Finite Frame Quantization

Liam Fowl

University of Maryland

August 21, 2018

Overview

Motivation

- Quantization has long been an area of study in electronic signal processing (going back to Bennett in 1949 [Bennett 49])
- Digital signal processing made possible the storage/transmission of audio and video signaling [Daubechies et al. 03].
- On the mathematical side, much work has been done by the likes of Daubechies and others on understanding analog audio signals [Daubechies et al. 03].
- More recently, much work has been done in the more general setting of all digital signals, where the structure of audio signals is absent [Bodmann et al. 07], [Benedetto et al. 06].

- The goal of such work is to find a good representation of the signal for storage, transmission, and recovery
- A common starting point for finding such a representation is to expand the signal x over a countable dictionary $\{e_n\}$ so that $x = \sum c_n e_n$
  - What's the best way to expand?
  - How to we deal with the coefficients, c_n?

To deal with the first question, we will introduce frames. The theory of frames was first introduced by Duffin and Schaeffer in 1952, and first applied to signal processing by Daubechies, Grossman, and Meyer in 1986. Frames are a specific example of a dictionary which have many benefits:

- in some settings, they are robust under additive noise [Benedetto et al. (3) 01].
- in the setting of finite frames, representations have been shown to be robust under partial data loss [Goyal et al. (2) 01]

Background

A collection $F = \{e_n\}_{n\in\Lambda}$ in a Hilbert space $\mathcal{H}$ is said to be a frame for $\mathcal{H}$ if there exists $0 < A \leq B < \infty$ so that for each $x \in \mathcal{H}$ we have:

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

Example

Roots of unity frame for $\mathbb{R}^2$: It is not hard to show that the N^{th} roots of unity defined by $e_n = [\cos(2\pi n/N), \sin(2\pi n/N)]^T$ form a normalized tight frame.

F is called *tight* if $A = B$, and *normalized* (or uniform) if $||e_n|| = 1\ \forall n$. From now on, we'll assume a finite frame (Λ is a finite collection).

Given a frame, we can define the frame operator, $S : \mathcal{H} \to \mathcal{H}$ is defined as $S(x) = \sum_{n=1}^{N} \langle x, e_n \rangle e_n$.
One can show S is invertible, $\{S^{-1} e_n\}$ is also a frame for $\mathcal{H}$, and the following decompositions hold:

$$\forall x \in \mathbb{R}^d, \quad x = \sum_n \langle x, \tilde{e_n} \rangle e_n = \sum_n \langle x, e_n \rangle \tilde{e_n}$$

where $\tilde{e_n} = S^{-1} e_n$. In general (not always), there are other frames that one can use to reconstruct. Any frame $\{f_n\}_{n=1}^{N}$ for $\mathbb{R}^d$ which satisfies:

$$x = \sum_{n=1}^{N} \langle x, e_n \rangle f_n$$

is called a *dual frame* to $\{e_n\}$. In the case of a normalized, tight frame for $\mathbb{R}^d$, the *canonical dual frame* $\{S^{-1} e_n\}$ will just be $\{\frac{d}{N} e_n\}$.

PCM Quantization

Perhaps the most natural quantization scheme is called PCM (Pulse Code Modulation). Let $\{e_n\}$ be a normalized tight frame for $\mathbb{R}^d$. We then get that for each $x \in \mathbb{R}^d$

$$x = \frac{d}{N} \sum_{n=1}^{N} x_n e_n$$

where $x_n = \langle x, e_n \rangle$. The $2K, \delta$ PCM quantization scheme takes an alphabet:

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

and a quantization function $Q(u) = argmin_{q \in \mathcal{A}_K^\delta} |u - q|$ to construct an approximation

$$\tilde{x} = \frac{d}{N} \sum_{n=1}^{N} Q(x_n) e_n$$

PCM/Bennett's white noise assumption

Question: What error estimates can we make about PCM?

One way we can utilize the redundancy of our frame is to make Bennett's white noise assumption:

Bennett's white noise assumption

The error sequence $\{x_n - q_n\}$ is well approximated by a sequence of independent, identically distributed (uniform on $[\frac{-\delta}{2}, \frac{\delta}{2}]$) random variables.

If we make this assumption, for a normalized, tight frame, we attain the following MSE estimate for reconstruction with the canonical dual:

$$MSE_{PCM} \leq E\left(||x - \tilde{x}||^2\right) = \frac{d^2\delta^2}{12N}$$

Bennett's white noise assumption

While Bennett's white noise assumption is sometimes justified, there are some shortcomings.

- The assumption only gives us an estimate on the *average* quantizer performance.
- There are cases where the assumptions are not rigorous.

