# Waveform design and Sigma-Delta quantization

John J. Benedetto

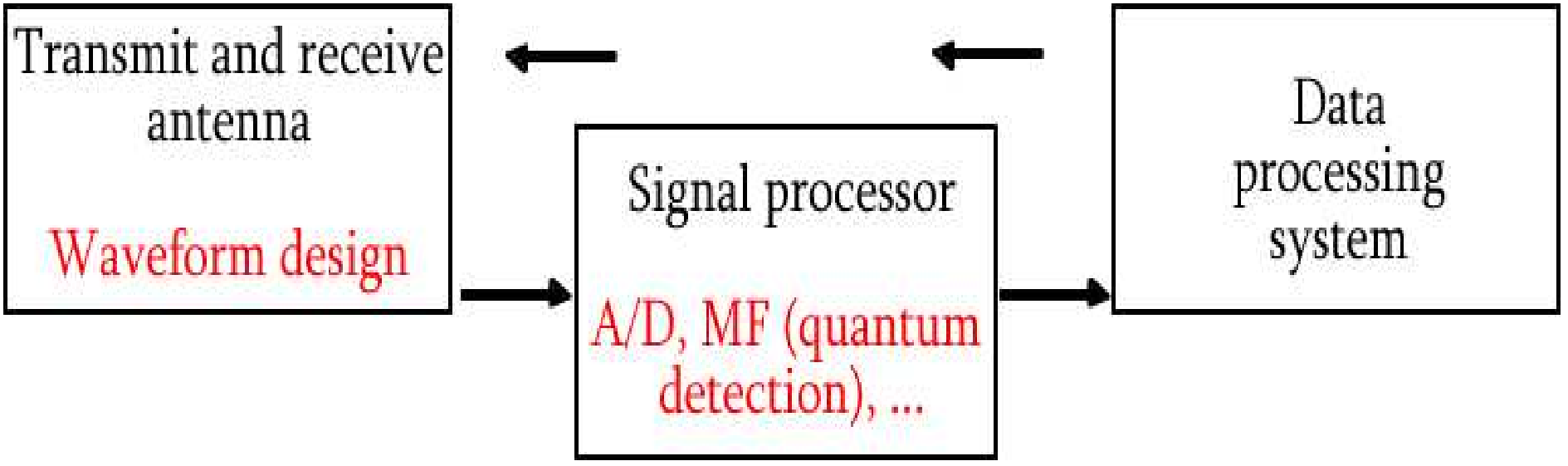
Norbert Wiener Center
Department of Mathematics
University of Maryland, College Park
*http://www.norbertwiener.umd.edu*

# Outline

# Collaborators

- Jeff Donatelli (Waveform Design)
- Matt Fickus (Frame Force)
- Alex Powell and Özgür Yilmaz (ΣΔ theory and implementation)
- Onur Oktay (ΣΔ pointwise comparison and implementation)
- Alex Powell, Aram Tangboondouangjit, and Özgür Yilmaz (ΣΔ number theory estimates)

# Processing


Transmit and receive antenna
Waveform design
Signal processor
A/D, MF (quantum detection), ...
Data processing system

# Outline

# CAZAC waveforms

## Definition of CAZAC waveforms

A $K$-periodic waveform $u : \mathbb{Z}_K = \{0, 1, \ldots, K-1\} \to \mathbb{C}$ is *Constant Amplitude Zero Autocorrelation (CAZAC)* if,

$$\text{for all } k \in \mathbb{Z}_K, |u[k]| = 1, \quad \text{(CA)}$$

and, for $m = 1, \ldots, K-1$, the *autocorrelation*

$$A_u[m] = 1/K \sum_{k=0}^{K-1} u[m+k]\overline{u}[k] \text{ is } 0. \quad \text{(ZAC)}$$

The *crosscorrelation* of $u, v : \mathbb{Z}_K \to \mathbb{C}$ is

$$C_{u,v}[m] = 1/K \sum_{k=0}^{K-1} u[m+k]\overline{v}[k]$$

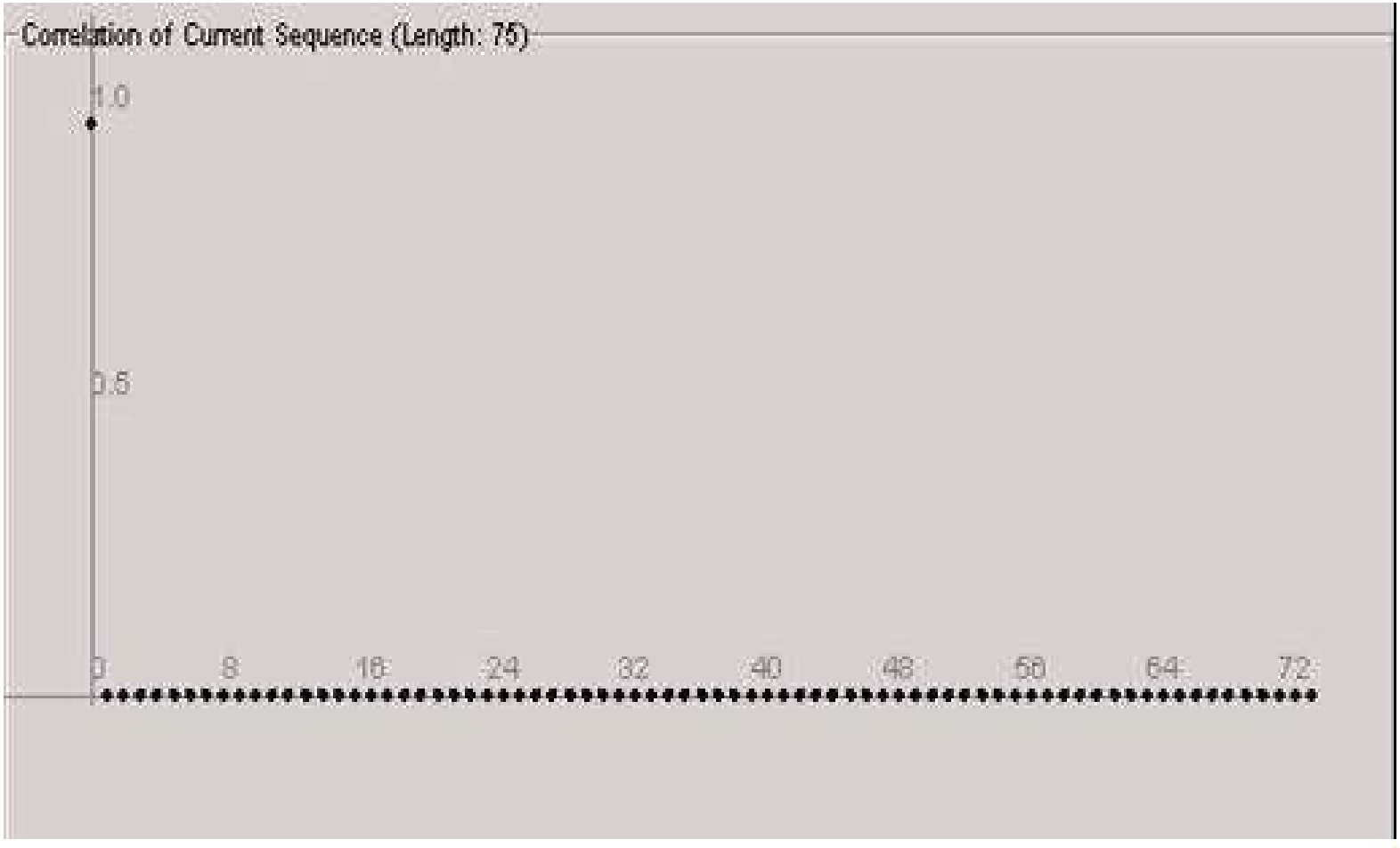
# Properties of CAZAC waveforms

- $u$ CAZAC $\Rightarrow$ $u$ is broadband (full bandwidth).
- There are different constructions of different CAZAC waveforms resulting in different behavior vis à vis Doppler, additive noises, and approximation by bandlimited waveforms.
- $u$ CA $\Leftrightarrow$ DFT of $u$ is ZAC off dc. (DFT of $u$ can have zeros)
- $u$ CAZAC $\Leftrightarrow$ DFT of $u$ is CAZAC.
- User friendly software: http://www.math.umd.edu/∼jjb/cazac

# Examples of CAZAC Waveforms

$$K = 75 : u(x) =$$
$$(1, 1, 1, 1, 1, 1, e^{2\pi i \frac{1}{15}}, e^{2\pi i \frac{2}{15}}, e^{2\pi i \frac{1}{5}}, e^{2\pi i \frac{4}{15}}, e^{2\pi i \frac{1}{3}}, e^{2\pi i \frac{7}{15}}, e^{2\pi i \frac{3}{5}},$$
$$e^{2\pi i \frac{11}{15}}, e^{2\pi i \frac{13}{15}}, 1, e^{2\pi i \frac{1}{5}}, e^{2\pi i \frac{2}{5}}, e^{2\pi i \frac{3}{5}}, e^{2\pi i \frac{4}{5}}, 1, e^{2\pi i \frac{4}{15}}, e^{2\pi i \frac{8}{15}}, e^{2\pi i \frac{4}{5}},$$
$$e^{2\pi i \frac{16}{15}}, e^{2\pi i \frac{1}{3}}, e^{2\pi i \frac{2}{3}}, e^{2\pi i}, e^{2\pi i \frac{4}{3}}, e^{2\pi i \frac{5}{3}}, 1, e^{2\pi i \frac{2}{5}}, e^{2\pi i \frac{4}{5}}, e^{2\pi i \frac{6}{5}},$$
$$e^{2\pi i \frac{8}{5}}, 1, e^{2\pi i \frac{7}{15}}, e^{2\pi i \frac{14}{15}}, e^{2\pi i \frac{7}{5}}, e^{2\pi i \frac{28}{15}}, e^{2\pi i \frac{1}{3}}, e^{2\pi i \frac{13}{15}}, e^{2\pi i \frac{7}{5}}, e^{2\pi i \frac{29}{15}},$$
$$e^{2\pi i \frac{37}{15}}, 1, e^{2\pi i \frac{3}{5}}, e^{2\pi i \frac{6}{5}}, e^{2\pi i \frac{9}{5}}, e^{2\pi i \frac{12}{5}}, 1, e^{2\pi i \frac{2}{3}}, e^{2\pi i \frac{4}{3}}, e^{2\pi i \cdot 2}, e^{2\pi i \frac{8}{3}},$$
$$e^{2\pi i \frac{1}{3}}, e^{2\pi i \frac{16}{15}}, e^{2\pi i \frac{9}{5}}, e^{2\pi i \frac{38}{15}}, e^{2\pi i \frac{49}{15}}, 1, e^{2\pi i \frac{4}{5}}, e^{2\pi i \frac{8}{5}}, e^{2\pi i \frac{12}{5}}, e^{2\pi i \frac{16}{5}},$$
$$1, e^{2\pi i \frac{13}{15}}, e^{2\pi i \frac{26}{15}}, e^{2\pi i \frac{13}{5}}, e^{2\pi i \frac{52}{15}}, e^{2\pi i \frac{1}{3}}, e^{2\pi i \frac{19}{15}}, e^{2\pi i \frac{11}{5}}, e^{2\pi i \frac{47}{15}}, e^{2\pi i \frac{61}{15}})$$

# Autocorrelation of CAZAC $K = 75$



# Perspective

Sequences for coding theory, cryptography, and communications (synchronization, fast start-up equalization, frequency hopping) include the following in the periodic case:

- Gauss, Wiener (1927), Zadoff (1963), Schroeder (1969), Chu (1972), Zhang and Golomb (1993)
- Frank (1953), Zadoff and Abourezk (1961), Heimiller (1961)
- Milewski (1983)
- Björck (1985) and Golomb (1992).

and their generalizations, both periodic and aperiodic, with some being equivalent in various cases.

# Finite ambiguity function

Given $K$-periodic waveform, $u : \mathbb{Z}_K \to \mathbb{C}$ let $e_j[k] = e^{-2\pi ikj/K}$.

- The *ambiguity function* of $u$, $A : \mathbb{Z}_K \times \mathbb{Z}_K \to \mathbb{C}$ is defined as

$$A_u[m, j] = C_{u,ue_j}[m] = 1/K \sum_{k=0}^{K-1} u[k+m]\overline{u}[k]e^{2\pi ikj/K}.$$

- Analogue ambiguity function for $u \leftrightarrow U$, $\|u\|_2 = 1$, on $\mathbb{R}$:

$$A_u(t, \gamma) = \int_{\widehat{\mathbb{R}}} U(\omega - \gamma/2)\overline{U(\omega + \gamma/2)}e^{2\pi it(\omega+\gamma/2)}d\omega$$

$$= \int u(s+t)\overline{u(s)}e^{2\pi is\gamma}ds.$$

# Rationale and theorem

Different CAZACs exhibit different behavior in their ambiguity plots, according to their construction method. Thus, the ambiguity function reveals localization properties of different constructions.

## Theorem

Given $K$ odd, $\zeta = e^{\frac{2\pi i}{K}}$, and $u[k] = \zeta^{k^2}$, $k \in \mathbb{Z}_K$

- $1 \leq k \leq K-2$ odd implies

$$A[m,k] = e^{\pi i (K-k)^2/K} \text{ for } m = \frac{1}{2}(K-k), \text{ and 0 elsewhere}$$

- $2 \leq k \leq K-1$ even implies

$$A[m,k] = e^{\pi i (2K-k)^2/K} \text{ for } m = \frac{1}{2}(2K-k), \text{ and 0 elsewhere}$$

# Rationale and theorem

## Theorem 1

Given $N \geq 1$. Let

$$M = \begin{cases} N, & \text{N odd,} \\ 2N, & \text{N even,} \end{cases}$$

and let $\omega$ be a primitive $M$th root of unity. Define the Wiener waveform $u : \mathbb{Z}_N \to \mathbb{C}$ by $u(k) = \omega^{k^2}$, $0 \leq k \leq N-1$. Then $u$ is a CAZAC waveform.

# Rationale and theorem

## Theorem 2

Let $j \in \mathbb{Z}$. Define $u_j : \mathbb{Z}_N \to \mathbb{C}$ by $u_j(k) = e^{2\pi ijk^2/M}$, where $M = 2N$ if $N$ is even and $M = N$ if $N$ is odd. If $N$ is even, then

$$A_{u_j}(m,n) = \begin{cases} e^{2\pi ijm^2/(2N)}, & jm + n \equiv 0 \ mod \ N, \\ 0, & \text{otherwise.} \end{cases}$$

If $N$ is odd

$$A_{u_j}(m,n) = \begin{cases} e^{2\pi ijm^2/N}, & 2jm + n \equiv 0 \ mod \ N, \\ 0, & \text{otherwise.} \end{cases}$$

# Rationale and theorem

*Proof.* Let $N$ be even, and set $u_j(k) = e^{\pi i j k^2/N}$. We calculate

$$\begin{aligned} A_{u_j}(m,n) &= \frac{1}{N}\sum_{k=0}^{N-1} u_j(m+k)\overline{u_j(k)}e^{2\pi i k n/N} \\ &= \frac{1}{N}\sum_{k=0}^{N-1} e^{(\pi i/N)(jm^2+2jkm+2kn)} = e^{\pi i j m^2/N}\frac{1}{N}\sum_{k=0}^{N-1} e^{2\pi i k(jm+n)/N}. \end{aligned}$$

If $jm + n \equiv 0 \bmod N$, then

$$\frac{1}{N}\sum_{k=0}^{N-1} e^{2\pi i k(jm+n)/N} = 1.$$

Otherwise, we have

$$\frac{1}{N}\sum_{k=0}^{N-1} e^{2\pi i k(jm+n)/N} = \frac{e^{(2\pi i(jm+n)/N)N} - 1}{e^{2\pi i(jm+n)/N} - 1} = 0.$$

# Rationale and theorem

*Proof.(Continued)* Let $N$ be odd, and set $u(k) = e^{2\pi ik^2/N}$. We calculate

$$\begin{aligned} A_{u_j}(m,n) &= \frac{1}{N}\sum_{k=0}^{N-1} u_j(m+k)\overline{u_j(k)}e^{2\pi ikn/N} \\ &= \frac{1}{N}\sum_{k=0}^{N-1} e^{(2\pi i/N)(jm^2+2jkm+kn)} = e^{2\pi ijm^2/N}\frac{1}{N}\sum_{k=0}^{N-1} e^{2\pi ik(2jm+n)/N}. \end{aligned}$$

If $2jm + n \equiv 0 \bmod N$, then

$$\frac{1}{N}\sum_{k=0}^{N-1} e^{2\pi ik(2jm+n)/N} = 1.$$

Otherwise, we have

$$\frac{1}{N}\sum_{k=0}^{N-1} e^{2\pi ik(2jm+n)/N} = \frac{e^{2\pi i(2m+n)/N)N} - 1}{e^{(2\pi i(2m+n)/N} - 1} = 0.$$

# Rationale and theorem

## Corollary

Let $\{u(k)\}_{k=0}^{N-1}$ be a Wiener CAZAC waveform as given in Theorem 1. (In particular, $\omega$ is a primitive $M$-th root of unity.)
If $N$ is even, then

$$A_u(m,n) = \begin{cases} \omega^{m^2}, & m \equiv -n \ mod \ N, \\ 0, & \text{otherwise.} \end{cases}$$

If $N$ is odd, then

$$A_u(m,n) = \begin{cases} \omega^{m^2}, & m \equiv -n(N+1)/2 \ mod \ N, \\ 0, & \text{otherwise.} \end{cases}$$

# Rationale and theorem

## Example

a. Let $N$ be odd and let $\omega = e^{2\pi i/N}$. Then, $u(k) = \omega^{k^2}$, $0 \leq k \leq N-1$, is a CAZAC waveform. By the Corollary, $|A_u(m,n)| = |\omega^{m^2}| = 1$ if $2m+n = l_{m,n}N$ for some $l_{m,n} \in \mathbb{Z}$ and $|A_u(m,n)| = 0$ otherwise, i.e., $A_u(m,n) = 0$ on $\mathbb{Z}_N \times \mathbb{Z}_N$ unless $2m+n \equiv 0 \bmod N$. In the case $2m+n = l_{m,n}N$ for some $l_{m,n} \in \mathbb{Z}$, we have the following phenomenon.

