Finite frames and Sigma-Delta quantization

John J. Benedetto
Norbert Wiener Center, Department of Mathematics
University of Maryland, College Park

http://www.norbertwiener.umd.edu

Norbert Wiener Center
Outline and collaborators

1. Finite frames
2. Sigma-Delta quantization – theory and implementation
3. Sigma-Delta quantization – number theoretic estimates

Collaborators: Matt Fickus (frame force); Alex Powell and Özugr Yilmaz ($\Sigma - \Delta$ quantization); Alex Powell, Aram Tangboondouangjit, and Özugr Yilmaz ($\Sigma - \Delta$ quantization and number theory).
Frames $F = \{ e_n \}_{n=1}^{N}$ for $d$-dimensional Hilbert space $H$, e.g., $H = K^d$, where $K = \mathbb{C}$ or $K = \mathbb{R}$.

Any spanning set of vectors in $K^d$ is a frame for $K^d$. 

Finite Frames
Frames

Frames $F = \{e_n\}_{n=1}^{N}$ for $d$-dimensional Hilbert space $H$, e.g., $H = \mathbb{K}^d$, where $\mathbb{K} = \mathbb{C}$ or $\mathbb{K} = \mathbb{R}$.

- Any spanning set of vectors in $\mathbb{K}^d$ is a frame for $\mathbb{K}^d$.
- $F \subseteq \mathbb{K}^d$ is $A$-tight if

$$\forall x \in \mathbb{K}^d, A\|x\|^2 = \sum_{n=1}^{N} |\langle x, e_n \rangle|^2$$
Finite Frames

Frames

$F = \{e_n\}_{n=1}^{N}$ for $d$-dimensional Hilbert space $H$, e.g., $H = K^d$, where $K = \mathbb{C}$ or $K = \mathbb{R}$.

- Any spanning set of vectors in $K^d$ is a frame for $K^d$.

$F \subseteq K^d$ is $A$-tight if

$$\forall x \in K^d, A\|x\|^2 = \sum_{n=1}^{N} |\langle x, e_n \rangle|^2$$

- If $\{e_n\}_{n=1}^{N}$ is a finite unit norm tight frame (FUN-TF) for $K^d$, with frame constant $A$, then $A = N/d$. 
Finite Frames

Frames

Frames $F = \{e_n\}_{n=1}^N$ for $d$-dimensional Hilbert space $H$, e.g., $H = \mathbb{K}^d$, where $\mathbb{K} = \mathbb{C}$ or $\mathbb{K} = \mathbb{R}$.

- Any spanning set of vectors in $\mathbb{K}^d$ is a frame for $\mathbb{K}^d$.

- $F \subseteq \mathbb{K}^d$ is $A$-tight if

$$\forall x \in \mathbb{K}^d, A\|x\|^2 = \sum_{n=1}^N |\langle x, e_n \rangle|^2$$

- If $\{e_n\}_{n=1}^N$ is a finite unit norm tight frame (FUN-TF) for $\mathbb{K}^d$, with frame constant $A$, then $A = N/d$.

- Let $\{e_n\}$ be an $A$-unit norm TF for any separable Hilbert space $H$. $A \geq 1$, and $A = 1 \Leftrightarrow \{e_n\}$ is an ONB for $H$ (Vitali).
The geometry of finite tight frames

The vertices of platonic solids are FUN-TFs.
The geometry of finite tight frames

- The vertices of platonic solids are FUN-TFs.
- Points that constitute FUN-TFs do not have to be equidistributed, e.g., ONBs and Grassmanian frames.
The geometry of finite tight frames

- The vertices of platonic solids are FUN-TFs.
- Points that constitute FUN-TFs do not have to be equidistributed, e.g., ONBs and Grassmanian frames.
- FUN-TFs can be characterized as minimizers of a “frame potential function” (with Fickus) analogous to

Finite frames and Sigma-Delta quantization – p.4/
The geometry of finite tight frames

- The vertices of platonic solids are FUN-TFs.
- Points that constitute FUN-TFs do not have to be equidistributed, e.g., ONBs and Grassmanian frames.
- FUN-TFs can be characterized as minimizers of a “frame potential function” (with Fickus) analogous to Coulomb’s Law.
Frame force and potential energy

\[ F : S^{d-1} \times S^{d-1} \setminus D \rightarrow \mathbb{R}^d \]

\[ P : S^{d-1} \times S^{d-1} \setminus D \rightarrow \mathbb{R}, \]

where \( P(a, b) = p(\|a - b\|), \quad p'(x) = -xf(x) \)

\[ CF(a, b) = (a - b)/\|a - b\|^3, \quad f(x) = 1/x^3 \]