Example Consider the normalized tight frame for $\mathbb{R}^2$ given by:

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

In addition to the shortcomings listed above, it is also known that PCM has poor robustness properties in other settings [Daubechies et al. 03]. This motivates an alternative quantization scheme which utilizes frame redundancy better.

$\Sigma\Delta$ quantization

- $\Sigma\Delta$ (SD) quantization has its roots in electrical engineering.
- SD quantization was first introduced by Yasuda in the 1960s [Yasuda et al. 62] as a way of improving classical Δ-modulation, a technique for AD conversion.
- The added Σ reflects the sum tracking feature of the algorithm. The basic idea is to include tracking of current, and past quantization differences.
- Some of the earliest work mathematically on SD quantization was done by Daubechies analyzing bandlimited functions.

Now, for the algorithm's construction: As before, take the same alphabet $\mathcal{A}_K^\delta$ and $Q(u)$. The first order SD quantization scheme is defined by the iteration:

$$u_n = u_{n-1} + x_{p(n)} - q_n$$

$$q_n = Q(u_{n-1} + x_{p(n)})$$

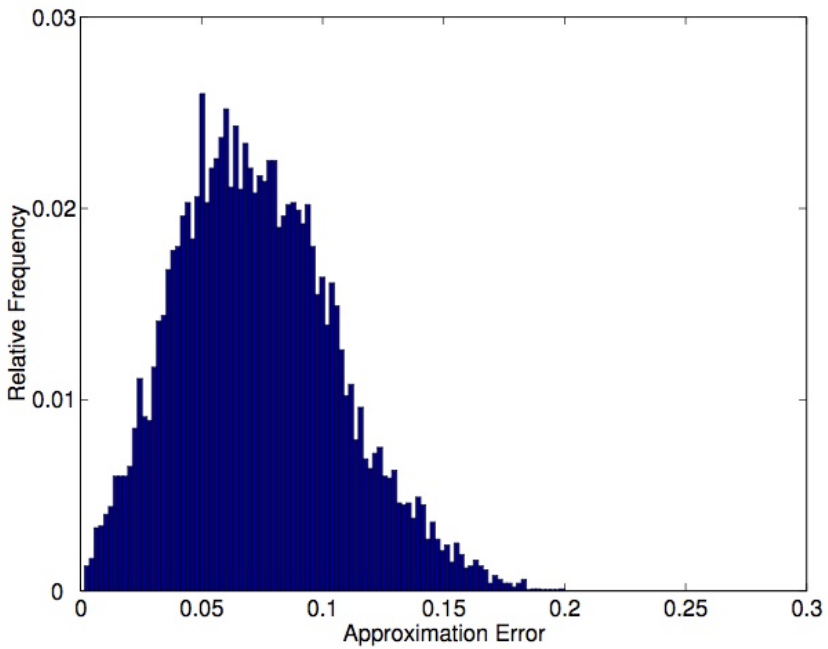
where u_0 is some constant (usually 0) and $p \in S_N$.

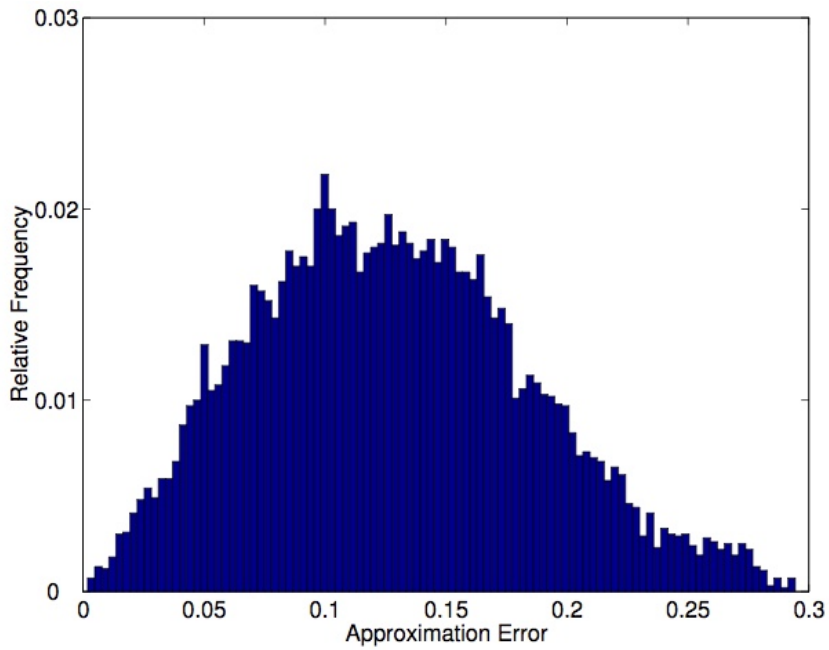
Proposition

Let K be a positive integer, $\delta > 0$ and consider the SD scheme defined above. If $u_0 \leq \frac{\delta}{2}$ and $\forall n$ we have $|x_n| \leq (K - \frac{1}{2})\delta$, then $\forall n$, $|u_n| \leq \frac{\delta}{2}$

An aspect of the algorithm of note is the inclusion of the permutation. Because of the inclusion of the internal state variable, ordering does affect the algorithm's performance.