# Rationale and theorem

## Example (Continued)

If $0 \leq m \leq \frac{N-1}{2}$ and $2m+n = l_{m,n}N$ for some $l_{m,n} \in \mathbb{Z}$, then $n$ is odd; and if $\frac{N+1}{2} \leq m \leq N-1$ and $2m+n = l_{m,n}N$ for some $l_{m,n} \in \mathbb{Z}$, then $n$ is even. Thus, the values $(m,n)$ in the domain of the discrete periodic ambiguity function $A_u$, for which $A_u(m,n) = 0$, appear as two parallel discrete lines. The line whose domain is $0 \leq m \leq \frac{N-1}{2}$ has odd function values $n$; and the line whose domain is $\frac{N+1}{2} \leq m \leq N-1$ has even function values $n$.

# Rationale and theorem

## Example

b. The behavior observed in (a) has extensions for primitive and non-primitive roots of unity.

Let $u : \mathbb{Z}_N \to \mathbb{C}$ be a Wiener waveform. Thus, $u(k) = \omega^{k^2}$, $0 \leq k \leq N-1$, and $\omega = e^{2\pi ij/M}$, $(j, M) = 1$, where $M$ is defined in terms of $N$ in Theorem 1. By the Corollary, for each fixed $n \in \mathbb{Z}_N$, the function $A_u(\bullet, n)$ of $m$ vanishes everywhere except for a *unique* value $m_n \in \mathbb{Z}_N$ for which $|A_u(m_n, n)| = 1$.

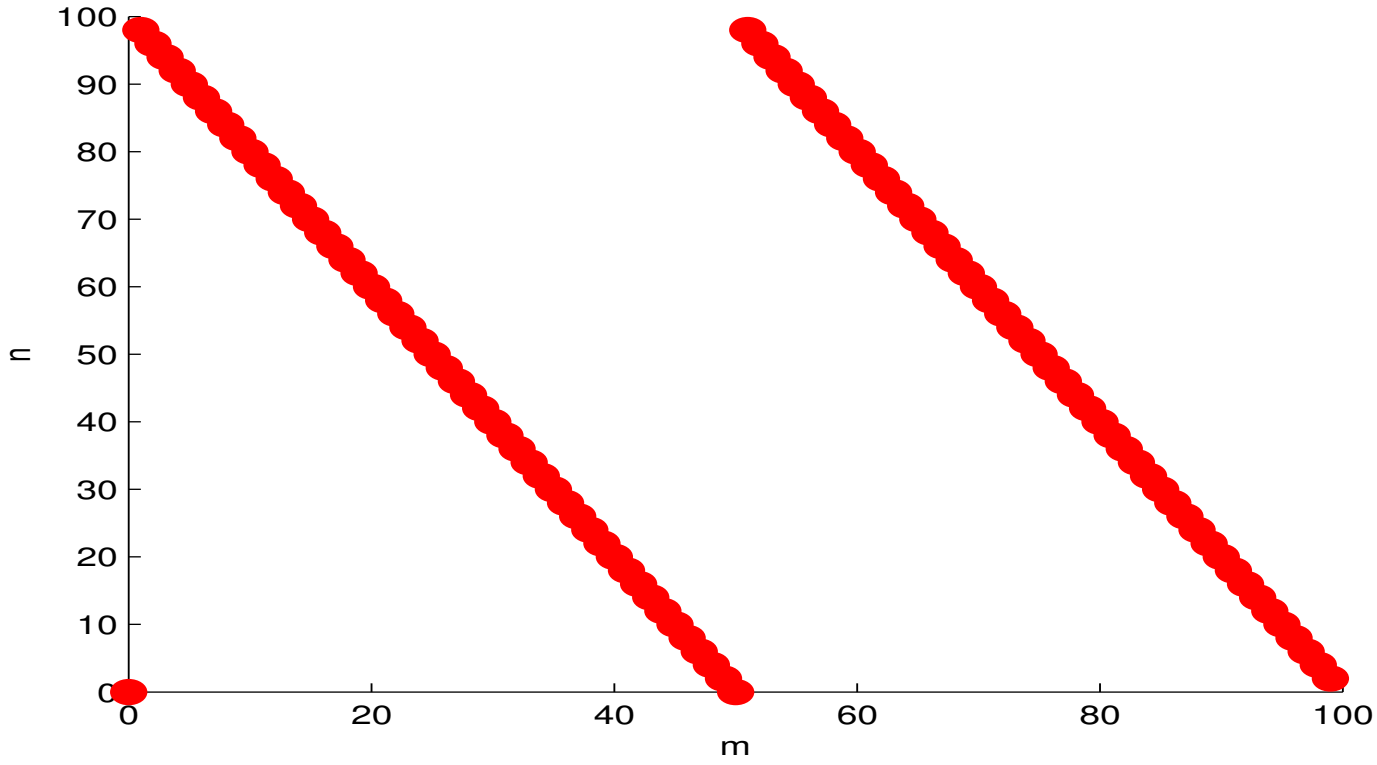
# Rationale and theorem

## Example (Continued)

The hypotheses of Theorem 2 do not assume that $e^{2\pi ij/M}$ is a primitive $M$th root of unity. In fact, in the case that $e^{2\pi ij/M}$ is *not* primitive, then, for certain values of $n$, $A_u(\bullet, n)$ will be identically 0 and, for certain values of $n$, $|A_u(\bullet, n)| = 1$ will have several solutions. For example, if $N = 100$ and $j = 2$, then, for each odd $n$, $A_u(\bullet, n) = 0$ as a function of $m$. If $N = 100$ and $j = 3$, then $(100, 3) = 1$ so that $e^{2\pi i3/100}$ is a primitive 100th root of unity; and, in this case, for each $n \in \mathbb{Z}_N$ there is a *unique* $m_n \in \mathbb{Z}_N$ such that $|A_u(m_n, n)| = 1$ and $A_u(m, n) = 0$ for each $m \neq m_n$.