Coulomb force
Frame force and potential energy

\[ F : S^{d-1} \times S^{d-1} \setminus D \rightarrow \mathbb{R}^d \]

\[ P : S^{d-1} \times S^{d-1} \setminus D \rightarrow \mathbb{R}, \]

where \( P(a, b) = p(\|a - b\|), \quad p'(x) = -xf(x) \)

- Coulomb force

\[ CF(a, b) = (a - b)/\|a - b\|^3, \quad f(x) = 1/x^3 \]

- Frame force

\[ FF(a, b) = \langle a, b \rangle (a - b), \quad f(x) = 1 - x^2/2 \]
Frame force and potential energy

\[ F : S^{d-1} \times S^{d-1} \setminus D \longrightarrow \mathbb{R}^d \]

\[ P : S^{d-1} \times S^{d-1} \setminus D \longrightarrow \mathbb{R}, \]

where \( P(a, b) = p(\|a - b\|), \quad p'(x) = -x f(x) \)

- **Coulomb force**
  
  \[ CF(a, b) = (a - b)/\|a - b\|^3, \quad f(x) = 1/x^3 \]

- **Frame force**
  
  \[ FF(a, b) = \langle a, b \rangle (a - b), \quad f(x) = 1 - x^2/2 \]

- **Total potential energy for the frame force**

\[ TFP(\{x_n\}) = \sum_{m=1}^{N} \sum_{n=1}^{N} \left| \langle x_m, x_n \rangle \right|^2 \]
Characterization of FUN-TFs

For the Hilbert space $H = \mathbb{R}^d$ and $N$, consider
${\{x_n\}}_1^N \in S^{d-1} \times ... \times S^{d-1}$ and

$$T F P({\{x_n\}}) = \sum_{m=1}^N \sum_{n=1}^N |< x_m, x_n |^2.$$ 

**Theorem** Let $N \leq d$. The minimum value of $T F P$, for the frame force and $N$ variables, is $N$; and the *minimizers* are precisely the orthonormal sets of $N$ elements for $\mathbb{R}^d$. 
Characterization of FUN-TFs

For the Hilbert space $H = \mathbb{R}^d$ and $N$, consider
\[
\{x_n\}_1^N \in S^{d-1} \times \ldots \times S^{d-1}
\] and

\[
T F P(\{x_n\}) = \sum_{m=1}^N \sum_{n=1}^N |< x_m, x_n >|^2.
\]

**Theorem** Let $N \leq d$. The minimum value of $T F P$, for the frame force and $N$ variables, is $N$; and the minimizers are precisely the orthonormal sets of $N$ elements for $\mathbb{R}^d$.

**Theorem** Let $N \geq d$. The minimum value of $T F P$, for the frame force and $N$ variables, is $N^2/d$; and the minimizers are precisely the FUN-TFs of $N$ elements for $\mathbb{R}^d$. 
Characterization of FUN-TFs

For the Hilbert space \( H = \mathbb{R}^d \) and \( N \), consider 
\[ \{x_n\}_1^N \in S^{d-1} \times \ldots \times S^{d-1} \] and 
\[
TFP(\{x_n\}) = \sum_{m=1}^{N} \sum_{n=1}^{N} |<x_m, x_n>|^2.
\]

**Theorem** Let \( N \leq d \). The minimum value of \( TFP \), for the frame force and \( N \) variables, is \( N \); and the minimizers are precisely the orthonormal sets of \( N \) elements for \( \mathbb{R}^d \).

**Theorem** Let \( N \geq d \). The minimum value of \( TFP \), for the frame force and \( N \) variables, is \( N^2/d \); and the minimizers are precisely the FUN-TFs of \( N \) elements for \( \mathbb{R}^d \).

**Problem** Find FUN-TFs analytically, effectively, computationally.
Sigma-Delta quantization — theory and implementation

Given $u_0$ and $\{x_n\}_{n=1}^{\infty}$

\[ u_n = u_{n-1} + x_n - q_n \]

\[ q_n = Q(u_{n-1} + x_n) \]
**A quantization problem**

**Qualitative Problem** Obtain *digital* representations for class $X$, suitable for storage, transmission, recovery.

**Quantitative Problem** Find dictionary $\{e_n\} \subseteq X$:

1. Sampling [continuous range $\mathbb{K}$ is not digital]

   $$\forall x \in X, \ x = \sum x_n e_n, \ x_n \in \mathbb{K} (\mathbb{R} \text{ or } \mathbb{C}).$$

2. Quantization. Construct finite alphabet $\mathcal{A}$ and

   $$Q : X \to \{\sum q_n e_n : q_n \in \mathcal{A} \subseteq \mathbb{K}\}$$

   such that $|x_n - q_n|$ and/or $\|x - Qx\|$ small.