Example: Let F be the 7^{th} roots of unity in $\mathbb{R}^2$, and consider $\mathcal{A}_4^{\frac{1}{4}}$. The approximation errors for $10,000$ randomly selected points is shown below [Benedetto et al. 06]:





SD error estimates

Mathematical error estimates [Benedetto et al. 06]:

Theorem

Let $F = \{e_n\}$ be a finite normalized frame for $\mathbb{R}^d$, and $p \in S_N$. Take $|u_0| < \delta/2$ and let $x \in \mathbb{R}^d$ satisfy $||x|| \leq (K - 1/2)\delta$. The approximation error of the SD scheme $||x - \tilde{x}||$ satisfies

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

where the frame variation, $\sigma(F, p)$ is defined as
$\sigma(F, p) = \sum_{n=1}^{N-1} ||e_{p(n)} - e_{p(n+1)}||$

In the case of tight frames, this estimate becomes:

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

Families of frames

From the previous error estimate, there are a few ways we can lower the approximation error:

- Increase the resolution of quantization
- Utilize redundant frames

The first option can become costly, but to adopt the second approach, it is important to find families of frames with uniformly bounded frame variation.

Example: (Roots of unity). For $\mathbb{R}^2$, take R_N to be the frame consisting of N^{th} roots of unity. Then $\{R_N\}$ is a family of normalized tight frames with uniform bound