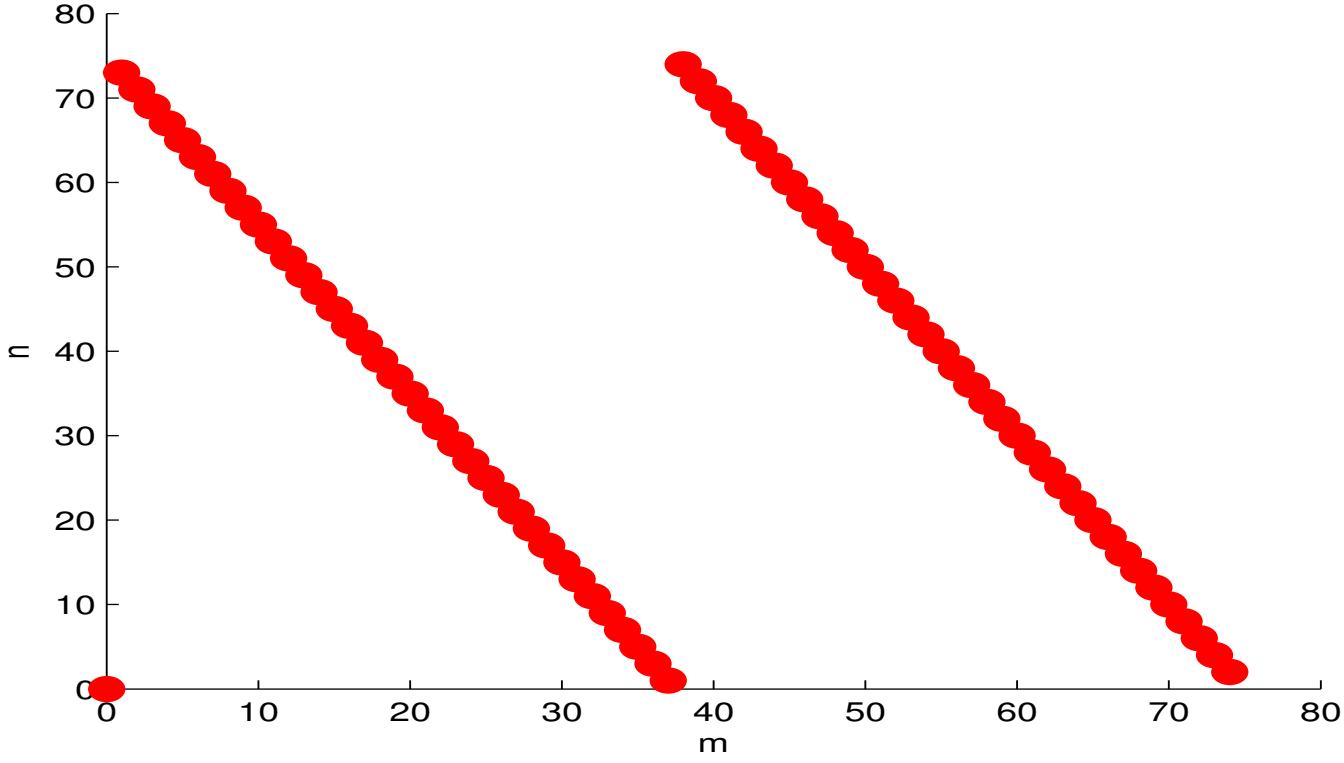
# Wiener CAZAC ambiguity domain

$K = 100, j = 2$



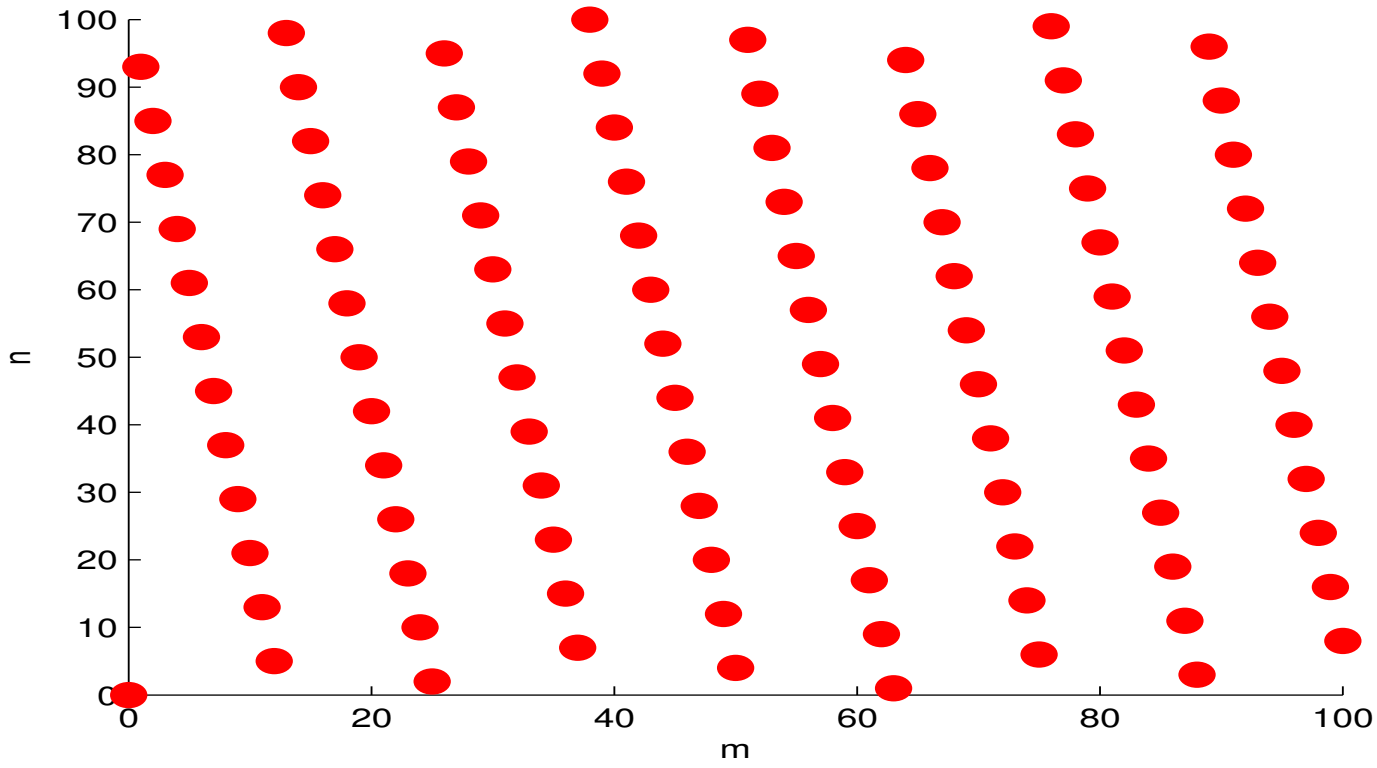
# Wiener CAZAC ambiguity domain

$K = 75, j = 1$



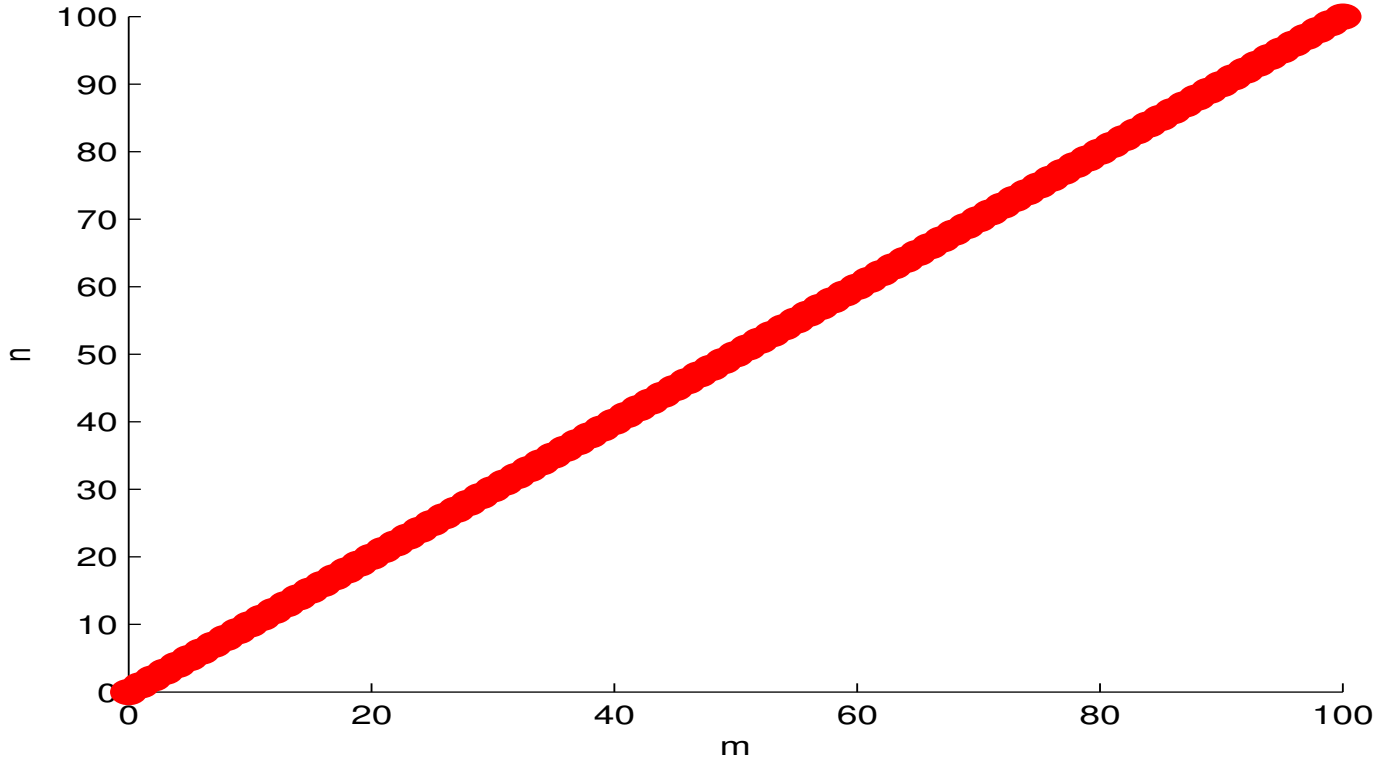
# Wiener CAZAC ambiguity domain

$K = 101, j = 4$



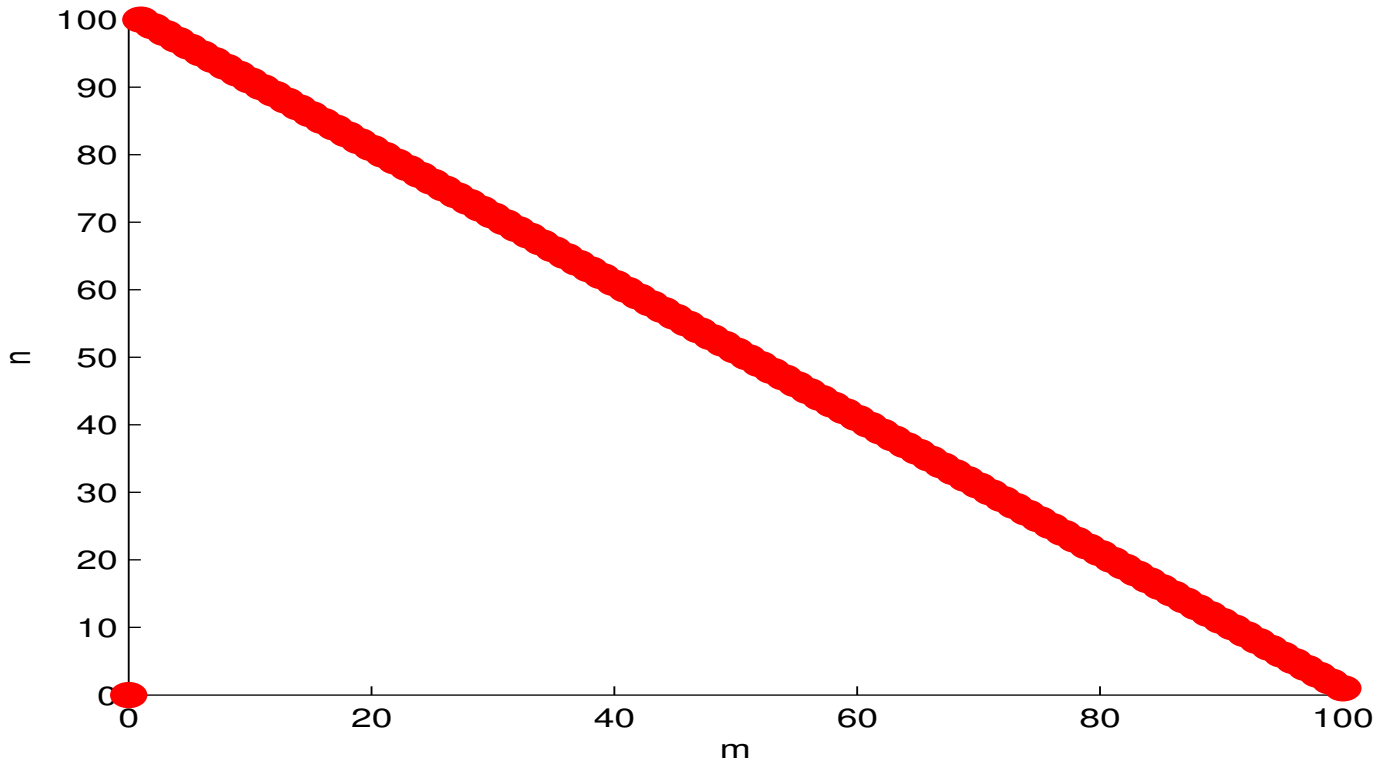
# Wiener CAZAC ambiguity domain

$K = 101, j = 50$



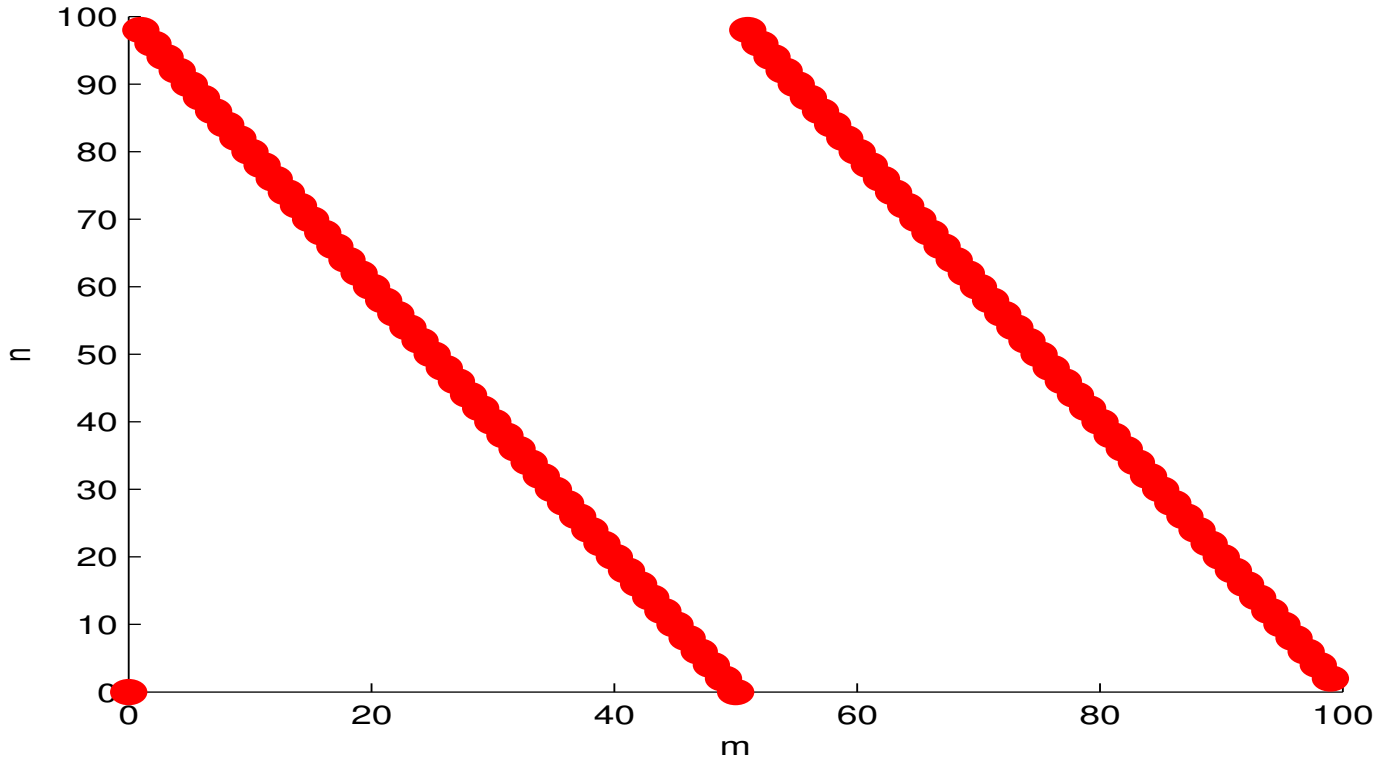
# Wiener CAZAC ambiguity domain

$K = 101, j = 51$



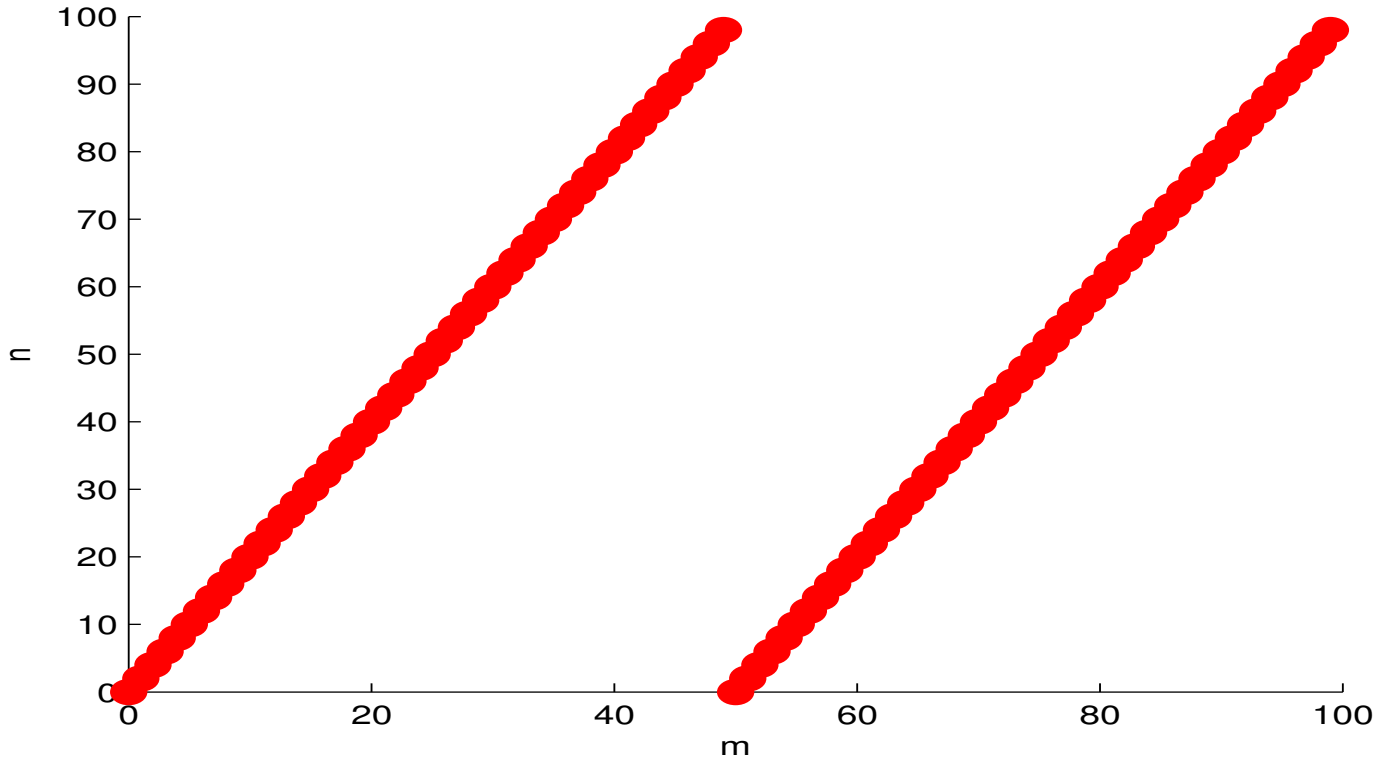
# Wiener CAZAC ambiguity domain

$K = 100, j = 2$



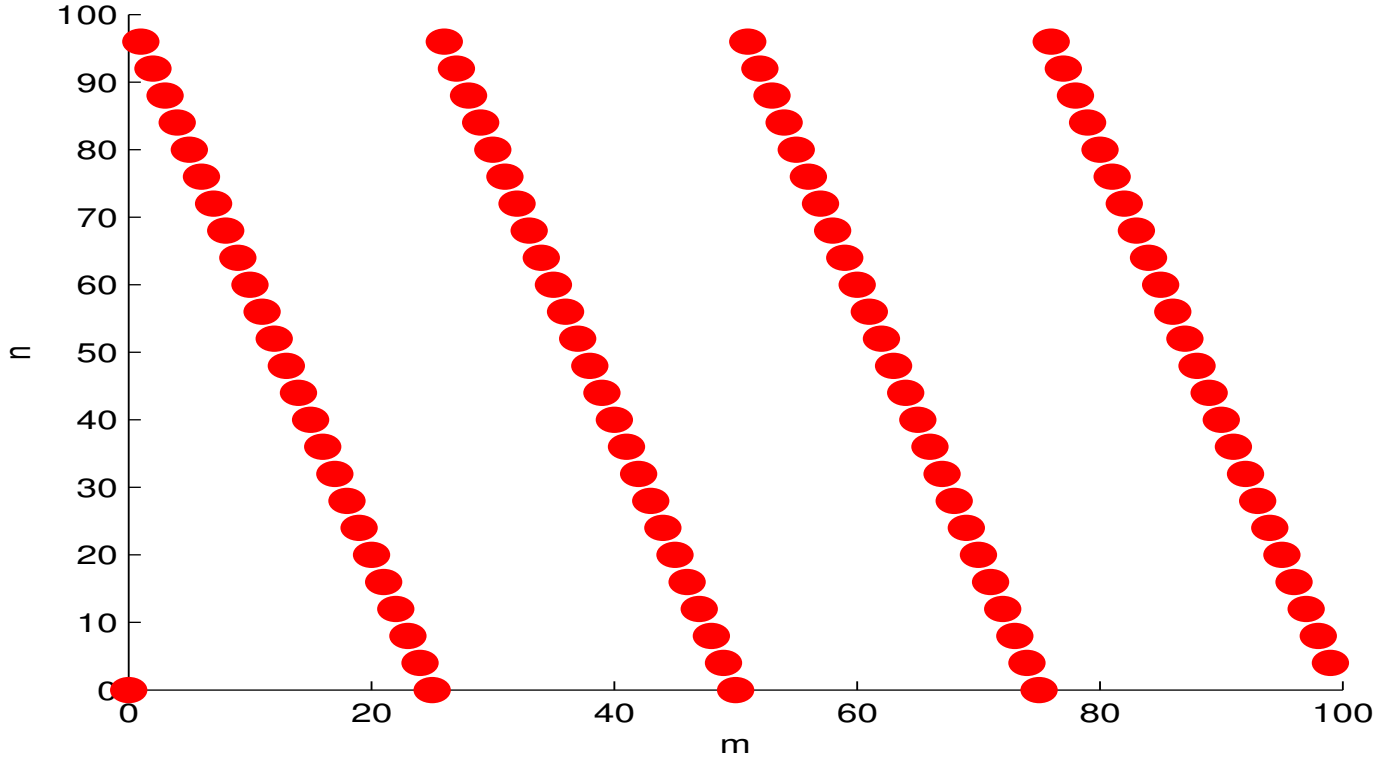
# Wiener CAZAC ambiguity domain

$K = 100, j = 98$



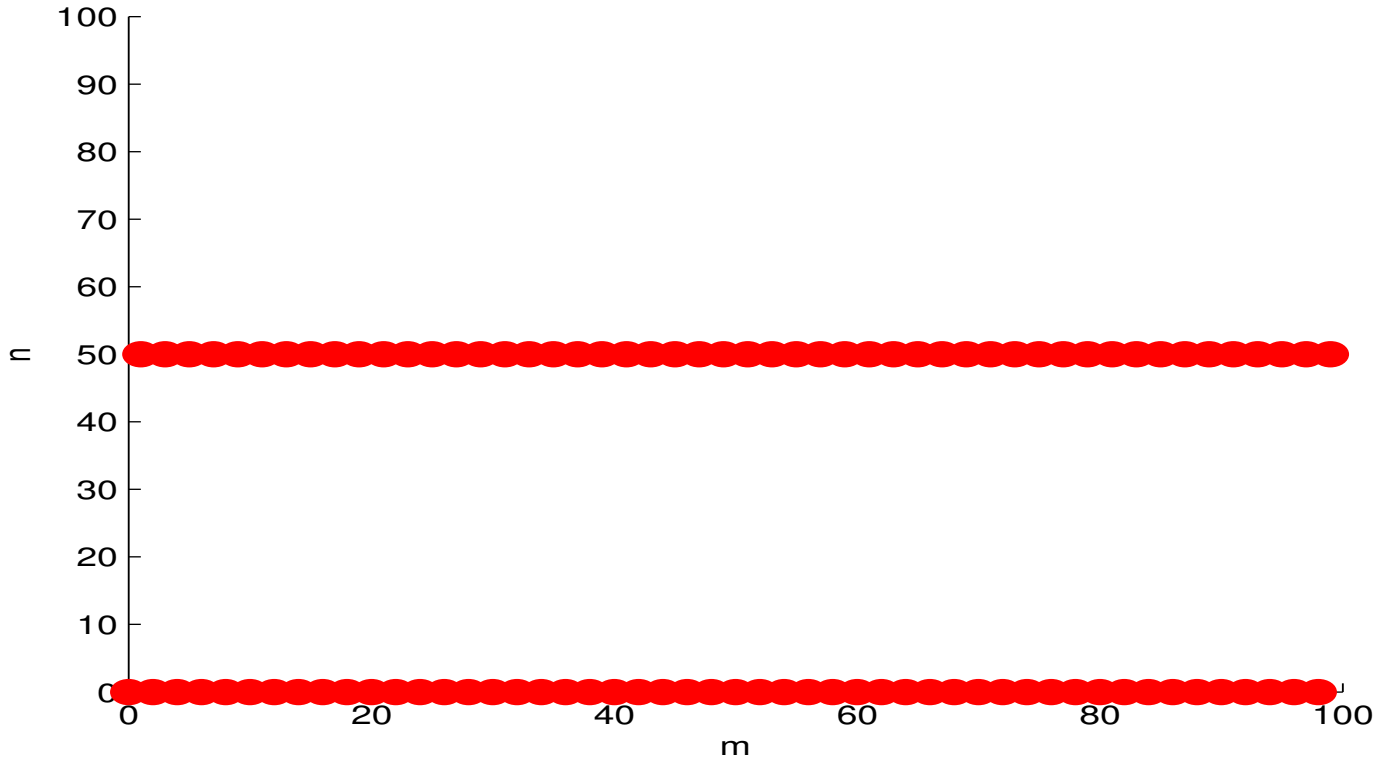
# Wiener CAZAC ambiguity domain

$K = 100, j = 4$



# Wiener CAZAC ambiguity domain

$K = 100, j = 50$



# Outline

# Frames

## Definition

A collection $(e_n)_{n\in\Lambda}$ in a Hilbert space $\mathcal{H}$ is a *frame* for $\mathcal{H}$ if there exist $0 < A \leq B < \infty$ such that

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

The constants $A$ and $B$ are the *frame bounds*. If $A = B$, the frame is an $A$-*tight* frame.