**Methods** Fine quantization, e.g., PCM. Take $q_n \in \mathcal{A}$ close to given $x_n$. Reasonable in 16-bit (65,536 levels) digital audio.

Coarse quantization, e.g., $\Sigma\Delta$. Use fewer bits to exploit redundancy.
Quantization

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

\[ Q(u) = \arg\min\{|u - q| : q \in \mathcal{A}_K^\delta\} = q_u \]
Setting

Let $x \in \mathbb{R}^d$, $\|x\| \leq 1$. Suppose $F = \{e_n\}_{n=1}^{N}$ is a FUN-TF for $\mathbb{R}^d$. Thus, we have

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

with $x_n = \langle x, e_n \rangle$. Note: $A = N/d$, and $|x_n| \leq 1$.

**Goal** Find a “good” quantizer, given

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

**Example** Consider the alphabet $\mathcal{A}_1^{2} = \{-1, 1\}$, and $E_7 = \{e_n\}_{n=1}^{7}$, with $e_n = \left( \cos\left(\frac{2n\pi}{7}\right), \sin\left(\frac{2n\pi}{7}\right) \right)$. 
$\mathcal{A}_1^2 = \{-1, 1\}$ and $E_7$

\[ \Gamma_{\mathcal{A}_1^2}(E_7) = \left\{ \frac{2}{7} \sum_{n=1}^{7} q_n e_n : q_n \in \mathcal{A}_1^2 \right\} \]
 PCM

Replace \( x_n \leftrightarrow q_n = \arg\{\min |x_n - q| : q \in A^\delta_K\} \). Then \( \tilde{x} = \frac{d}{N} \sum_{n=1}^{N} q_n e_n \) satisfies

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

Not good!

Bennett’s “white noise assumption”

Assume that \( (\eta_n) = (x_n - q_n) \) is a sequence of independent, identically distributed random variables with mean 0 and variance \( \frac{\delta^2}{12} \). Then the mean square error (MSE) satisfies

\[
\text{MSE} = E\|x - \tilde{x}\|^2 \leq \frac{d}{12A} \delta^2 = \frac{(d\delta)^2}{12N}
\]
Remarks

1. Bennett’s “white noise assumption” is not rigorous, and not true in certain cases.

2. The MSE behaves like $C/A$. In the case of $\Sigma\Delta$ quantization of bandlimited functions, the MSE is $O(A^{-3})$ (Gray, Güntürk and Thao, Bin Han and Chen). PCM does not utilize redundancy efficiently.

3. The MSE only tells us about the average performance of a quantizer.
\[ A_1^2 = \{-1, 1\} \text{ and } E_7 \]

Let \( x = \left( \frac{1}{3}, \frac{1}{2} \right) \), \( E_7 = \{ (\cos\left( \frac{2n\pi}{7} \right), \sin\left( \frac{2n\pi}{7} \right)) \}_{n=1}^7 \). Consider quantizers with \( A = \{-1, 1\} \).
\[ \mathcal{A}_1^2 = \{-1, 1\} \text{ and } E_7 \]

Let \( x = \left( \frac{1}{3}, \frac{1}{2} \right) \), \( E_7 = \{(\cos(\frac{2n\pi}{7}), \sin(\frac{2n\pi}{7}))\}_{n=1}^7 \). Consider quantizers with \( \mathcal{A} = \{-1, 1\} \).
$A_1^2 = \{-1, 1\}$ and $E_7$

Let $x = \left(\frac{1}{3}, \frac{1}{2}\right)$, $E_7 = \{(\cos(\frac{2n\pi}{7}), \sin(\frac{2n\pi}{7}))\}_{n=1}^7$. Consider quantizers with $A = \{-1, 1\}$. 
\[ A_1^2 = \{-1, 1\} \] and \( E_7 \)

Let \( x = \left( \frac{1}{3}, \frac{1}{2} \right) \), \( E_7 = \{ (\cos(\frac{2n\pi}{7}), \sin(\frac{2n\pi}{7})) \}_{n=1}^7 \). Consider quantizers with \( A = \{-1, 1\} \).
Let \( F = \{ e_n \}_{n=1}^{N} \) be a frame for \( \mathbb{R}^d \), \( x \in \mathbb{R}^d \).