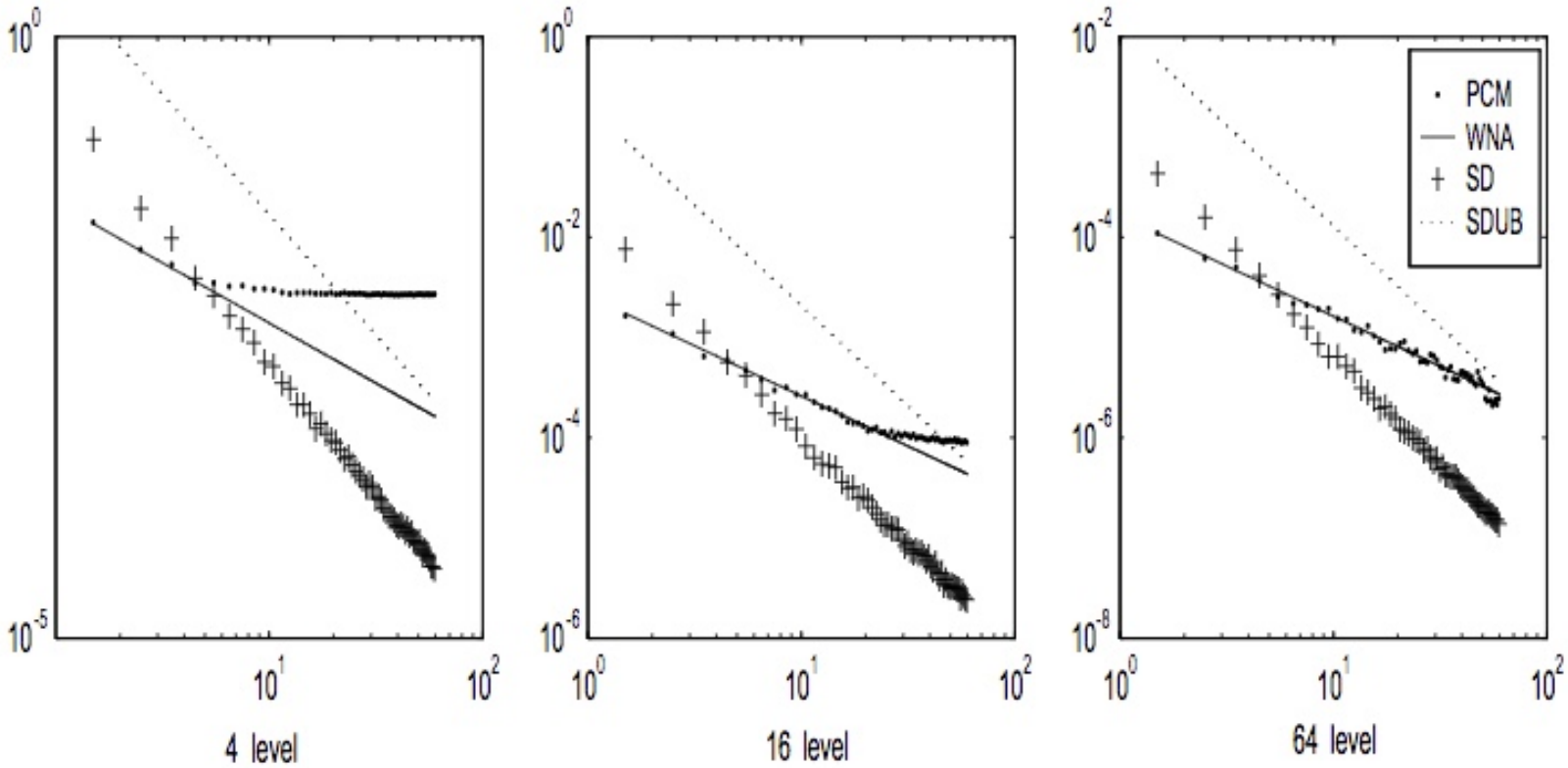
$$\sigma(R_N, id) \leq 2\pi$$

Comparison with PCM

We recall that even under Bennett's white noise assumption, $MSE_{PCM} = \frac{d^2\delta^2}{12N}$. We have as a direct consequence of our previous error estimates that:

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

PCM vs. SD quantization MSE of 100 randomly selected points with $2K$-level quantization. The N^{th} roots of unity frame was chosen [Benedetto et al. 06].



Higher order SD quantization

For first order SD quantization, the auxilary sequence $\{u_n\}$ was introduced. A natural question arises: can we improve performance by adding more internal state variables to better track differences?

r^{th} order SD quantization

$$q_n = Q(F(u_{n-1}^1, u_{n-1}^2, \ldots, u_{n-1}^r, x_n))$$
$$u_n^1 = u_{n-1}^1 + x_n - q_n,$$
$$u_n^2 = u_{n-1}^2 + u_n^1,$$
$$\vdots$$
$$u_n^r = u_{n-1}^r + u_n^{r-1}$$

Where $u_0^1 = u_0^2 = \ldots = u_0^r = 0$, $n = 1, \ldots, N$, and $F : \mathbb{R}^{r+1} \to \mathbb{R}$ is a fixed function.

As before, a natural first question is: what about stability?

Definition

We will say that the r^{th} order scheme as defined before is stable if there are constants $C_1, C_2 > 0$ so that for any $N > 0$ and any $x = \left[x_n\right]_{n=1}^{N} \subset \mathbb{R}^n$ we have that if $\forall n,\ |x_n| < C_1$, then $\forall n, \forall j\ |u_n^j| < C_2$

As a basic example, it has been shown that the following 1-bit second order SD scheme is stable [Benedetto et al. (2) 06]:

$$q_n = sign(u_{n-1}^1 + \frac{1}{2}u_{n-1}^2),$$
$$u_n^1 = u_{n-1}^1 + x_n - q_n,$$
$$u_n^2 = u_{n-1}^2 + u_n^1$$

What about higher order multibit schemes?

Example

The general r^{th} order SD scheme with

$$q_n = Q\left(u^r_{n-1} + u^{r-1}_{n-1} + \ldots + u^1_{n-1} + x_n\right)$$

is stable in the sense that $|u^j_{n-1}| \leq 2^{r-j}\delta \Rightarrow |u^j_n| \leq 2^{r-j}\delta$ provided that $|x_n| < 1 - (2^r - 1)\delta$

Note that we need δ to be fairly small in this formulation as r increases.

Now what about error estimation? The following lemma [Lammers et al.10] will be helpful for further error bounds.

Lemma

Consider a stable r^{th} order SD scheme with stability constants C_1, C_2. Suppose $||x|| < C_1$ has frame expansion $x = \sum_1^N \langle x, e_n \rangle f_n$. The linear reconstruction $\left(\tilde{x} = \sum_{n=1}^N q_n f_n\right)$ the SD scheme satisfies:

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

Where Δ is the forward difference operator defined as:
$\Delta^0 e_n = e_n,\ \Delta e_n = e_n - e_{n+1},\ \Delta^j e_n = \Delta \cdot \Delta^{j-1} e_n.$

The first term in this sum will be referred to as the main error term, and the second term as the boundary error term.

Alternative dual frames

In the previous lemma, we kept arbitrary the choice of dual frame $\{f_n\}$.