# Frames

## Definition

- Bessel (analysis) operator L: $\mathcal{H} \to \ell^2(\Lambda)$

$$Lx = (\langle x, e_n \rangle)$$

- Synthesis operator $L^*$, the Hilbert space adjoint of $L$
- Frame operator $S = L^*L : \mathcal{H} \to \mathcal{H}$,

$$Sx = \sum \langle x, e_n \rangle e_n.$$

  By the definition of frames, $S$ satisfies $AI \leq S \leq BI$.
- Grammian operator $G = LL^* : \ell^2(\Lambda) \to \ell^2(\Lambda)$.

# Frames

$AI \leq S \leq BI$ implies that $S$ is invertible and that $B^{-1}I \leq S^{-1} \leq A^{-1}I$.

## Definition

Let $F = \{e_n\}$ be a frame, and let $\widetilde{e}_n = S^{-1}e_n$. $\widetilde{F} = \{\widetilde{e}_n\}$ is the *dual frame* of $F$.

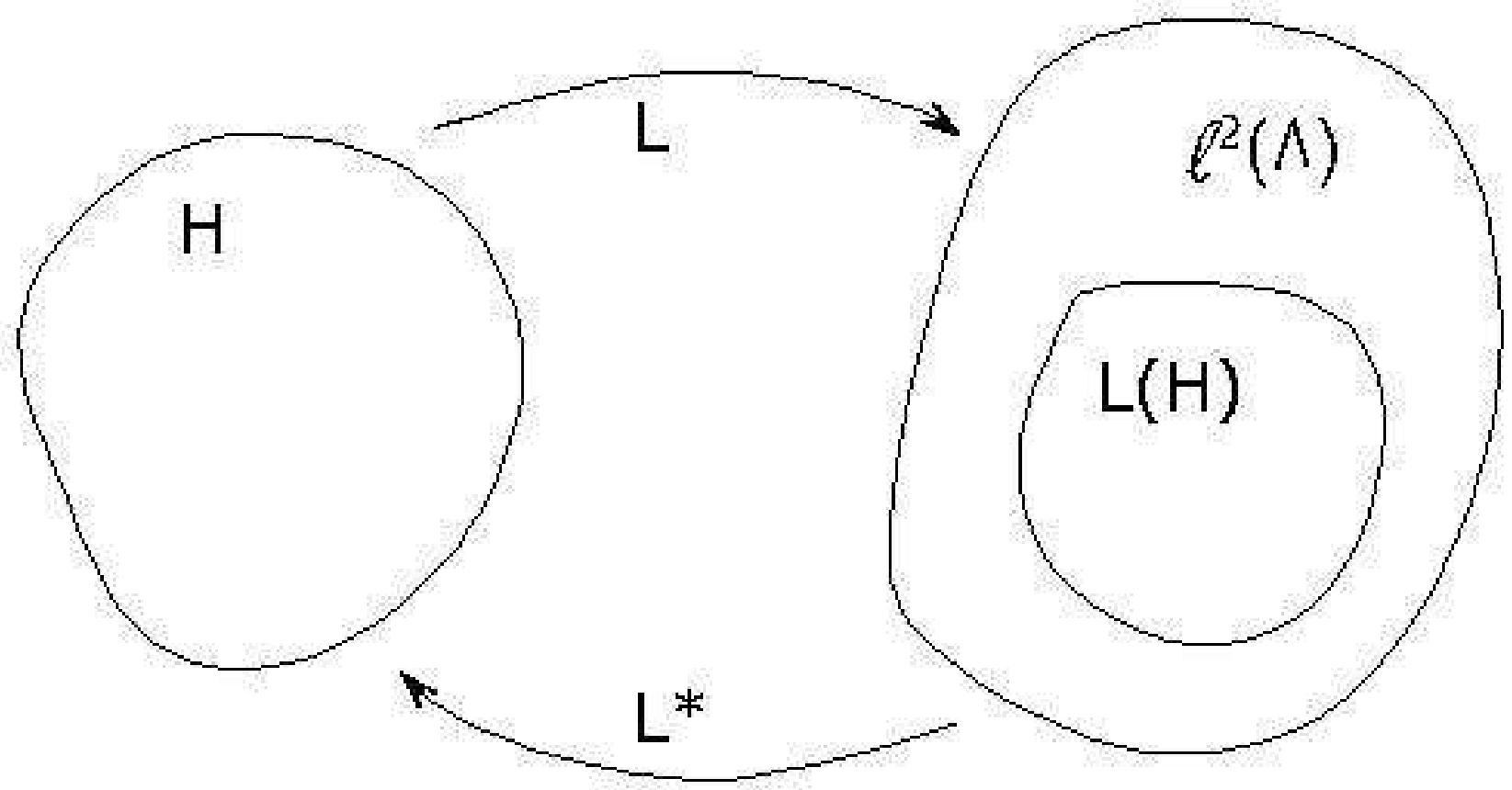
- $\sum \langle x, e_n \rangle \widetilde{e}_n = S^{-1} \sum \langle x, e_n \rangle \widetilde{e}_n = S^{-1} S x = x.$
- $\sum \langle x, \widetilde{e}_n \rangle e_n = \sum \langle S^{-1}x, e_n \rangle e_n = S S^{-1} x = x.$
- The frame operator of $\widetilde{F}$ is $S^{-1}$ since

$$\sum \langle x, \widetilde{e}_n \rangle \widetilde{e}_n = S^{-1} \sum \langle S^{-1}x, e_n \rangle e_n = S^{-1} S S^{-1} x = S^{-1} x.$$

- $\sum |\langle x, \widetilde{e}_n \rangle|^2 = \langle S^{-1}x, x \rangle$. Then,

$$B^{-1}\|x\|^2 \leq \sum |\langle x, \widetilde{e}_n \rangle|^2 \leq A^{-1}\|x\|^2.$$

# Frames



# Frames

**Theorem**

Let $H$ be a Hilbert space.

$$\{e_n\}_{n\in\Lambda} \subseteq H \text{ is } A\text{-tight} \Leftrightarrow S = AI,$$

where $I$ is the identity operator.

*Proof.* ($\Rightarrow$) If $S = L^*L = AI$, then $\forall x \in H$

$$\begin{aligned} A\|x\|^2 &= A\langle x, x\rangle = \langle Ax, x\rangle = \langle Sx, x\rangle \\ &= \langle L^*Lx, x\rangle = \langle Lx, Lx\rangle \\ &= \|Ly\|^2_{l^2(\Lambda)} \\ &= \sum_{i\in\Lambda} |\langle x, e_i\rangle|^2. \end{aligned}$$

# Frames

*Proof.* ($\Leftarrow$) If $\{e_i\}_{i\in\Lambda}$ is $A$-tight, then $\forall x \in H$, $A\langle x, x\rangle$ is

$$A\|x\|^2 = \sum_{i\in K} |\langle x, e_i\rangle|^2 = \sum_{i\in K} \langle x, e_i\rangle\langle e_i, x\rangle = \left\langle \sum_{i\in K} \langle x, e_i\rangle e_i, x \right\rangle = \langle Sx, x\rangle.$$

Therefore,

$$\forall x \in H, \quad \langle (S - AI)x, x\rangle = 0.$$

In particular, $S - AI$ is Hermitian and positive semi-definite, so

$$\forall x, y \in H, \quad |\langle (S - AI)x, y\rangle| \leq \sqrt{\langle (S - AI)x, x\rangle\langle (S - AI)y, y\rangle} = 0.$$

Thus, $(S - AI) = 0$, so, $S = AI$.

# Frames

**Theorem (Vitali, 1921)**

Let $H$ be a Hilbert space, $\{e_n\} \subseteq H$, $\|e_n\| = 1$.

$$\{e_n\} \text{ is 1-tight} \Leftrightarrow \{e_n\} \text{ is an ONB.}$$

*Proof.* If $\{e_n\}$ is 1-tight, then $\forall y \in H$

$$\|y\|^2 = \sum_n |\langle y, e_n \rangle|^2.$$

Since each $\|e_n\| = 1$, we have

$$1 = \|e_n\|^2 = \sum_k |\langle e_n, e_k \rangle|^2 = 1 + \sum_{k,k \neq n} |\langle e_n, e_k \rangle|^2$$

$$\Rightarrow \sum_{k \neq n} |\langle e_n, e_k \rangle|^2 = 0 \Rightarrow \forall n \neq k, \ \langle e_n, e_k \rangle = 0$$

# Finite frames

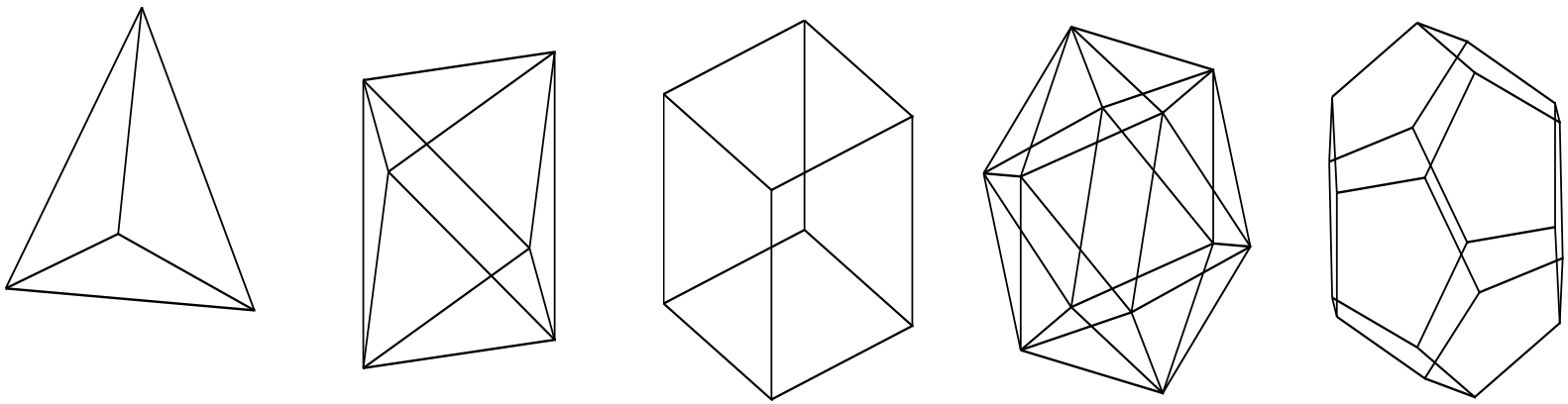
Frames $F = \{e_n\}_{n=1}^N$ for $d$-dimensional Hilbert space $H$, e.g., $H = \mathbb{K}^d$, where $\mathbb{K} = \mathbb{C}$ or $\mathbb{K} = \mathbb{R}$.

- Any spanning set of vectors in $\mathbb{K}^d$ is a *frame* for $\mathbb{K}^d$.
- If $\{e_n\}_{n=1}^N$ is a finite unit norm tight frame (FUNTF) for $\mathbb{K}^d$, with frame constant $A$, then $A = N/d$.
- $\{e_n\}_{n=1}^d$ is a $A$-tight frame for $\mathbb{K}^d$, then it is a $\sqrt{A}$-normed orthogonal set.

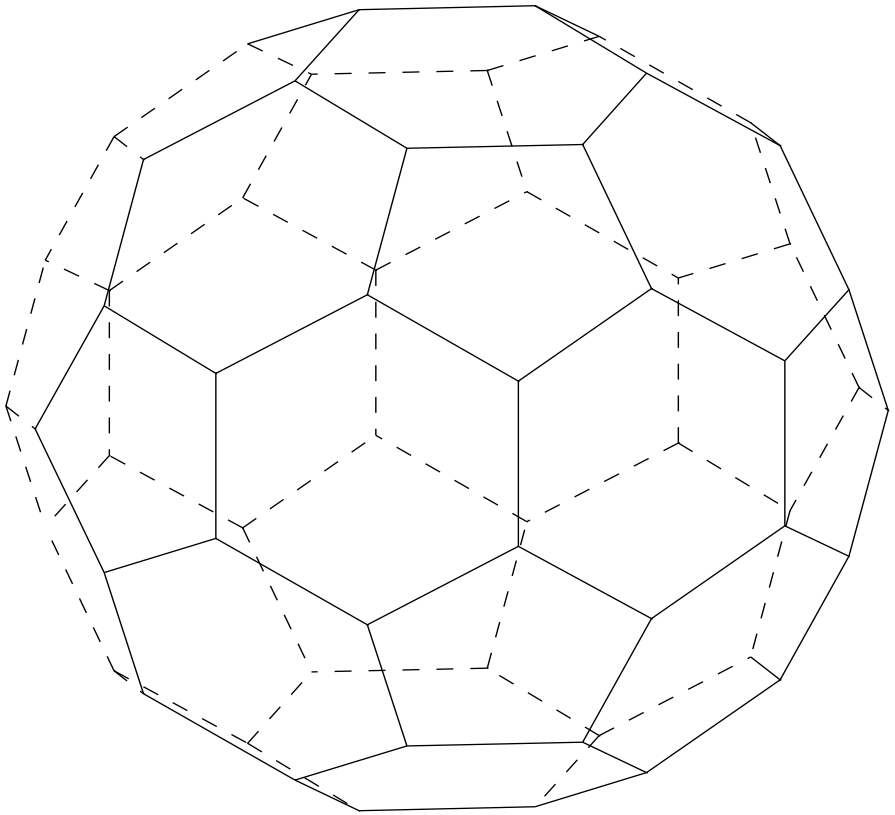
# Properties and examples of FUNTFs

- Frames give redundant signal representation to compensate for hardware errors, to ensure numerical stability, and to minimize the effects of noise.
- Thus, if certain types of noises are known to exist, then the FUNTFs are constructed using this information.
- Orthonormal bases, vertices of Platonic solids, kissing numbers (sphere packing and error correcting codes) are FUNTFs.
- The vector-valued CAZAC – FUNTF problem: Characterize $u : \mathbb{Z}_K \longrightarrow \mathbb{C}^d$ which are CAZAC FUNTFs.