Define \( x_n = \langle x, e_n \rangle \).

Fix the ordering \( p \), a permutation of \( \{1, 2, \ldots, N\} \).

Quantizer alphabet \( \mathcal{A}_K^\delta \)

Quantizer function \( Q(u) = \operatorname{arg}\{\min \{ |u - q| : q \in \mathcal{A}_K^\delta \} \} \)

Define the first-order \( \Sigma \Delta \) quantizer with ordering \( p \) and with the quantizer alphabet \( \mathcal{A}_K^\delta \) by means of the following recursion.

\[
\begin{align*}
  u_n - u_{n-1} &= x_{p(n)} - q_n \\
  q_n &= Q(u_{n-1} + x_{p(n)})
\end{align*}
\]

where \( u_0 = 0 \) and \( n = 1, 2, \ldots, N \).
Stability

The following stability result is used to prove error estimates.

**Proposition** If the frame coefficients \( \{x_n\}_{n=1}^N \) satisfy

\[
|x_n| \leq (K - 1/2)\delta, \quad n = 1, \ldots, N,
\]

then the state sequence \( \{u_n\}_{n=0}^N \) generated by the first-order \( \Sigma\Delta \) quantizer with alphabet \( A_{\delta K}^\delta \) satisfies \( |u_n| \leq \delta/2, n = 1, \ldots, N. \)

The first-order \( \Sigma\Delta \) scheme is equivalent to

\[
u_n = \sum_{j=1}^n x_{p(j)} - \sum_{j=1}^n q_j, \quad n = 1, \ldots, N.
\]
Stability

The following stability result is used to prove error estimates.

**Proposition** If the frame coefficients \( \{ x_n \}_{n=1}^{N} \) satisfy

\[
|x_n| \leq (K - 1/2)\delta, \quad n = 1, \ldots, N,
\]

then the state sequence \( \{ u_n \}_{n=0}^{N} \) generated by the first-order \( \Sigma\Delta \) quantizer with alphabet \( A_{K}^{\delta} \) satisfies \( |u_n| \leq \delta/2, n = 1, \ldots, N \).

- The first-order \( \Sigma\Delta \) scheme is equivalent to

\[
u_n = \sum_{j=1}^{n} x_{p(j)} - \sum_{j=1}^{n} q_j, \quad n = 1, \ldots, N.
\]

- Stability results lead to tiling problems for higher order schemes.
**Definition** Let \( F = \{e_n\}_{n=1}^N \) be a frame for \( \mathbb{R}^d \), and let \( p \) be a permutation of \( \{1, 2, \ldots, N\} \). The variation \( \sigma(F, p) \) is

\[
\sigma(F, p) = \sum_{n=1}^{N-1} \|e_p(n) - e_p(n+1)\|
\]
Error estimate

**Definition** Let \( F = \{e_n\}_{n=1}^N \) be a frame for \( \mathbb{R}^d \), and let \( p \) be a permutation of \( \{1, 2, \ldots, N\} \). The *variation* \( \sigma(F, p) \) is

\[
\sigma(F, p) = \sum_{n=1}^{N-1} \|e_{p(n)} - e_{p(n+1)}\|.
\]

**Theorem** Let \( F = \{e_n\}_{n=1}^N \) be an \( A\)-FUN-TF for \( \mathbb{R}^d \). The approximation

\[
\tilde{x} = d \frac{1}{N} \sum_{n=1}^N q_n e_{p(n)}
\]

generated by the first-order \( \Sigma \Delta \) quantizer with ordering \( p \) and with the quantizer alphabet \( A_K^{\delta} \) satisfies

\[
\|x - \tilde{x}\| \leq \frac{(\sigma(F, p) + 1)d}{N} \frac{\delta}{2}.
\]
Let $E_7$ be the FUN-TF for $\mathbb{R}^2$ given by the 7th roots of unity. Randomly select 10,000 points in the unit ball of $\mathbb{R}^2$. Quantize each point using the $\Sigma\Delta$ scheme with alphabet $\mathcal{A}^{1/4}$. The figures show histograms for $||x - \tilde{x}||$ when the frame coefficients are quantized in their natural order $x_1, x_2, x_3, x_4, x_5, x_6, x_7$ (left) and order $x_1, x_4, x_7, x_3, x_6, x_2, x_5$ (right).
Even – odd

$$E_N = \{e_n^N\}_{n=1}^N, e_n^N = (\cos(2\pi n/N), \sin(2\pi n/N)).$$
Let $$x = \left(\frac{1}{\pi}, \sqrt{\frac{3}{17}}\right).$$

