more general way.

Theorem 11 *Let $F = \{e_n\}_{n=1}^N$ be a uniform tight frame for $\mathbf{R}^d$, and let p be as above. Suppose that K is a positive integer and $\delta > 0$. Consider q_n that are produced by the first order $\Sigma\Delta$ quantizer defined by (8) and (7) using the sequence $\{\langle x, e_{p(n)}\rangle\}_{n=1}^N$ as the input. Then the approximation $\widetilde{x} := dN^{-1}\sum_{n=1}^N q_n e_{p(n)} = dN^{-1}\sum_{n=1}^N q_{p^{-1}(n)} e_n$ satisfies*

$$\|x - \widetilde{x}\| \leq \frac{d}{N}(\sigma(F, p) + 2)\frac{\delta}{2}. \tag{12}$$

Theorem 11 shows that the performance of the first-order $\Sigma\Delta$ algorithm in quantizing a given frame expansion in $\mathbf{R}^d$ depends on the variation of the frame with respect to the order in which the coefficients are quantized.

The *harmonic frames* for $\mathbf{R}^d$ [4] provide an infinite family of uniform tight frames with arbitrarily high redundancy for which we can derive uniform bounds on the frame variation.

Theorem 12 *Let F_N be a harmonic frame for $\mathbf{R}^d$ and let p be the identity map on $\{1, 2, \ldots, N\}$. Then $\sigma(F_N, p) \leq \pi d(d+1)$.*

Theorems 11 and 12 show that first-order $\Sigma\Delta$ schemes have approximation error which is $O(N^{-1})$ as $N \to \infty$ when used to quantize harmonic frame expansions for $\mathbf{R}^d$. N is the number of frame elements. Thus the MSE corresponding to the first-order $\Sigma\Delta$ quantization of harmonic frames for $\mathbf{R}^d$ behaves like N^{-2}. This is the theoretical asymptotic lower bound for the MSE for PCM given in [3]. Furthermore, the upper bound on the MSE for the first-order $\Sigma\Delta$ quantization performs better than the MSE for PCM computed under Bennett's white noise assumption if N is sufficiently large.

See [10] for further results on finite frame $\Sigma\Delta$ quantization, as well as proofs of the results presented in this paper.

7. REFERENCES

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

[2] V.K. Goyal, J. Kovačević, and J.A. Kelner, "Quantized frame expansions with erasures," *Appl. Comput. Harmon. Anal.*, vol. 10, pp. 203–233, 2001.

[3] V.K. Goyal, M. Vetterli, and N.T. Thao, "Quantized overcomplete expansions in R^n: Analysis, synthesis, and algorithms," *IEEE Transactions on Information Theory*, vol. 44, no. 1, pp. 16–31, January 1998.

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

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

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

[7] I. Daubechies and R. DeVore, "Reconstructing a bandlimited function from very coarsely quantized data: A family of stable sigma-delta modulators of arbitrary order," *Annals of Mathematics*, to appear.

[8] C.S. Güntürk, "Approximating a bandlimited function using very coarsely quantized data: Improved error estimates in sigma-delta modulation," *J. Amer. Math. Soc.*, to appear.

[9] W. Chen and B. Han, "Improving the accuracy estimate for the first order sigma-delta modulator," *J. Amer. Math. Soc.*, submitted in 2003.

[10] J.J. Benedetto, A.M. Powell, and Ö. Yılmaz, "Finite frame Sigma-Delta quantization," *Preprint*, 2003.