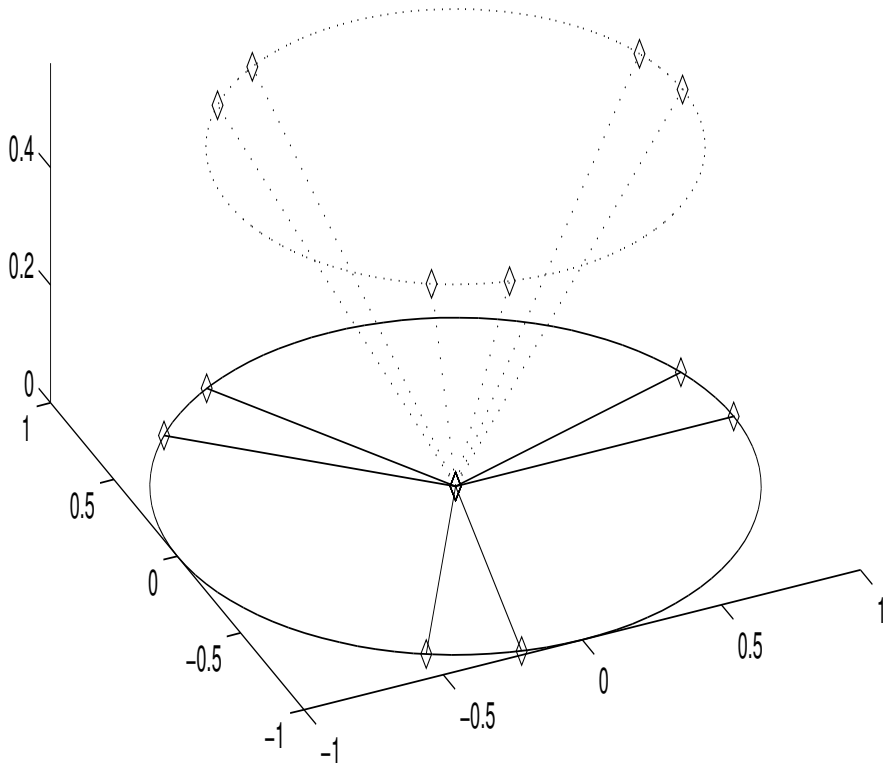
# Recent applications of FUNTFs



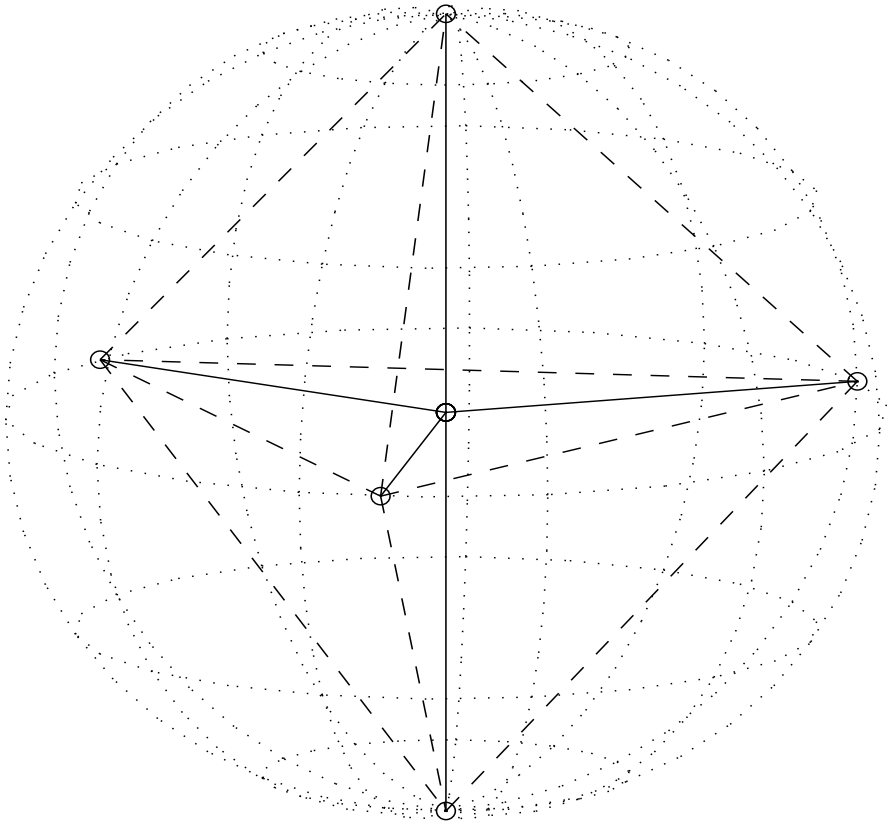
# Recent applications of FUNTFs



# Recent applications of FUNTFs



# Recent applications of FUNTFs



# Recent applications of FUNTFs

discuss kissing numbers

# Recent applications of FUNTFs

- Robust transmission of data over erasure channels such as the internet [Casazza, Goyal, Kelner, Kovačević]
- Multiple antenna code design for wireless communications [Hochwald, Marzetta,T. Richardson, Sweldens, Urbanke]
- Multiple description coding [Goyal, Heath, Kovačević, Strohmer,Vetterli]
- Quantum detection [Bölcskei, Eldar, Forney, Oppenheim, Kebo, B]
- Grassmannian ”min-max” waveforms [Calderbank, Conway, Sloane, et al., Kolesar, B]

# DFT FUNTFs

- $N \times d$ submatrices of the $N \times N$ DFT matrix are FUNTFs for $\mathbb{C}^d$. These play a major role in finite frame $\Sigma\Delta$-quantization.

$$N = 8,\ d = 5 \qquad \frac{1}{\sqrt{5}} \begin{bmatrix} * & * & \cdot & \cdot & * & * & * & \cdot \\ * & * & \cdot & \cdot & * & * & * & \cdot \\ * & * & \cdot & \cdot & * & * & * & \cdot \\ * & * & \cdot & \cdot & * & * & * & \cdot \\ * & * & \cdot & \cdot & * & * & * & \cdot \\ * & * & \cdot & \cdot & * & * & * & \cdot \\ * & * & \cdot & \cdot & * & * & * & \cdot \\ * & * & \cdot & \cdot & * & * & * & \cdot \end{bmatrix}$$

$$x_m = \frac{1}{5}(e^{2\pi i \frac{m}{8}}, e^{2\pi i m \frac{2}{8}}, e^{2\pi i m \frac{5}{8}}, e^{2\pi i m \frac{6}{8}}, e^{2\pi i m \frac{7}{8}})$$

$$m = 1, \ldots, 8.$$

- Sigma-Delta Super Audio CDs - but not all authorities are fans.

# Naimark Theorem

## Definition

Let $H$ be a Hilbert space, $V \subseteq H$ a closed subspace, and

$$V^{\perp} = \{z \in H : \forall y \in V,\ \langle z, y \rangle = 0\}$$

be its orthogonal complement. Then, for every $x \in H$, there is a unique $y \in V$ satisfying

$$\|x - y\| = \min\{\|x - y'\| : y' \in V\},$$

and a unique $z \in V^{\perp}$ such that $x = y + z$.
The map $P_V : H \to V$, $P_V x = y$ is the *orthogonal projection* on $V$.

If $\{v_n\}$ is an orthonormal basis for $V$, then $P_V$ can be expressed as

$$\forall x \in H, \quad P_V x = \sum_n \langle x, v_n \rangle v_n.$$

# Naimark Theorem

Can we make tight frames for $H = \mathbb{F}^d$ ($\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$) with prescribed redundancy?

Yes. Take an $N \times N$ unitary matrix $U$, and choose any $d$ columns of it to form an $N \times d$ matrix $L$. Then, $L^*L = I$, which means, the rows of $L$ form a 1-tight frame for $\mathbb{F}^d$.

How about FUNTFs?

Yes, we shall explain how to generate FUNTFs by using the frame potential.

# Naimark Theorem

If $\{e_n\}_{n=1}^N$ is an A-tight frame for $\mathbb{F}^d$, and $L$ is its Bessel map, then $L^*L = AI$, i.e., the set of the columns of L, $\{c_1, \ldots, c_d\}$ is a $\sqrt{A}$-normed orthogonal set in $\mathbb{F}^N$. Let $V = span\{c_1, \ldots, c_d\}$, and let $\{c_{d+1}, \ldots, c_N\}$ be a $\sqrt{A}$-normed orthogonal basis for $V^\perp$. Then, the matrix

$$U = A^{-1/2}[c_1 \ldots c_N]$$

is a unitary matrix, since its columns give an ONB for $\mathbb{F}^d$. Then, the rows of $U$ also give an ONB for $\mathbb{F}^d$. Let $\widetilde{e}_k$ be the $k$th row of $A^{1/2}U$. Then,

1. $\{\widetilde{e}_k\}$ is a $\sqrt{A}$-normed orthogonal basis for $\mathbb{F}^N$,
2. $e_k = P\widetilde{e}_k$, where $P : \mathbb{F}^N \to \mathbb{F}^d$,

$$P(x[1], \ldots x[N]) = (x[1], \ldots, x[d]).$$

# Naimark Theorem

**Theorem (Naimark)**

Let $H$ be a $d$-dimensional Hilbert space, $\{e_n\}_{n=1}^N$ be an $A$-tight frame for $H$. Then there exists an $N$-dimensional Hilbert space $\widetilde{H}$, and orthogonal $A$-normed set $\{\widetilde{e}_n\}_{n=1}^N \subseteq \widetilde{H}$ such that

$$P_H \widetilde{e}_n = e_n$$

where $P_H$ is the orthogonal projection onto $H$.

# Naimark Theorem

## Proposition

Let $H$ be an $N$-dimensional Hilbert space. Let $\{e_i\}_{i=1}^N$ be an orthonormal basis for $H$. Define the unit normed element

$$\xi = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} e_i \in H$$

and the subspace

$$V = (span\{\xi\})^{\perp}.$$

Denote by $P_V$ the orthogonal projection from $H$ onto $V$. Then,

$$\left\{ \sqrt{\frac{N}{N-1}} P_V e_i \right\}_{i=1}^{N}$$

is an $N$ element FUNTF for the $N-1$ dimensional space $V$.

# Naimark Theorem

*Proof.* By Naimark Theorem, $F = \{P_V e_i\}$ is a tight frame for $V$. We now show that each element of $F$ has the same length. Let $\{b_i\}_{i=1}^{N-1}$ be an ONB for $V$. Then, $\{b_i\}_{i=1}^{N-1} \cup \{\xi\}$ is an ONB for $H$ and

$$e_i = \langle \xi, e_i \rangle \xi + \sum_{j=1}^{N-1} \langle b_j, e_i \rangle b_j = \langle \xi, e_i \rangle \xi + P_V e_i.$$

Note that by the definition of $\xi$, and using the fact that $\{e_i\}_{i=1}^{N}$ is an ONB for $H$, we have,

$$\langle \xi, e_i \rangle = \frac{1}{\sqrt{N}} \sum_{j=1}^{N} \langle e_j, e_i \rangle = \frac{1}{\sqrt{N}}.$$

Hence, combining the above equations gives us,

$$1 = \|e_i\|^2 = \|\langle \xi, e_i \rangle \xi\|^2 + \|P_V e_i\|^2 = \frac{1}{N} + \|P_V e_i\|^2$$

# Naimark Theorem

*Proof. (Continued)* Hence, combining the above equations gives us,

$$1 = \|e_i\|^2 = \|\langle \xi, e_i \rangle \xi\|^2 + \|P_V e_i\|^2 = \frac{1}{N} + \|P_V e_i\|^2$$

so,

$$\|P_V e_i\|^2 = 1 - \frac{1}{N} = \frac{N-1}{N}$$

and taking the square root on both sides gives us,

$$\|P_V e_i\| = \sqrt{\frac{N-1}{N}},$$

so we see that all elements of $F$ have the same norm. By normalizing each element, it now follows that,

$$\left\{ \sqrt{\frac{N}{N-1}} P_V e_i \right\}_{i=1}^{N}$$

is a FUNTF for $V$.

# Naimark Theorem

## Proposition

Let $d < N$, and let $H$ be an $N$-dimensional Hilbert space with orthonormal basis $\{e_i\}_{i=1}^N$. Then, there exist $N - d$ vectors of the form,

$$\xi = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \pm e_i.$$

such that the orthogonal complement of the span of those vectors is a $d$ dimensional vector space $V$, and

$$\left\{ \sqrt{\frac{N}{d}} P_V e_i \right\}_{i=1}^{N}$$

is an $N$ element FUNTF for $V$.

# Naimark Theorem

## Example

We shall construct a 6 element FUNTF for $\mathbb{R}^4$. Let $\{e_i\}_{i=1}^6$ be the standard ONB for $\mathbb{R}^6$, and let

$$\xi_1 = \frac{1}{\sqrt{6}} \sum_{i=1}^{6} e_i = \frac{1}{\sqrt{6}} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \xi_2 = -\frac{1}{\sqrt{6}} \sum_{i=1}^{3} e_i + \frac{1}{\sqrt{6}} \sum_{i=4}^{6} e_i = \frac{1}{\sqrt{6}} \begin{bmatrix} -1 \\ -1 \\ -1 \\ 1 \\ 1 \\ 1 \end{bmatrix}.$$

Let $V = Span\{\xi_1, \xi_2\}^{\perp}$.

# Naimark Theorem

## Example

$$e_1' = P_V e_1 = e_1 - \langle \xi_1, e_1 \rangle \xi_1 - \langle \xi_2, e_1 \rangle \xi_2 = e_1 - \frac{1}{6}\xi_1 + \frac{1}{6}\xi_2 = \begin{bmatrix} 2/3 \\ -1/3 \\ -1/3 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$e_2' = P_V e_2 = e_2 - \langle \xi_1, e_2 \rangle \xi_1 - \langle \xi_2, e_2 \rangle \xi_2 = e_2 - \frac{1}{6}\xi_1 + \frac{1}{6}\xi_2 = \begin{bmatrix} -1/3 \\ 2/3 \\ -1/3 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

enter
plications

# Naimark Theorem

## Example

$$e_3' = P_V e_3 = e_3 - \langle \xi_1, e_3 \rangle \xi_1 - \langle \xi_2, e_3 \rangle \xi_2 = e_3 - \frac{1}{6}\xi_1 + \frac{1}{6}\xi_2 = \begin{bmatrix} -1/3 \\ -1/3 \\ 2/3 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$e_4' = P_V e_4 = e_4 - \langle \xi_1, e_4 \rangle \xi_1 - \langle \xi_2, e_4 \rangle \xi_2 = e_4 - \frac{1}{6}\xi_1 - \frac{1}{6}\xi_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 2/3 \\ -1/3 \\ -1/3 \end{bmatrix}$$

# Naimark Theorem

## Example

$$e_5' = P_V e_5 = e_5 - \langle \xi_1, e_5 \rangle \xi_1 - \langle \xi_2, e_5 \rangle \xi_2 = e_5 - \frac{1}{6}\xi_1 - \frac{1}{6}\xi_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ -1/3 \\ 2/3 \\ -1/3 \end{bmatrix}$$

$$e_6' = P_V e_4 = e_6 - \langle \xi_1, e_5 \rangle \xi_1 - \langle \xi_2, e_6 \rangle \xi_2 = e_6 - \frac{1}{6}\xi_1 - \frac{1}{6}\xi_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ -1/3 \\ -1/3 \\ 2/3 \end{bmatrix}$$

enter
plications

# Naimark Theorem

## Example

By the previous proposition, $F = \{P_V e_i\}_{i=1}^6$ is an equal normed tight frame for $V$. We now rewrite the elements of $F$ in terms of an orthonormal basis for $V$. The set $\{\xi_1, \xi_2, e_1, e_3, e_4, e_5\}$ is linearly independent, and $\xi_1$ and $\xi_2$ are orthogonal. We use Gram Schmit to orthogonalize this set. We obtain,

$$\overline{e}_1 = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ -1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \overline{e}_3 = \frac{1}{\sqrt{6}}\begin{bmatrix} -1 \\ -1 \\ 2 \\ 0 \\ 0 \end{bmatrix}, \overline{e}_4 = \frac{1}{\sqrt{6}}\begin{bmatrix} 0 \\ 0 \\ 0 \\ 2 \\ -1 \end{bmatrix}, \overline{e}_5 = \frac{1}{\sqrt{2}}\begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ -1 \end{bmatrix}.$$

and $\xi_1$, $\xi_2$ remain the same.

# Naimark Theorem

## Example

The set $\mathcal{B} = \{\overline{e}_1, \overline{e}_3, \overline{e}_4, \overline{e}_5\}$ forms an ONB for $V$. We rewrite the frame vectors $F$ in terms of the orthonormal basis $\mathcal{B}$ and obtain,

$$[e_1']_{\mathcal{B}} = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{6}} \\ 0 \\ 0 \end{bmatrix}, [e_2']_{\mathcal{B}} = \begin{bmatrix} -\frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{6}} \\ 0 \\ 0 \end{bmatrix}, [e_3']_{\mathcal{B}} = \begin{bmatrix} 0 \\ \frac{2}{\sqrt{6}} \\ 0 \\ 0 \end{bmatrix},$$

# Naimark Theorem

## Example

The set $\mathcal{B} = \{\overline{e}_1, \overline{e}_3, \overline{e}_4, \overline{e}_5\}$ forms an ONB for $V$. We rewrite the frame vectors $F$ in terms of the orthonormal basis $\mathcal{B}$ and obtain,

$$[e_4']_{\mathcal{B}} = \begin{bmatrix} 0 \\ 0 \\ \frac{2}{\sqrt{6}} \\ 0 \end{bmatrix}, [e_5']_{\mathcal{B}} = \begin{bmatrix} 0 \\ 0 \\ -\frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}, [e_6']_{\mathcal{B}} = \begin{bmatrix} 0 \\ 0 \\ -\frac{1}{\sqrt{6}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix}.$$

$\left\{ \frac{\sqrt{6}}{2}[e_i']_{\mathcal{B}} \right\}_{i=1}^{6}$ is a 6 element FUNTF for $\mathbb{R}^4$.