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

Let $$\tilde{x}_N$$ be the approximation given by the 1st order $\Sigma\Delta$ quantizer with alphabet $$\{-1, 1\}$$ and natural ordering. Log-log plot of $$\|x - \tilde{x}_N\|.$$
**Improved estimates**

\[ E_N = \{e_n^N\}_{n=1}^N, \text{ } N \text{th roots of unity FUN-TFs for } \mathbb{R}^2, x \in \mathbb{R}^2, \]
\[ ||x|| \leq (K - 1/2)\delta. \]

Quantize \[ x = \frac{d}{N} \sum_{n=1}^{N} x_n^N e_n^N, \quad x_n^N = \langle x, e_n^N \rangle \]
using 1st order \( \Sigma\Delta \) scheme with alphabet \( A_{K}^{\delta} \).

**Theorem** If \( N \) is even and large then \[ ||x - \tilde{x}|| \lesssim \frac{\delta \log N}{N^{5/4}}. \]
If \( N \) is odd and large then \[ \frac{\delta}{N} \lesssim ||x - \tilde{x}|| \leq \frac{(2\pi+1)d}{N} \frac{\delta}{2}. \]

**Remark** The proof uses the analytic number theory approach developed by Sinan Güntürk, and the theorem is true more generally.
Harmonic frames

Zimmermann and Goyal, Kelner, Kovačević, Thao, Vetterli.

$H = \mathbb{C}^d$. An harmonic frame $\{e_n\}_{n=1}^N$ for $H$ is defined by the rows of the Bessel map $L$ which is the complex $N$-DFT $N \times d$ matrix with $N - d$ columns removed.
Harmonic frames

Zimmermann and Goyal, Kelner, Kovačević, Thao, Vetterli.

- $H = \mathbb{C}^d$. An harmonic frame $\{e_n\}_{n=1}^N$ for $H$ is defined by the rows of the Bessel map $L$ which is the complex $N$-DFT $N \times d$ matrix with $N - d$ columns removed.

- $H = \mathbb{R}^d$, $d$ even. The harmonic frame $\{e_n\}_{n=1}^N$ is defined by the Bessel map $L$ which is the $N \times d$ matrix whose $n$th row is

$$e_n^N = \sqrt{\frac{2}{d}} \left( \cos\left(\frac{2\pi n}{N}\right), \sin\left(\frac{2\pi n}{N}\right), \ldots, \cos\left(\frac{2\pi (d/2)n}{N}\right), \sin\left(\frac{2\pi (d/2)n}{N}\right) \right).$$
Harmonic frames

Zimmermann and Goyal, Kelner, Kovačević, Thao, Vetterli.

- $H = \mathbb{C}^d$. An harmonic frame $\{e_n\}_{n=1}^N$ for $H$ is defined by the rows of the Bessel map $L$ which is the complex $N$-DFT $N \times d$ matrix with $N - d$ columns removed.

- $H = \mathbb{R}^d$, $d$ even. The harmonic frame $\{e_n\}_{n=1}^N$ is defined by the Bessel map $L$ which is the $N \times d$ matrix whose $n$th row is

$$e_n^N = \sqrt{\frac{2}{d}} \left( \cos\left(\frac{2\pi n}{N}\right), \sin\left(\frac{2\pi n}{N}\right), \ldots, \cos\left(\frac{2\pi (d/2)n}{N}\right), \sin\left(\frac{2\pi (d/2)n}{N}\right) \right).$$

- Harmonic frames are FUN-TFs.
Zimmermann and Goyal, Kelner, Kovačević, Thao, Vetterli.

- **$H = \mathbb{C}^d$.** An harmonic frame $\{e_n\}_{n=1}^N$ for $H$ is defined by the rows of the Bessel map $L$ which is the complex $N$-DFT $N \times d$ matrix with $N - d$ columns removed.

- **$H = \mathbb{R}^d$, $d$ even.** The harmonic frame $\{e_n\}_{n=1}^N$ is defined by the Bessel map $L$ which is the $N \times d$ matrix whose $n$th row is

$$e_n^N = \sqrt{\frac{2}{d}} \left( \cos\left(\frac{2\pi n}{N}\right), \sin\left(\frac{2\pi n}{N}\right), \ldots, \cos\left(\frac{2\pi (d/2)n}{N}\right), \sin\left(\frac{2\pi (d/2)n}{N}\right) \right).$$

- Harmonic frames are FUN-TFs.