- Later, we will see alternative dual frames are useful for reconstruction
- In the first order case, we reconstructed with the canonical dual frame and achieved reasonable results.
- Can we expect the same for the higher order scheme?

When reconstructing with the canonical dual frame, the higher order SD algorithm does not achieve ideal results. The following lemma [Lammers et al.10] illustrates this.

Lemma

Given a stable rth order SD scheme with $r \geq 3$, and $\{e_n\}_{n=1}^N$ a unit norm tight frame for $\mathbb{R}^d$ that satisfies the zero sum condition: $\sum_{n=1}^N e_n = 0$, and also satisfies: $A/N^j \leq ||\Delta^j e_n|| \leq B/N^j$, then given $||x|| \leq C_1$ in $\mathbb{R}^d$, reconstruction using the canonical dual yields lower bounds:

$$N \text{ odd} \Rightarrow \frac{d\delta}{2N} - \frac{3dC_2B}{N^2} \leq ||x - \tilde{x}||$$

$$N \text{ even} \Rightarrow \frac{dA|u_{N-1}^2|}{N^2} - \frac{2dC_2B}{N^3} \leq ||x - \tilde{x}||$$

It is also shown in [Lammers et al.10] that similar lower bounds exist for stable 1 bit quantization schemes using canonical reconstruction.

The previous lemma illustrates a fundamental problem with canonical reconstruction.

- The boundary error term seems to be too large
- If we were to use alternative dual frames for reconstruction, how can we minimize the boundary error term?
- What asymptotic behavior can we expect?

Motivation for the construction below [Lammers et al.10] can be found in [Bodmann et al. 07]. We will look at dual frames that come from sampling a frame path.

Definition

(Frame path property) [Lammers et al.10] Fix $r > 0$ and let $E = \{E_N\}_{N=d}^{\infty}$ be a collection of unit norm frames of order N for $\mathbb{R}^d$. Suppose there is a family of dual frames $F = \{F_N\}_{N=d}^{\infty}$ for E so that

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

For some real valued functions $\psi_i \in C^r[0,1]$. Also suppose there is some $C_\psi(r)$ independent of N so that such derivatives of ψ_i satisfy:

$$\forall i = 1, \ldots, d \ \forall j = 1, \ldots, r-1, \ ||\psi_i^{(j)}||_{L^\infty[(N-j)/N,1]} \leq \frac{C_\psi(r)}{N^{r-j-1}}$$

We'll also enforce the condition that $\psi_i(1) = 0 \ \forall i$

With this property we can prove the following estimation:

Theorem

Take a stable r^{th} order SD scheme and frames E_N, F_N satisfying the frame path property. Fix $x \in \mathbb{R}^d$ with $||x|| < C_1$, then if $\tilde{x} = \sum_{n=0}^{N-1} q_n(x) f_n$, then

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

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

- The asymptotic behavior is $\sim \frac{1}{N^r}$ which is much better than reconstruction with the canonical dual
- Specifically, the boundary error term in the estimation lemma is minimized better.
- Difficulties in higher order SD quantization arise in the reconstruction phase, not the encoding phase [Lammers et al.10].
- Some work has been done by Bodmann and Paulsen [Bodmann et al. 07] on reconstructing and encoding with the same frame, however this has limited applicability [Lammers et al.10].

Application of this theorem to the roots of unity frames for $\mathbb{R}^2$

- It was shown in a previous lemma that for the roots of unity frame, reconstructing with the canonical dual frame performs poorly.
- For example, how do we construct dual frames for the roots of unity family so that the dual frames satisfy the frame path property?

Example

Let $E_N = \{e_n\}_1^N$ be the N^{th} roots of unity frame as defined before. Define

$$g_n^N = \left[a_0 + \sum_{l=1}^{k} a_l \cos(2\pi(l+1)n/N), \sum_{l=1}^{k} b_l \sin(2\pi(l+1)n/N)\right]^T$$

where $k = k(r)$, $\{a_l\}, \{b_l\}$ are constants that will be further explained later. Then setting $f_n^N = \frac{1}{N}(2e_n^N + g_n^N)$, discrete orthogonality relations get us that $\sum_n \langle x, e_n \rangle g_n = 0$, and so F_N defined by this will form a dual frame to E_N [Lammers et al.10].

It is still left to show that the alternative dual frame defined before has the frame path property. We have that

$$\psi_1(t) = 2\cos(2\pi t) + a_0 + \sum_{l=1}^{k} a_l \cos((l+1)2\pi t)$$

and

$$\psi_2(t) = 2\sin(2\pi t) + \sum_{l=1}^{k} b_l \sin((l+1)2\pi t)$$

so that $f_n^N = \frac{1}{N}[\psi_1(n/N), \psi_2(n/N)]$.