# The geometry of finite tight frames

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

# Frame force and potential energy

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

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

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

- Coulomb force

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

- Frame force

$$FF(a,b) = \langle a,b\rangle(a-b), \quad f(x) = 1 - x^2/2$$

- Total potential energy for the frame force

$$TFP(\{x_n\}) = \sum_{m=1}^{N}\sum_{n=1}^{N} |\langle x_m, x_n\rangle|^2$$

# Characterization of FUNTFs

**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 $\mathbb{R}^d$.

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 FUNTFs of $N$ elements for $\mathbb{R}^d$.

**Problem**

Find FUNTFs analytically, effectively, computationally.

# Construction of FUNTFs

Suppose we want to construct a FUNTF for $\mathbb{F}^d$.

- If $\mathbb{F} = \mathbb{R}$, Let $(x_1, x_2, \ldots, x_N)$ denote a point in $\mathbb{R}^{Nd}$, where each $x_k \in \mathbb{R}^d$. The solutions of the following constrained minimization problem are FUNTFs.

$$\begin{aligned} &\textit{minimize} && TFP(x_1, x_2, \ldots, x_N) = \sum_{m=1}^{N}\sum_{n=1}^{N} |\langle x_m, x_n\rangle|^2 \qquad (1) \\ &\textit{subject to} && \|x_n\|^2 = 1, \quad \forall n = 1, \ldots, N. \end{aligned}$$

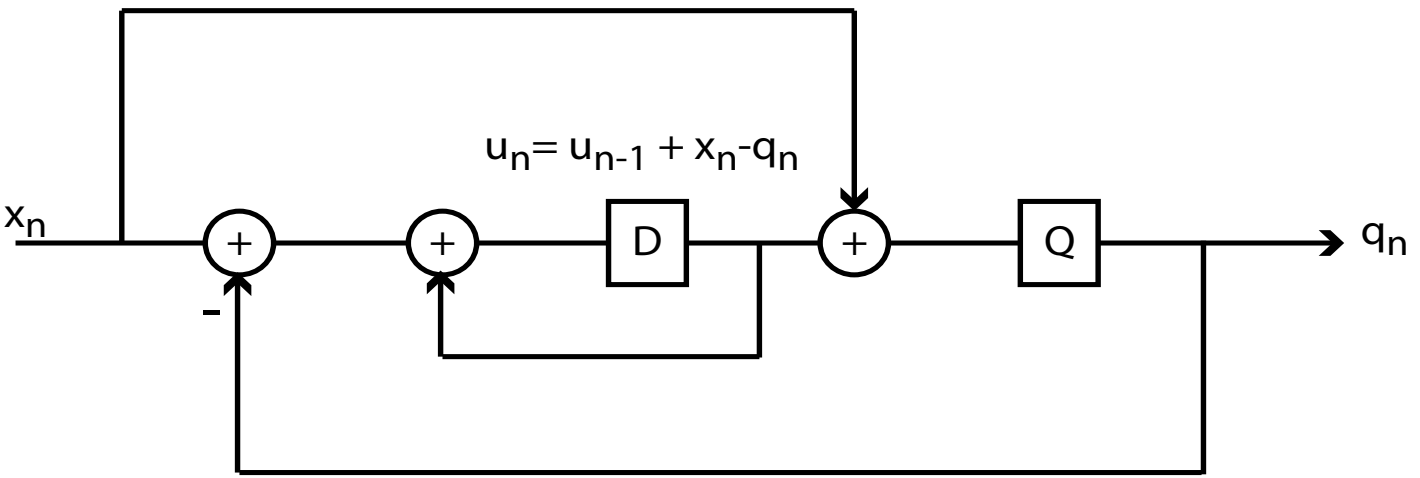
  If we view TFP as a function from $\mathbb{R}^{Nd}$ into $\mathbb{R}$, then it is twice differentiable in each argument, so are the constraints. We can solve this problem numerically, e.g., by using Conjugate Gradient minimization algorithm.

- If $\mathbb{F} = \mathbb{C}$, we let $(Re(x_1), Im(x_1), \ldots, Re(x_N), Im(x_N))$ denote a point in $\mathbb{R}^{2Nd}$, view TFP as a function from $\mathbb{R}^{2Nd}$ into $\mathbb{R}$, and solve (1) as in the real case.

# Outline

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

$u_n = u_{n-1} + x_n - q_n$
$q_n = Q(u_{n-1} + x_n)$




First Order $\Sigma\Delta$

# 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]

$$\forall x \in X, \quad x = \sum x_n e_n, \quad x_n \in \mathbb{K}.$$

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

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

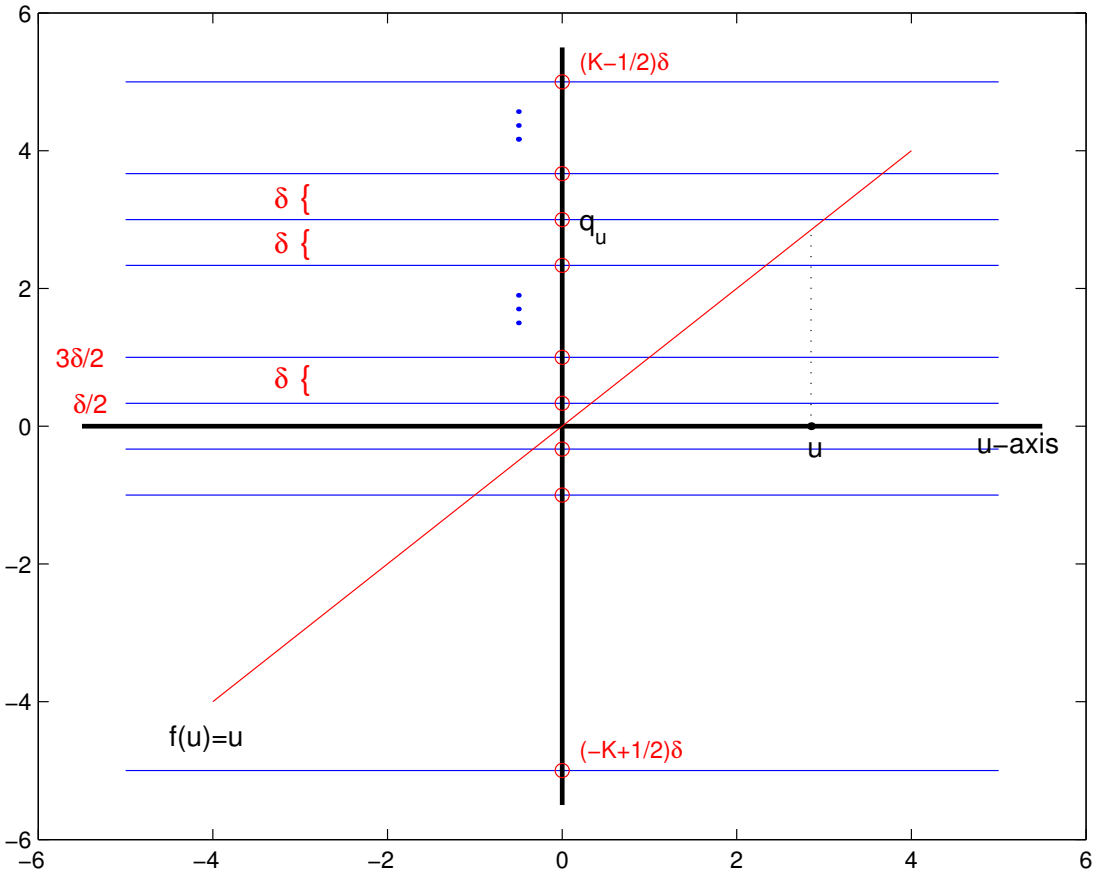
such that $|x_n - q_n|$ and/or $\|x - Qx\|$ small.

**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

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




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

# PCM

Replace $x_n \leftrightarrow q_n = \arg\{\min |x_n - q| : q \in \mathcal{A}_K^\delta\}$. 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

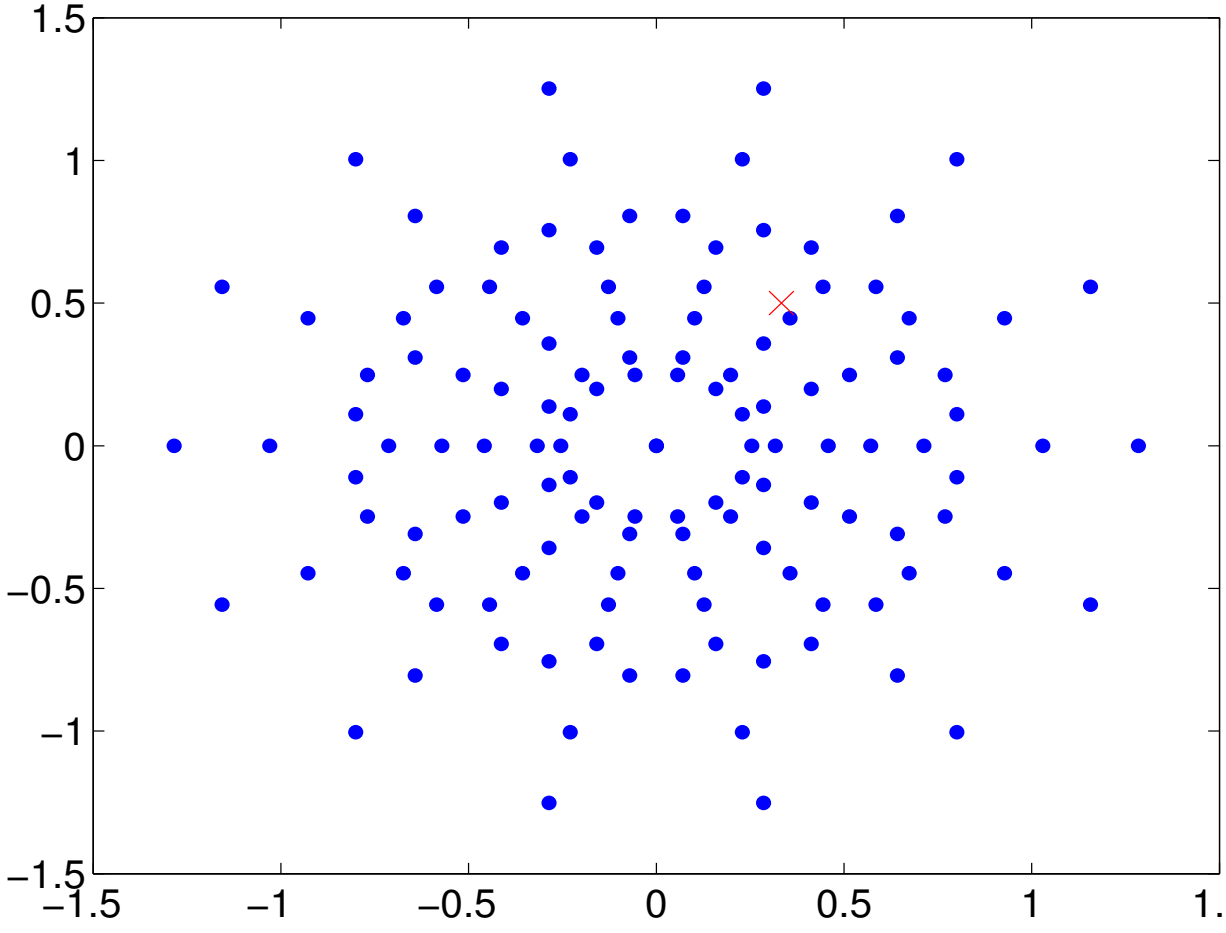
$$\text{MSE} = E\|x - \tilde{x}\|^2 \leq \frac{d}{12A} \delta^2 = \frac{(d\delta)^2}{12N}$$

# $\mathcal{A}_1^2 = \{-1, 1\}$ and $E_7$

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

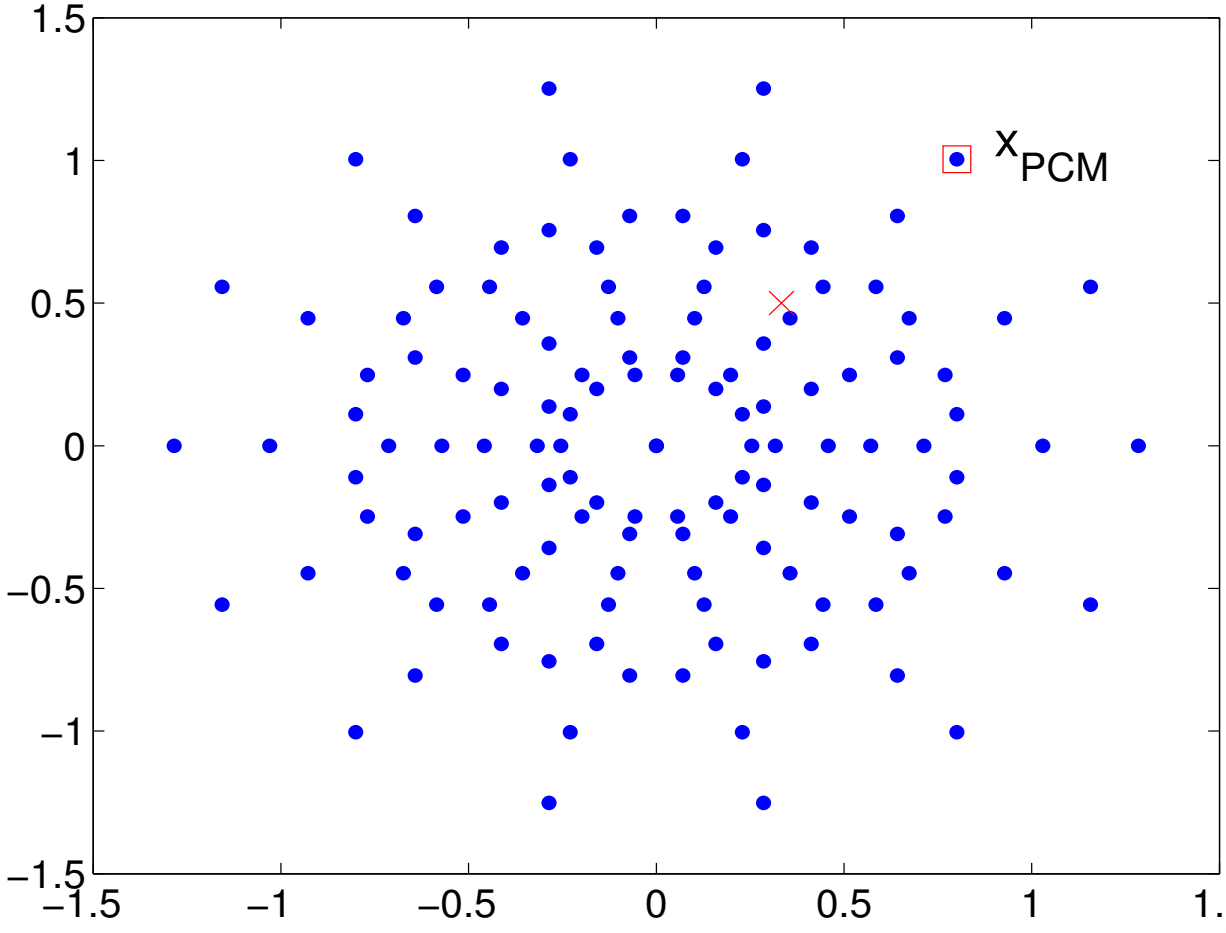
# $\mathcal{A}_1^2 = \{-1, 1\}$ and $E_7$