- Let $E_N$ be the harmonic frame for $\mathbb{R}^d$ and let $p_N$ be the identity permutation. Then

$$\forall N, \sigma(E_N, p_N) \leq \pi d(d + 1).$$
Error estimate for harmonic frames

**Theorem** Let $E_N$ be the harmonic frame for $\mathbb{R}^d$ with frame bound $N/d$. Consider $x \in \mathbb{R}^d$, $\|x\| \leq 1$, and suppose the approximation $\tilde{x}$ of $x$ is generated by a first-order $\Sigma \Delta$ quantizer as before. Then

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

Hence, for harmonic frames (and all those with bounded variation),

$$\text{MSE}_{\Sigma \Delta} \leq \frac{C_d}{N^2} \delta^2.$$
Error estimate for harmonic frames

**Theorem** Let $E_N$ be the harmonic frame for $\mathbb{R}^d$ with frame bound $N/d$. Consider $x \in \mathbb{R}^d$, $\|x\| \leq 1$, and suppose the approximation $\tilde{x}$ of $x$ is generated by a first-order $\Sigma\Delta$ quantizer as before. Then

$$\|x - \tilde{x}\| \leq \frac{d^2(d + 1) + d\delta}{N} \cdot \frac{\delta}{2}.$$ 

Hence, for harmonic frames (and all those with bounded variation),

$$\text{MSE}_{\Sigma\Delta} \leq \frac{Cd}{N^2} \delta^2.$$ 

This bound is clearly superior asymptotically to

$$\text{MSE}_{\text{PCM}} = \frac{(d\delta)^2}{12N}.$$
The digital encoding

\[ \text{MSE}_{\text{PCM}} = \frac{(d\delta)^2}{12N} \]

in PCM format leaves open the possibility that decoding (reconstruction) could lead to

\[ \text{"MSE}_{\text{PCM}}^{\text{opt}} \ll O\left(\frac{1}{N}\right). \]

Goyal, Vetterli, Thao (1998) proved

\[ \text{"MSE}_{\text{PCM}}^{\text{opt}} \sim \frac{\tilde{C}_d}{N^2} \delta^2. \]

**Theorem** The first order \( \Sigma\Delta \) scheme achieves the asymptotically optimal \( \text{MSE}_{\text{PCM}} \) for harmonic frames.
Proof of Improved Estimates theorem

If $N$ is even and large then $||x - \tilde{x}|| \lesssim \frac{\delta \log N}{N^{3/4}}$.

If $N$ is odd and large then $\frac{\delta}{N} \lesssim ||x - \tilde{x}|| \leq \frac{(2\pi + 1)d \delta}{N}$. 
Sigma-Delta quantization–number theoretic estimates

Proof of Improved Estimates theorem

- If $N$ is even and large then $||x - \tilde{x}|| \lesssim \frac{\delta \log N}{N^{3/4}}$.
- If $N$ is odd and large then $\frac{\delta}{N} \lesssim ||x - \tilde{x}|| \leq \frac{(2\pi + 1)d \delta}{N}$.

$\forall N$, $\{e_n^N\}_{n=1}^N$ is a FUN-TF.
Sigma-Delta quantization–number theoretic estimates

Proof of Improved Estimates theorem

- If $N$ is even and large then $||x - \tilde{x}|| \lesssim \frac{\delta \log N}{N^{5/4}}$.
- If $N$ is odd and large then $\frac{\delta}{N} \lesssim ||x - \tilde{x}|| \leq \frac{(2\pi+1)d \delta}{2}.$
- $\forall N, \{e^n_N\}_{n=1}^N$ is a FUN-TF.

\[
x - \tilde{x}_N = \frac{d}{N} \left( \sum_{n=1}^{N-2} v^n_N (f^n_N - f^n_{n+1}) + v^n_{N-1} f^N_{N-1} + u^N_N e^N_N \right)
\]

\[
f^n_N = e^n_N - e^n_{n+1}, \quad v^n_N = \sum_{j=1}^{n} u^N_j, \quad \tilde{u}_n = \frac{u^N_n}{\delta}
\]
Sigma-Delta quantization–number theoretic estimates

**Proof of Improved Estimates theorem**

- If \( N \) is even and large then \( \|x - \tilde{x}\| \lesssim \frac{\delta \log N}{N^{3/4}} \).
- If \( N \) is odd and large then \( \frac{\delta}{N} \lesssim \|x - \tilde{x}\| \leq \frac{(2\pi + 1)d\delta}{N} \).