We now need to select the $\{a_l\}, \{b_l\}$ cleverly to bound the derivatives.

We'll start by looking at the power series expansions of ψ_1, ψ_2 around 0. These are given by:

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

Where $\beta_0 = 2 + \sum_{l=0}^{k} a_l$, and

$$\beta_{2n} = \frac{(-1)^n (2\pi)^{2n}}{(2n)!} \left(2 + \sum_{l=1}^{k} a_l (l+1)^{2n}\right)$$

$$\beta_{2n+1} = \frac{(-1)^n (2\pi)^{2n+1}}{(2n+1)!} \left(2 + \sum_{l=1}^{k} b_l (l+1)^{2n+1}\right)$$

The goal now is to eliminate the first $2k$ terms.

To do this, take the Vandermonde matrix

$$V_k = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 2^2 & 3^2 & \cdots & (k+1)^2 \\ \vdots & \vdots & \vdots & \vdots \\ 2^{2k-2} & 3^{2k-2} & \cdots & (k+1)^{2k-2} \end{pmatrix}$$

and the diagonal matrix M_k with $\{2, 3, 4, \ldots, (k+1)\}$ along the diagonal. Then by setting $\vec{a} = [a_1, \ldots, a_k]$ to be the solution to

$$V_k M_k^2 \vec{a} = [-2, \ldots, -2]^T$$

and $\vec{b} = [b_1, \ldots, b_k]$ to be the solution to

$$V_k M_k \vec{b} = [-2, \ldots, -2]^T$$

Then by construction, we have $\beta_n = 0 \ \forall 0 \leq n \leq 2k$.
(Note $a_0 = -2 - (a_1 + \ldots + a_k)$).

Now, all we have to do is take $k = \lceil r/2 \rceil - 1$, and we've constructed a dual frame with the frame path property. The following figure illustrates the new dual frame geometrically [Lammers et al.10].

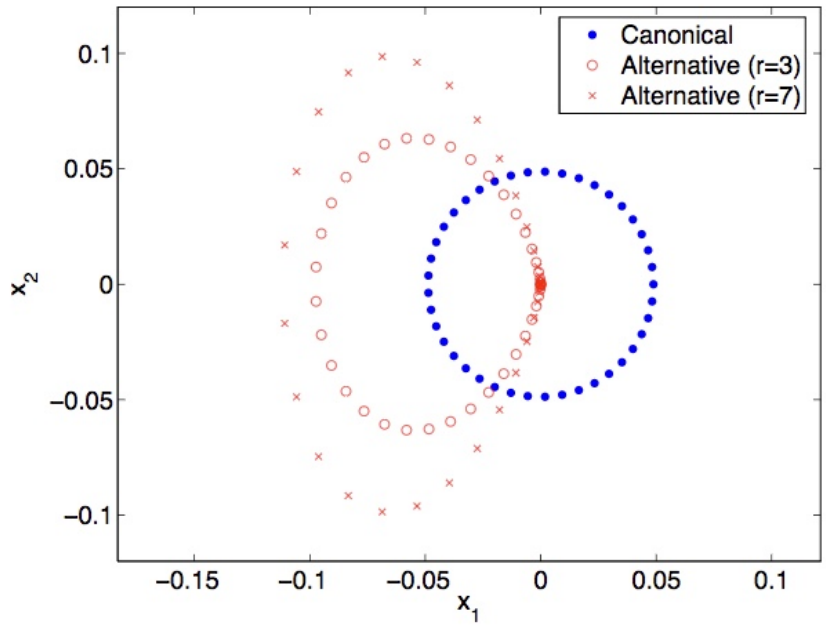


FIGURE 1. The canonical dual frame of E_{41} and two alternative dual frames $F_{41}(3)$ and $F_{41}(7)$.

Below, the boundary error terms are plotted [Lammers et al.10]:

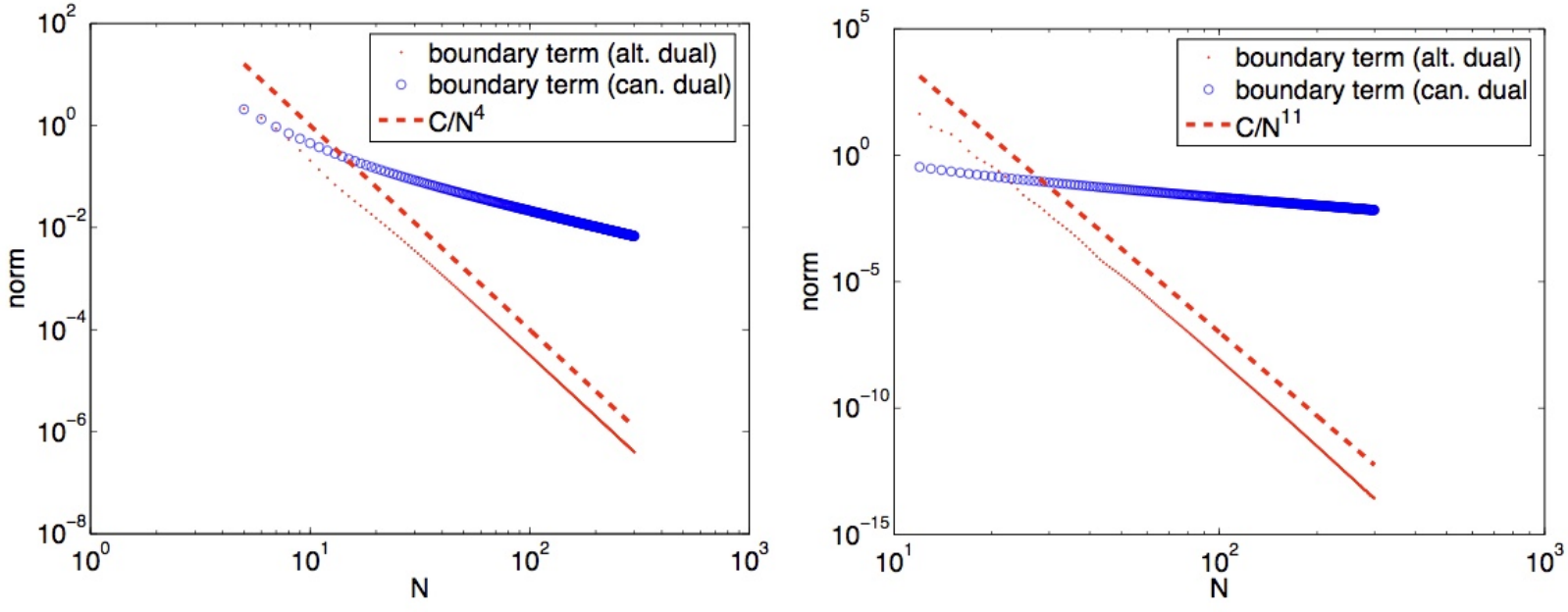


FIGURE 2. Parts (a) and (b) show log-log plots of the norm of the boundary terms in (6.5) for the frames $F_N(3)$ and $F_N(11)$, respectively. For comparison, boundary terms for the canonical dual frame of E_N are also plotted.

And finally, the total error on some test point [Lammers et al.10]:

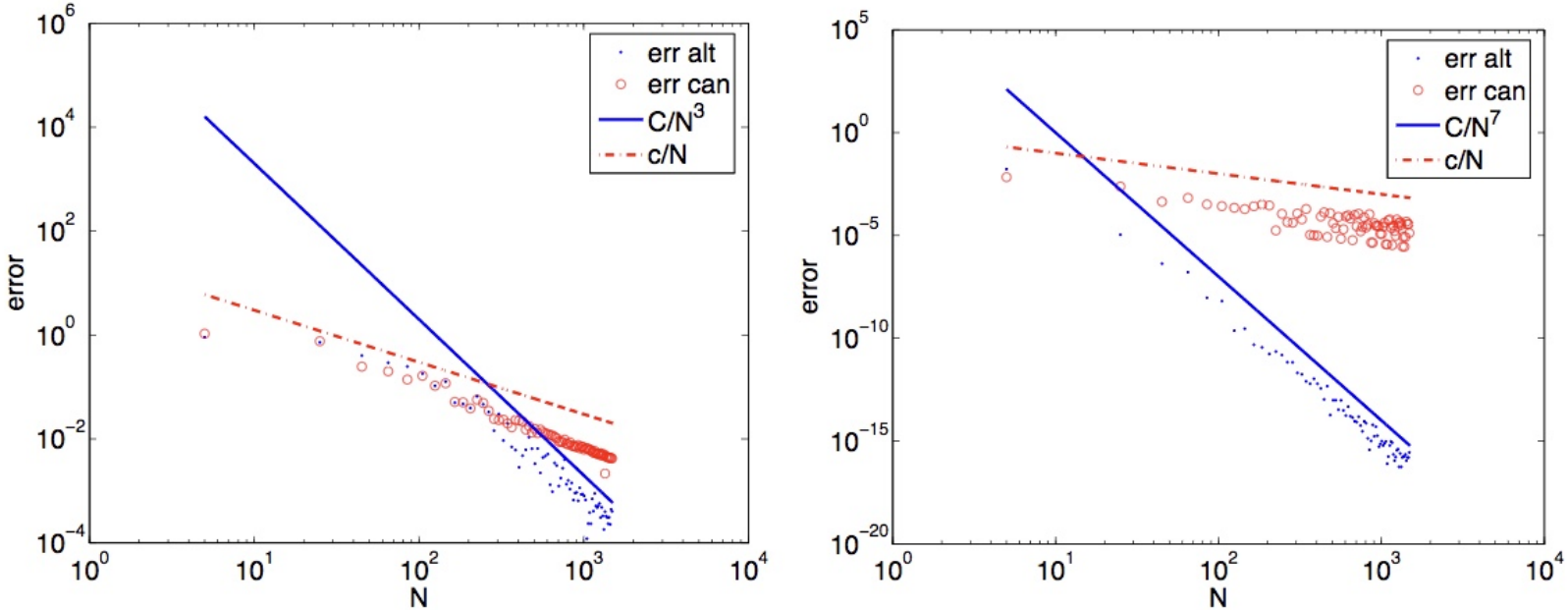


FIGURE 3. The frame expansions of $x = (1/\pi, \sqrt{3/17})$ with respect to E_N are quantized using: (a) the 3rd order scheme of [14], (b) the 7th order $\Sigma\Delta$ scheme from Example 5.2 with $\delta = 0.0039$. Parts (a) and (b) show log-log plots of the approximation error $\|x - \widetilde{x}\|$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

Future Work

On the SD quantization side:

- Are there other dual frame conditions that guarantee $\frac{1}{N^r}$ behavior?
- What about other SD algorithms? Would adding nonlinearity change error estimations?
- Can there be better bounds on the error? Gunturk [Gunturk 03] has shown in the bandlimited case that error can be bounded above by $2^{-.07\lambda}$ for a certain SD construction. Can this somehow be translated to the more general side?

On the machine learning side:

- Training quantized neural networks is an area of interest in machine learning [Goldstein et al. 17].
- Current algorithms achieve convergence (over training iterations) of $\log(N)/N$.
- The real issue for training and deployment is backpropogation and floating point calculations [Goldstein et al. 17].
- It is reasonable to think that 1-bit SD quantization schemes could improve the deployment, and maybe the training of quantized nets.

References I


J. Benedetto, A. Powell, O. Yilmaz,
Sigma-Delta quantization and finite Frames.
IEEE Inform. Theory, 52 (2006), 1990-2005


M. Lammers, A. Powell, O. Yilmaz,
Alternative dual frames for digital-to-analog conversion in $\Sigma\Delta$ quantization.
Advances in Computational Mathematics, 2010, Volume 32, Number 1, pp. 73


J. Benedetto, A. Powell, O. Yilmaz,
Second order Sigma-Delta ($\Sigma\Delta$) quantization and finite frames.
Applied and Computational Harmonic Analysis, 20 (2006), 126-148


J. Benedetto, O Treiber,
Wavelet frames: Multiresolution analysis and extension principles.
Wavelet Transforms and Time-Frequency Signal Analysis, Birkhauser 2001


W. Bennett,
Spectra of quantized signals
Bell System Technical Journal, 27 (1949), 446-472

References II


B.G. Bodmann, V.I. Paulsen,
Frame paths and error bounds for sigma-delta quantization
Applied and Computational Harmonic Analysis, 22 (2007), 176-197


I. Daubechies, R. DeVore.,
Reconstructing a bandlimited function from very coarsely quantized data: A family of stable sigma-delta modulators of arbitrary order,
Annals of Mathematics, vol. 158, no. 2, pp. 679-710 (2003)


T. Goldstein et al.,
Training Quantized Nets: A Deeper Understanding.
NIPS 2017.


V. Goyal, M. Vetterli, and N. Thao,
Quantized overcomplete expansions in $\mathbb{R}^n$: Analysis, synthesis, and algorithms
IEEE vol. 44, no 1, pp. 16-31

References III

V. Goyal, J. Kovacevic, and J. Kelner,
Quantized frame expansions with erasures.
Appl. Comput. Harm. Anal., vol. 14, no 3, pp 257-275, 2003
IEEE Transactions on Acoustics, Speech, and Signal Processing, vol 25, no 5 pp. 442-448. 1977

C.S. Gunturk
One-bit sigma-delta quantization with exponential accuracy,
Communications on Pure and Applied Mathematics, 56 (2003), no. 11, pp. 229-242

H. Inose, Y. Yasuda, J. Murakami,
A Telemetering System by Code Manipulation – $\Sigma\Delta$ Modulation
IRE Trans on Space Electronics and Telemetry, Sep. 1962, pp. 204-209