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



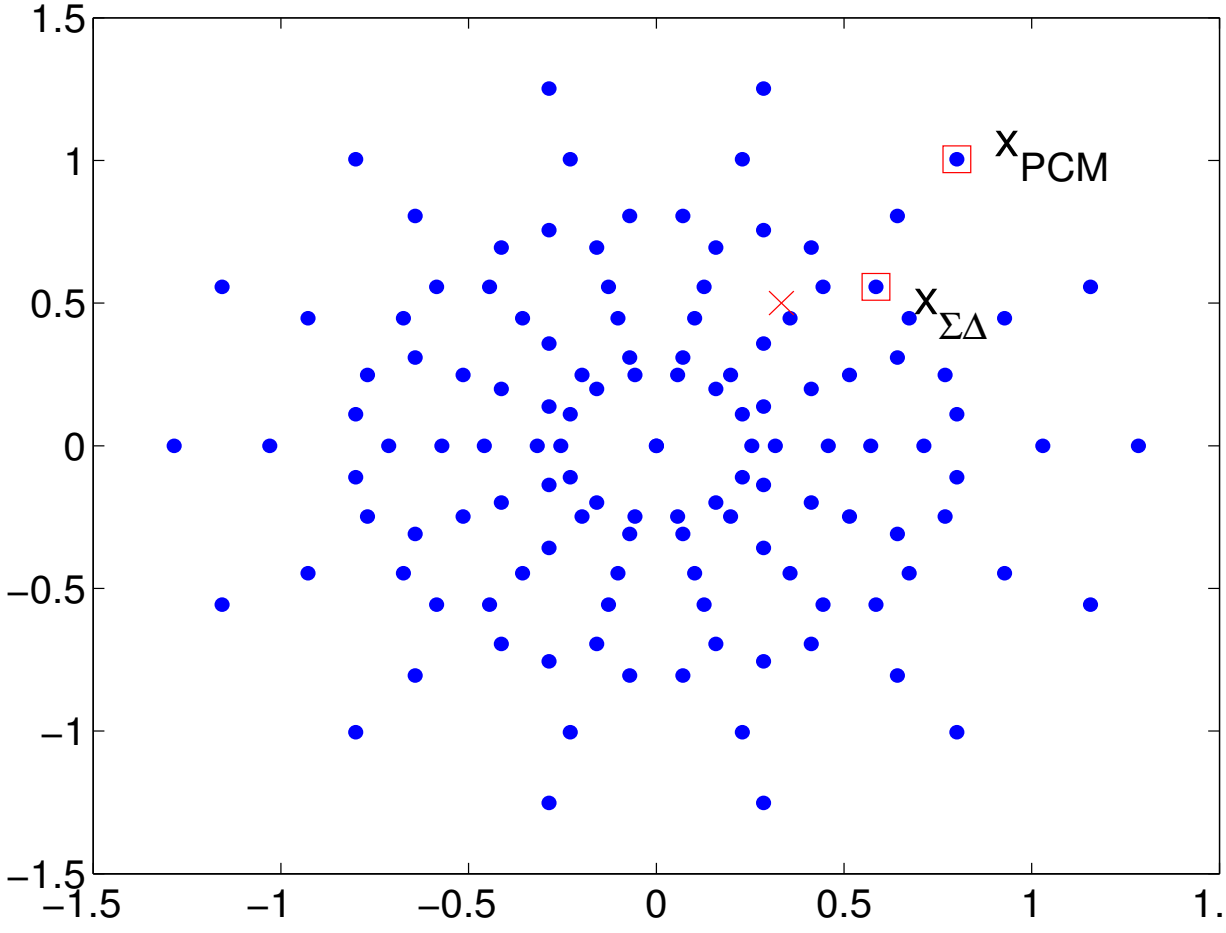
# $\mathcal{A}_1^2 = \{-1, 1\}$ and $E_7$

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



# $\mathcal{A}_1^2 = \{-1, 1\}$ and $E_7$

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



# $\Sigma\Delta$ quantizers for finite frames

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

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.

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

where $u_0 = 0$ and $n = 1, 2, \ldots, N$.

# Sigma-Delta quantization – background

- History from 1950s.
- Treatises of Candy, Temes (1992) and Norsworthy, Schreier, Temes (1997).
- PCM for finite frames and $\Sigma\Delta$ for $PW_\Omega$: Bölcskei, Daubechies, DeVore, Goyal, Güntürk, Kovačevič, Thao, Vetterli.
- Combination of $\Sigma\Delta$ and finite frames: Powell, Yılmaz, and B.
- Subsequent work based on this $\Sigma\Delta$ finite frame theory: Bodman and Paulsen; Boufounos and Oppenheim; Jimenez and Yang Wang; Lammers, Powell, and Yılmaz.
- Genuinely apply it.

# Stability

The following stability result is used to prove error estimates.

**Proposition**

If the frame coefficients $\{x_n\}_{n=1}^N$ satisfy

$$|x_n| \leq (K - 1/2)\delta, \quad n = 1, \cdots, N,$$

then the state sequence $\{u_n\}_{n=0}^N$ generated by the first-order $\Sigma\Delta$ quantizer with alphabet $\mathcal{A}_K^\delta$ satisfies $|u_n| \leq \delta/2, n = 1, \cdots, N.$

- The first-order $\Sigma\Delta$ scheme is equivalent to

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

- Stability results lead to tiling problems for higher order schemes.

# 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

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

# Error estimate

**Theorem**

Let $F = \{e_n\}_{n=1}^N$ be an $A$-FUNTF for $\mathbb{R}^d$. The approximation

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

generated by the first-order $\Sigma\Delta$ quantizer with ordering $p$ and with the quantizer alphabet $\mathcal{A}_K^\delta$ satisfies

$$\|x - \tilde{x}\| \leq \frac{(\sigma(F,p)+1)d}{N} \frac{\delta}{2}.$$

# Harmonic frames

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

## Definition

$H = \mathbb{C}^d$. 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

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

- Harmonic frames are FUNTFs.
- 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),

$$\mathrm{MSE}_{\Sigma\Delta} \leq \frac{C_d}{N^2} \delta^2.$$

- This bound is clearly superior asymptotically to

$$\mathrm{MSE}_{\mathrm{PCM}} = \frac{(d\delta)^2}{12N}.$$

# $\Sigma\Delta$ and "optimal" PCM

**Theorem**

The first order $\Sigma\Delta$ scheme achieves the asymptotically optimal $\mathrm{MSE}_{\mathrm{PCM}}$ for harmonic frames.

The digital encoding

$$\mathrm{MSE}_{\mathrm{PCM}} = \frac{(d\delta)^2}{12N}$$

in PCM format leaves open the possibility that decoding (consistent nonlinear reconstruction, with additional numerical complexity this entails) could lead to

$$\text{“}\mathrm{MSE}_{\mathrm{PCM}}^{\mathrm{opt}}\text{”} \ll O(\frac{1}{N}).$$

Goyal, Vetterli, Thao (1998) proved

$$\text{“}\mathrm{MSE}_{\mathrm{PCM}}^{\mathrm{opt}}\text{”} \sim \frac{\tilde{C}_d}{N^2}\delta^2.$$

# Complex $\Sigma\Delta$ - Alphabet

Let $K \in \mathbb{N}$ and $\delta > 0$. The *midrise* quantization alphabet is

$$\mathcal{A}_K^\delta = \left\{ \left( m + \frac{1}{2} \right) \delta + in\delta : m = -K, \ldots, K-1,\ n = -K, \ldots, K \right\}$$

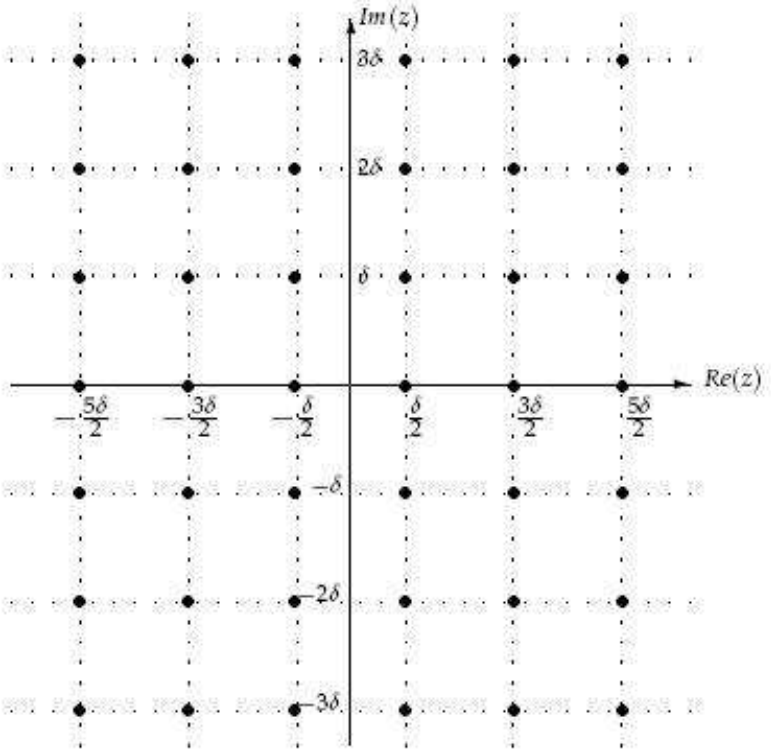



**Figure:** $\mathcal{A}_K^\delta$ for $K = 3\delta$.

# Complex ΣΔ

The *scalar uniform quantizer* associated to $\mathcal{A}_K^\delta$ is

$$Q_\delta(a+ib) = \delta\left(\frac{1}{2} + \left\lfloor \frac{a}{\delta} \right\rfloor + i\left\lfloor \frac{b}{\delta} \right\rfloor\right),$$

where $\lfloor x \rfloor$ is the largest integer smaller than $x$.
For any $z = a + ib$ with $|a| \leq K$ and $|b| \leq K$, $Q$ satisfies

$$|z - Q_\delta(z)| \leq \min_{\zeta \in \mathcal{A}_K^\delta} |z - \zeta|.$$

Let $\{x_n\}_{n=1}^N \subseteq \mathbb{C}$ and let $p$ be a permutation of $\{1, \ldots, N\}$. Analogous to the real case, the first order ΣΔ quantization is defined by the iteration

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

# Complex ΣΔ

The following theorem is analogous to BPY

## Theorem

Let $F = \{e_n\}_{n=1}^N$ be a finite unit norm frame for $\mathbb{C}^d$, let $p$ be a permutation of $\{1, \ldots, N\}$, let $|u_0| \leq \delta/2$, and let $x \in \mathbb{C}^d$ satisfy $\|x\| \leq (K - 1/2)\delta$. The ΣΔ approximation error $\|x - \widetilde{x}\|$ satisfies

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

where $S^{-1}$ is the inverse frame operator. In particular, if $F$ is a FUNTF, then

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

# Complex $\Sigma\Delta$

Let $\{F_N\}$ be a family of FUNTFs, and $p_N$ be a permutation of $\{1,\ldots,N\}$. Then the frame variation $\sigma(F_N, p_N)$ is a function of $N$. If $\sigma(F_N, p_N)$ is bounded, then

$$\|x - \widetilde{x}\| = \mathcal{O}(N^{-1}) \text{ as } N \to \infty.$$

Wang gives an upper bound for the frame variation of frames for $\mathbb{R}^d$, using the results from the Travelling Salesman Problem.

**Theorem YW**

Let $S = \{v_j\}_{j=1}^N \subseteq [-\frac{1}{2}, \frac{1}{2}]^d$ with $d \geq 3$. There exists a permutation $p$ of $\{1,\ldots,N\}$ such that

$$\sum_{j=1}^{N-1} \|v_{p(j)} - v_{p(j+1)}\| \leq 2\sqrt{d+3}N^{1-\frac{1}{d}} - 2\sqrt{d+3}.$$

enter
plications

# Complex ΣΔ

**Theorem**

Let $F = \{e_n\}_{n=1}^N$ be a FUNTF for $\mathbb{R}^d$, $|u_0| \leq \delta/2$, and let $x \in \mathbb{R}^d$ satisfy $\|x\| \leq (K - 1/2)\delta$. Then, there exists a permutation $p$ of $\{1, 2, \ldots, N\}$ such that the approximation error $\|x - \widetilde{x}\|$ satisfies

$$\|x - \widetilde{x}\| \leq \sqrt{2}\delta d \left( (1 - \sqrt{d+3})N^{-1} + \sqrt{d+3}N^{-\frac{1}{d}} \right)$$

This theorem guarantees that

$$\|x - \widetilde{x}\| \leq \mathcal{O}(N^{-\frac{1}{d}}) \text{ as } N \to \infty$$

for FUNTFs for $\mathbb{R}^d$.

# Complex $\Sigma\Delta$ - Algorithms

## Algorithm YW

1. Start with a permutation $p_0$. If this permutation meets the upper bound given in Theorem YW, we are done.
2. If $p$ does not meet the bound, divide $[-\frac{1}{2}, \frac{1}{2}]^d$ into $2^d$ subcubes, pick the nonempty subcubes (say $\{C_k\}$), and find permutations (say $\{p_k\}$) in these smaller subcubes (using *Algorithm 1*).
3. If a $p_k$ does not meet the bound given in Theorem YW, divide $C_k$ further into smaller subcubes. Proceed in this way until the bound in Theorem YW is met in each subcube.
4. Let $p$ be the union of these smaller permutations.

# Complex $\Sigma\Delta$ - Algorithms

**Algorithm 1**

Let $(x_k)_{k=1}^N$ be a FUNTF for $\mathbb{C}^d$.

1. Let $k_1 = 1$, $y_1 = x_{k_1}$, $J_1 = \{k_1\}$,
2. Let $k_n = argmax_{k \notin J_{n-1}} |Re\langle y_{n-1}, x_k \rangle|$, $J_n = J_{n-1} \cup \{k_n\}$,
3. Let $y_n = sign(Re\langle y_{n-1}, x_{k_n} \rangle) x_{k_n}$.

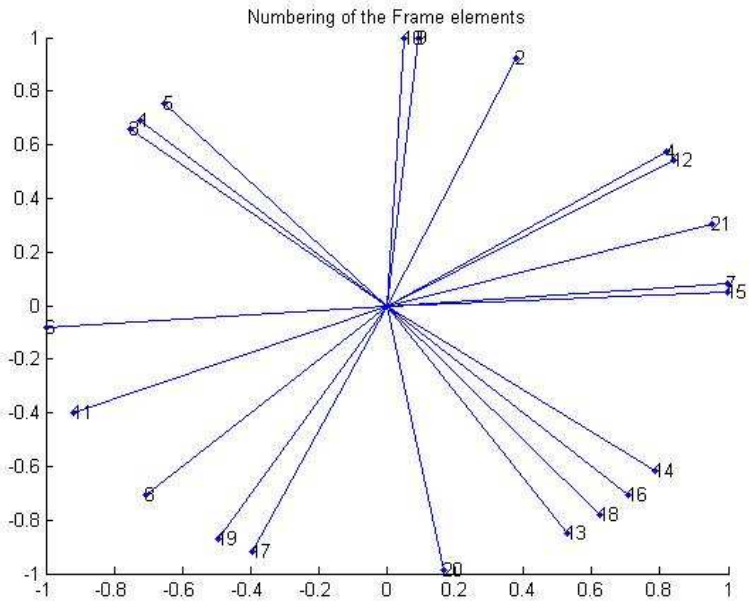
$(y_k)_{k=1}^N$ is unitarily equivalent to $(x_k)_{k=1}^N$, up to a multiplication of frame elements by $\pm 1$.