\[ \forall N, \{e_n^N\}^N_{n=1} \text{ is a FUN-TF.} \]

\[
x - \tilde{x}_N = \frac{d}{N} \left( \sum_{n=1}^{N-2} v_n^N (f_n^N - f_{n+1}^N) + v_{N-1}^N f_{N-1}^N + u_N^N e_N^N \right)
\]

\[
f_n^N = e_n^N - e_{n+1}^N, \quad v_n^N = \sum_{j=1}^{n} u_j^N, \quad \tilde{u}_n = \frac{u_n^N}{\delta}
\]

- To bound \( v_n^N \).
Koksma Inequality

**Discrepancy**

The discrepancy $D_N$ of a finite sequence $x_1, \ldots, x_N$ of real numbers is

$$D_N = D_N(x_1, \ldots, x_N) = \sup_{0 \leq \alpha < \beta \leq 1} \left| \frac{1}{N} \sum_{n=1}^{N} 1_{[\alpha, \beta]}(\{x_n\}) - (\beta - \alpha) \right|,$$

where $\{x\} = x - [x]$. 
Koksma Inequality

**Discrepancy**

The discrepancy $D_N$ of a finite sequence $x_1, \ldots, x_N$ of real numbers is

$$D_N = D_N(x_1, \ldots, x_N) = \sup_{0 \leq \alpha < \beta \leq 1} \left| \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}_{[\alpha, \beta]}(\{x_n\}) - (\beta - \alpha) \right|,$$

where $\{x\} = x - \lfloor x \rfloor$.

**Koksma Inequality**

$g : [-1/2, 1/2) \to \mathbb{R}$ of bounded variation and

$\{\omega_j\}_{j=1}^{n} \subset [-1/2, 1/2) \implies$

$$\left| \frac{1}{n} \sum_{j=1}^{n} g(\omega_j) - \int_{-\frac{1}{2}}^{\frac{1}{2}} g(t) dt \right| \leq \text{Var}(g) \text{Disc}\left(\{\omega_j\}_{j=1}^{n}\right).$$
Koksma Inequality

Discrepancy
The discrepancy $D_N$ of a finite sequence $x_1, \ldots, x_N$ of real numbers is

$$D_N = D_N(x_1, \ldots, x_N) = \sup_{0 \leq \alpha < \beta \leq 1} \left| \frac{1}{N} \sum_{n=1}^{N} 1_{[\alpha, \beta)}(\{x_n\}) - (\beta - \alpha) \right|,$$

where $\{x\} = x - [x]$.

Koksma Inequality
$g : [-1/2, 1/2) \rightarrow \mathbb{R}$ of bounded variation and
$\{\omega_j\}_{j=1}^{n} \subset [-1/2, 1/2) \implies$

$$\left| \frac{1}{n} \sum_{j=1}^{n} g(\omega_j) - \int_{-1/2}^{1/2} g(t) dt \right| \leq \text{Var}(g) \text{Disc}(\{\omega_j\}_{j=1}^{n}).$$

With $g(t) = t$ and $\omega_j = \tilde{u}_j^N$, $|v_n^N| \leq n\delta \text{Disc}(\{\tilde{u}_j^N\}_{j=1}^{n})$. 
Erdős-Turán Inequality

\[ \exists C > 0, \forall K, \text{Disc}\left(\{\tilde{u}_n^N\}_{n=1}^j\right) \leq C \left( \frac{1}{K} + \frac{1}{j} \sum_{k=1}^K \frac{1}{k} \left| \sum_{n=1}^j e^{2\pi i k \tilde{u}_n^N} \right| \right). \]
Erdös-Turán Inequality

\[ \exists C > 0, \forall K, \text{Disc}\left(\{\tilde{u}_{n}^{N}\}_{n=1}^{j}\right) \leq C\left(\frac{1}{K} + \frac{1}{j} \sum_{k=1}^{K} \frac{1}{k} \left| \sum_{n=1}^{j} e^{2\pi i k \tilde{u}_{n}^{N}} \right| \right). \]

To approximate the exponential sum.
Approximation of Exponential Sum

(1) Güntürk’s Proposition
\[ \forall N, \exists X_N \in \mathcal{B}_{\Omega/N} \]
such that \( \forall n = 0, \ldots, N \),

\[ X_N(n) = u_n^N + c_n \frac{\delta}{2}, \quad c_n \in \mathbb{Z} \]

and \( \forall t, \left| X'_N(t) - h\left(\frac{t}{N}\right)\right| \lesssim \frac{1}{N} \)

(2) Bernstein’s Inequality
If \( x \in \mathcal{B}_\Omega \), then \( \|x^{(r)}\|_\infty \leq \Omega^r \|x\|_\infty \)

\( \widehat{\mathcal{B}}_\Omega = \{T \in A'(\mathbb{R}) : \text{supp}T \subseteq [-\Omega, \Omega]\} \)

\( \mathcal{M}_\Omega = \{h \in \mathcal{B}_\Omega : h' \in L^\infty(\mathbb{R}) \text{ and all zeros of } h' \text{ on } [0, 1] \text{ are simple}\} \)

We assume \( \exists h \in \mathcal{M}_\Omega \) such that \( \forall N \) and \( \forall 1 \leq n \leq N \), \( h(n/N) = x_n^N \).
Approximation of Exponential Sum

(1) Güntürk’s Proposition
\[ \forall N, \exists X_N \in \mathcal{B}_{\Omega/N} \]
such that \( \forall n = 0, \ldots, N, \)
\[ X_N(n) = u_n^N + c_n \frac{\delta}{2}, \ c_n \in \mathbb{Z} \]
and \( \forall t, \left| X'_N(t) - h\left(\frac{t}{N}\right)\right| \lesssim \frac{1}{N} \)

(2) Bernstein’s Inequality
If \( x \in \mathcal{B}_\Omega, \) then \( \|x^{(r)}\|_\infty \leq \Omega^r \|x\|_\infty \)

\[ \forall t, \left| X''_N(t) - \frac{1}{N} h'(\frac{t}{N}) \right| \lesssim \frac{1}{N^2} \]

\[ \widehat{\mathcal{B}}_\Omega = \{ T \in A'(\mathbb{R}) : \text{supp}T \subseteq [-\Omega, \Omega] \} \]
\[ \mathcal{M}_\Omega = \{ h \in \mathcal{B}_\Omega : h' \in L^\infty(\mathbb{R}) \text{ and all zeros of } h' \text{ on } [0, 1] \text{ are simple} \} \]
We assume \( \exists h \in \mathcal{M}_\Omega \) such that \( \forall N \) and \( \forall 1 \leq n \leq N, \) \( h(n/N) = x_n^N. \)
Van der Corput Lemma

Let $a, b$ be integers with $a < b$, and let $f \in C^2([a, b])$ with $f''(x) \geq \rho > 0$ for all $x \in [a, b]$ or $f''(x) \leq -\rho < 0$ for all $x \in [a, b]$ then

$$\left| \sum_{n=a}^{b} e^{2\pi i f(n)} \right| \leq \left( |f'(b) - f'(a)| + 2 \right) \left( \frac{4}{\sqrt{\rho}} + 3 \right).$$
Van der Corput Lemma

Let $a, b$ be integers with $a < b$, and let $f \in C^2([a, b])$ with $f''(x) \geq \rho > 0$ for all $x \in [a, b]$ or $f''(x) \leq -\rho < 0$ for all $x \in [a, b]$ then

$$\left| \sum_{n=a}^{b} e^{2\pi i f(n)} \right| \leq \left( |f'(b) - f'(a)| + 2 \right) \left( \frac{4}{\sqrt{\rho}} + 3 \right).$$

\[\downarrow\]

\[\forall 0 < \alpha < 1, \exists N_\alpha \text{ such that } \forall N \geq N_\alpha, \]

$$\left| \sum_{n=1}^{j} e^{2\pi i k\tilde{u}^N_n} \right| \lesssim N^{\alpha} + \frac{\sqrt{k} N^{1 - \alpha}}{\sqrt{\delta}} + \frac{k}{\delta}.$$
Choosing appropriate $\alpha$ and $K$

Putting $\alpha = 3/4$, $K = N^{1/4}$ yields

$$\exists \tilde{N} \text{ such that } \forall N \geq \tilde{N}, \text{Disc}\left(\{\tilde{u}_n^N\}_{n=1}^j\right) \lesssim \frac{1}{N^{\frac{1}{4}}} + \frac{N^{\frac{3}{4}} \log(N)}{j}$$
Choosing appropriate $\alpha$ and $K$

Putting $\alpha = 3/4$, $K = N^{1/4}$ yields

$$\exists \tilde{N} \text{ such that } \forall N \geq \tilde{N}, \text{Disc}\left( \{\tilde{u}_n^N\}_{n=1}^j \right) \lesssim \frac{1}{N^{1/4}} + \frac{N^{3/4} \log(N)}{j}$$

↓

Conclusion

$$\forall n = 1, \ldots, N, |v_n^N| \lesssim \delta N^{3/4} \log N$$
That's all folks!

Norbert Wiener Center