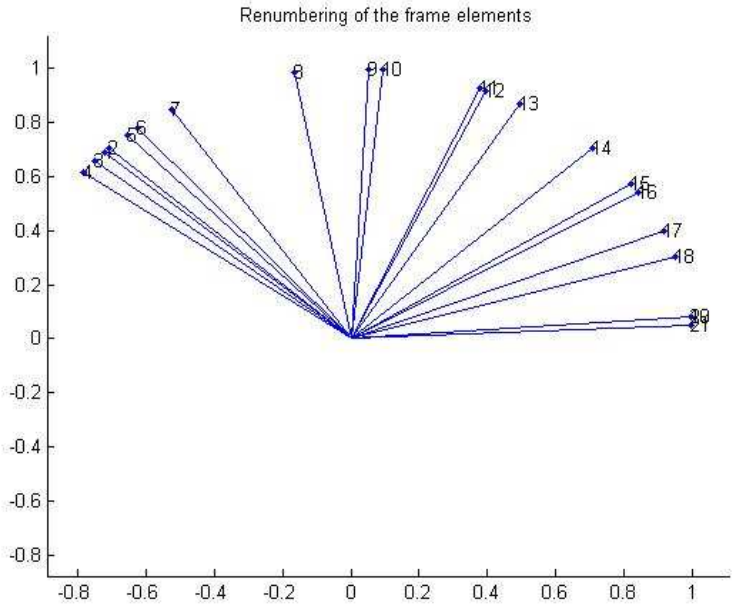
$$\begin{pmatrix}
-0.7228 & 0.69106 \\
0.37772 & 0.92592 \\
-0.75132 & 0.65994 \\
0.81978 & 0.57267 \\
-0.65436 & 0.75618 \\
-0.99649 & -0.083726 \\
0.99664 & 0.081908 \\
-0.70853 & -0.70568 \\
0.093379 & 0.99563 \\
0.05196 & 0.99865 \\
-0.91746 & -0.39784 \\
0.84169 & 0.53996 \\
0.52699 & -0.84987 \\
0.78632 & -0.61782 \\
0.99884 & 0.048212 \\
0.7079 & -0.70631 \\
-0.39328 & -0.91942 \\
0.62502 & -0.78061 \\
-0.49371 & -0.86962 \\
0.1665 & -0.98604 \\
0.95312 & 0.30258
\end{pmatrix}$$

The rows of this matrix form a FUNTF for $\mathbb{R}^2$ with $N = 21$ elements. We calculated this FUNTF by solving

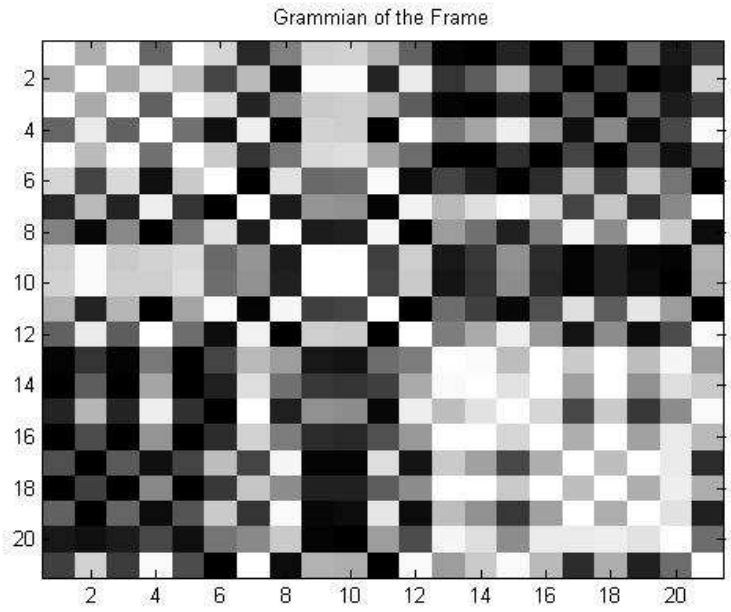
$$\begin{aligned}
&\textit{minimize} && \sum_{m=1}^{N}\sum_{n=1}^{N} |\langle x_m, x_n\rangle|^2 \\
&\textit{subject to} && \|x_n\|^2 = 1, \quad \forall n.
\end{aligned}$$

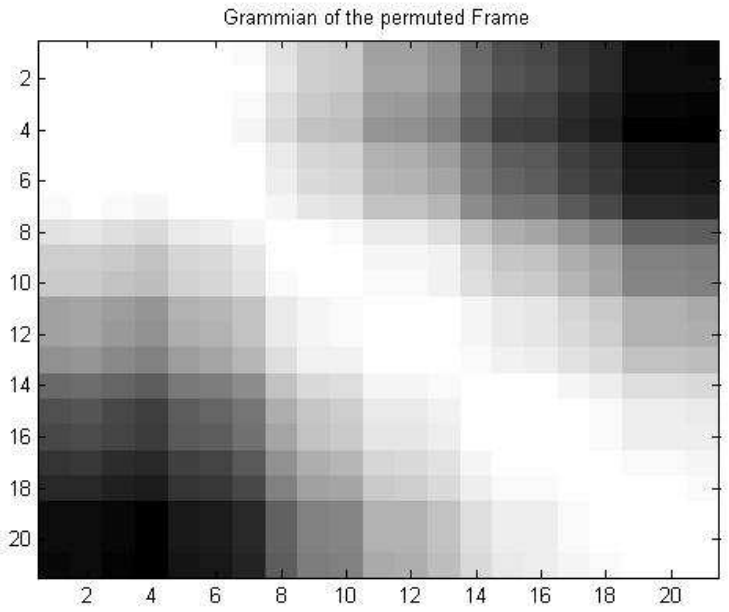
# Complex ΣΔ


Numbering of the Frame elements


Renumbering of the frame elements

# Complex $\Sigma\Delta$




Frame Variation of the original frame $= 24.4779$

Frame Variation of the new frame $= 2.56091$

$2\sqrt{d+3}N^{1-\frac{1}{d}} - 2\sqrt{d+3} = 16.0218$

Norbert Wiener Center for Harmonic Analysis and Applications

# Comparison of 1-bit PCM and 1-bit $\Sigma\Delta$

Let $x \in \mathbb{C}^d$, $||x|| \leq 1$.

## Definition

- $q_{PCM}(x)$ is the sequence to which $x$ is mapped by PCM.
- $q_{\Sigma\Delta}(x)$ is the sequence to which $x$ is mapped by $\Sigma\Delta$.
- $$\mathtt{err}_{PCM}(x) = ||x - \frac{d}{N} L^* q_{PCM}(x)||$$
$$\mathtt{err}_{\Sigma\Delta}(x) = ||x - \frac{d}{N} L^* q_{\Sigma\Delta}(x)||$$

Fickus question: We shall analyze to what extent $\mathtt{err}_{\Sigma\Delta}(x) < \mathtt{err}_{PCM}(x)$ beyond our results with Powell and Yilmaz.

# Comparison of 1-bit PCM and 1-bit $\Sigma\Delta$

**Theorem**

Let $x \in \mathbb{C}^d$ satisfy $0 < \|x\| \leq 1$, and let $F = (e_n)_{n=1}^N$ be a FUNTF for $\mathbb{C}^d$. Then,

$$\mathtt{err}_{PCM}(x) \geq \alpha_F + 1 - \|x\|$$

where

$$\alpha_F := \inf_{\|x\|=1} \frac{d}{N} \sum_{n=1}^{N} |Re(\langle x, e_n \rangle)| + |Im(\langle x, e_n \rangle)| - 1 \geq 0.$$

# Comparison of 1-bit PCM and 1-bit $\Sigma\Delta$

*Proof.* First, note that $Re(Q(a+ib)(a-ib)) = |a| + |b|$. Next,

$$\begin{aligned}
\mathtt{err}_{PCM}(x) &= \|x - \frac{d}{N}\sum_{n=1}^{N} Q(\langle x, e_n\rangle)e_n\| \\
&\geq \|x - \frac{d}{N}\frac{1}{\|x\|^2}\sum_{n=1}^{N} Q(\langle x, e_n\rangle)\langle e_n, x\rangle x\| \\
&= \|x\|\left|\frac{d}{N\|x\|^2}\sum_{n=1}^{N} Q(\langle x, e_n\rangle)\overline{\langle x, e_n\rangle} - 1\right| \\
&\geq \|x\|\left|\frac{d}{N\|x\|^2}\sum_{n=1}^{N} |Re(\langle x, e_n\rangle)| + |Im(\langle x, e_n\rangle)| - 1\right| \\
&= \frac{d}{N}\sum_{n=1}^{N}\frac{|Re(\langle x, e_n\rangle)| + |Im(\langle x, e_n\rangle)|}{\|x\|} - \|x\| \\
&\geq \alpha_F + 1 - \|x\|.
\end{aligned}$$

# Comparison of 1-bit PCM and 1-bit $\Sigma\Delta$

## Theorem (BPY)

Let $||x|| \leq 1$, $(e_n)_{n=1}^N$ be a FUNTF for $\mathbb{R}^d$. Then

$$\mathtt{err}_{\Sigma\Delta}(x) \leq \frac{d}{N}\left(1 + \sum_{n=1}^{N-1} ||e_n - e_{n+1}||\right).$$

## Theorem (BOT)

Let $F = \{e_n\}_{n=1}^N$ be a finite unit norm frame for $\mathbb{C}^d$, let $p$ be a permutation of $\{1, \ldots, N\}$, let $|u_0| \leq \delta/2$, and let $x \in \mathbb{C}^d$ satisfy $\|x\| \leq (K - 1/2)\delta$. The $\Sigma\Delta$ approximation error $\|x - \widetilde{x}\|$ satisfies

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

where $S^{-1}$ is the inverse frame operator for $F$.

enter
lications

# Comparison of 1-bit PCM and 1-bit $\Sigma\Delta$

Let $\{F_N = (e_n^N)_{n=1}^N\}$ be a family of FUNTFs satisfying:

$$\exists M \ \ such \ \ that \ \ \forall N, \ \ \sum_{n=1}^{N-1} ||e_n^N - e_{n+1}^N|| \leq M,$$

e.g., harmonic frames. Then $\mathtt{err}_{\Sigma\Delta}(x) \leq \frac{d(M+1)}{N}$.

On the other hand, we just showed that

$$\mathtt{err}_{PCM}(x) \geq \alpha_{F_N} + 1 - ||x|| \geq 1 - ||x||.$$

# Comparison of 1-bit PCM and 1-bit $\Sigma\Delta$

**Definition**

A frame $F$ is *robust to 1-erasure* if for any $x \in F$, $F\backslash\{x\}$ is a frame.

**Theorem**

FUNTFs are robust to 1-erasure.

*Proof.* Let $F$ be a FUNTF for $\mathbb{C}^d$, and let $F_{-x} = \Phi\backslash\{x\}$ for a fixed $x \in F$. Then, for every $y$,

$$\begin{aligned}\sum_{\phi\in F_{-x}} |\langle y,\phi\rangle|^2 &= \sum_{\phi\in F} |\langle y,\phi\rangle|^2 - |\langle y,x\rangle|^2 \\ &= \frac{N}{d}\|y\|^2 - |\langle y,x\rangle|^2 \\ &\geq \left(\frac{N}{d}-1\right)\|y\|^2.\end{aligned}$$

Therefore, $F_{-x}$ is a frame with $A = \frac{N}{d} - 1$, $B = \frac{N}{d}$.

# Comparison of 1-bit PCM and 1-bit $\Sigma\Delta$

**Lemma**

Let $\{v_k : k = 1, \dots, n\} \subseteq \mathbb{C}^d \backslash \{0\}$, and $\sum_{k=1}^n \|v_k\| = \|\sum_{k=1}^n v_k\|$. Then,

$$\exists w \in \mathbb{C}^d \text{ such that } \forall k = 1, \dots, n,\ v_k = w.$$

*Proof.*

$$\sum_{k,l=1}^n \langle v_k, v_l \rangle = \|\sum_{k=1}^n v_k\|^2 = \left(\sum_{k=1}^n \|v_k\|\right)^2 = \sum_{k,l=1}^n \|v_k\|\|v_l\|,$$

which is possible only if $\langle v_k, v_l \rangle = \|v_k\|\|v_l\|$ for every $k$ and $l$. Then

$$v_k \neq 0 \Rightarrow v_k = v_l \ \ \forall k, l = 1, \dots, n.$$

# Comparison of 1-bit PCM and 1-bit $\Sigma\Delta$

**Lemma**

Let $F = (e_n)_{n=1}^N$ be a FUNTF for $\mathbb{C}^d$ with distinct elements, and with the property

$$e_k \in F \quad \text{and} \quad |\lambda| = 1 \Rightarrow \lambda e_k \notin F.$$

Then $\alpha_F > 0$.

*Proof.* For every $n$ and $\|x\| = 1$, $|\langle x, e_n\rangle| \leq 1$, so

$$|Re(\langle x, e_n\rangle)| \geq |Re(\langle x, e_n\rangle)|^2, \quad |Im(\langle x, e_n\rangle)| \geq |Im(\langle x, e_n\rangle)|^2. \tag{2}$$

$$\begin{aligned} \alpha_F &= \inf_{\|x\|=1} \frac{d}{N} \sum_{n=1}^{N} |Re(\langle x, e_n\rangle)| + |Im(\langle x, e_n\rangle)| - 1 \qquad (3) \\ &= \inf_{\|x\|=1} \frac{d}{N} \sum_{n=1}^{N} |Re(\langle x, e_n\rangle)| - |Re(\langle x, e_n\rangle)|^2 \\ &\quad + |Im(\langle x, e_n\rangle)| - |Im(\langle x, e_n\rangle)|^2 \geq 0. \end{aligned}$$

# Comparison of 1-bit PCM and 1-bit $\Sigma\Delta$

By compactness of $\{x \in \mathbb{C}^d : \|x\| = 1\}$, either $\alpha_F > 0$, or there is an $x_0$, $\|x_0\| = 1$ such that

$$0 = \alpha_F = \sum_{n=1}^{N} |Re(\langle x_0, e_n\rangle)| - |Re(\langle x_0, e_n\rangle)|^2 + |Im(\langle x_0, e_n\rangle)| - |Im(\langle x_0, e_n\rangle)|^2.$$

In the latter case, we must have

$$\forall n = 1, \ldots, N, \quad |Re(\langle x_0, e_n\rangle)| = 0 \text{ or } 1 \text{ and } |Im(\langle x_0, e_n\rangle)| = 0 \text{ or } 1$$

by (2). Then, since

$$1 \geq |\langle x_0, e_n\rangle|^2 = |Re(\langle x_0, e_n\rangle)|^2 + |Im(\langle x_0, e_n\rangle)|^2,$$

either $|Re(\langle x_0, e_n\rangle)| = 0$ or $|Im(\langle x_0, e_n\rangle)| = 0$ or both. Hence,

# Comparison of 1-bit PCM and 1-bit $\Sigma\Delta$

$$|\langle x_0, e_n\rangle| = |Re(\langle x_0, e_n\rangle)| + |Im(\langle x_0, e_n\rangle)| \tag{4}$$

Then, (3) and (4) imply

$$\sum_{n=1}^{N} |\langle x, e_n\rangle| = \frac{N}{d}\|x_0\| = \|\sum_{n=1}^{N}\langle x_0, e_n\rangle e_n\|.$$

Then, by the previous Lemma, there is a $w$ such that $\langle x_0, e_n\rangle e_n = w$ if $\langle x_0, e_n\rangle \neq 0$. Then, all $e_n$ that are not orthogonal to $x_0$ are equal up to a multiplication by a $\lambda, |\lambda| = 1$. But, by the hypothesis, there is only one such frame element nonorthogonal to $x_0$. Erasing this element, remaining vectors would not span $\mathbb{C}^d$, i.e., $F$ would not be robust. Contradiction.
Therefore, $\alpha_F > 0$.

# Comparison of 1-bit PCM and 1-bit $\Sigma\Delta$

## Theorem 1

Let $\{F_N = (e_n^N)_{n=1}^N\}$ be a family of FUNTFs for $\mathbb{C}^d$ such that

$$\exists M > 0 \quad \forall N > 0 \quad \sum_{n=1}^{N-1} \|e_n^N - e_{n+1}^N\| \leq M.$$

Then,

$$\forall \varepsilon > 0 \quad \exists N_0 > 0 \quad \forall N \geq N_0 \quad \mathtt{err}_{\Sigma\Delta}(x) \leq \mathtt{err}_{PCM}(x)$$

for every $0 < \|x\| \leq 1 - \varepsilon$. (each $\mathtt{err}$ depends on $N$).

*Proof.* $\mathtt{err}_{\Sigma\Delta}(x) \leq \frac{d(1+M)}{N}$ for any $N$ and $||x|| \leq 1$. Then,

$$\forall \varepsilon > 0 \;\; \exists N_0 \;\; \textit{such that} \;\; \forall\, 0 < ||x|| \leq 1 - \varepsilon \;\; \textit{and} \;\; \forall N \geq N_0,$$

$$\mathtt{err}_{\Sigma\Delta}(x) \leq \frac{d(1+M)}{N} \leq \varepsilon \leq 1 - ||x|| \leq \mathtt{err}_{PCM}(x).$$

# Comparison of 1-bit PCM and 1-bit $\Sigma\Delta$

Let $F_N = (e_n^N)_{n=1}^N, N = d, d+1, \ldots$ be a family of FUNTFs. If there is a positive uniform lower bound for $(\alpha_{F_N})$, then we can replace the conclusion of Theorem 1 by the assertion that

$$\exists N_0 > 0,\ \forall N \geq N_0 \text{ and } \forall 0 < \|x\| \leq 1\ \mathtt{err}_{\Sigma\Delta}(x) \leq \mathtt{err}_{PCM}(x).$$

The families $(F_N)$ such that $\alpha_{F_N} \to 0$ are the pathological cases that we describe in the following Theorem.

# Comparison of 1-bit PCM and 1-bit $\Sigma\Delta$

**Theorem**

Let $F_N = (e_n^N)_{n=1}^N, N = d, d+1, \ldots$ be a family of FUNTFs for $\mathbb{C}^d$. If $\alpha_{F_N} \to 0$, then there is $\|x_0\| = 1$ such that

$$\forall \varepsilon > 0, \quad \lim_{N\to\infty} \frac{card\{n \in \{1,\ldots,N\} : |\langle x_0, e_n^N\rangle| - |\langle x_0, e_n^N\rangle|^2 \le \varepsilon\}}{N} = 1.$$

# Comparison of 1-bit PCM and 1-bit $\Sigma\Delta$

*Proof.*

$$\alpha_{F_N} \geq \inf_{\|x\|=1} \frac{d}{N} \sum_{n=1}^{N} |\langle x, e_n^N \rangle| - 1 > 0.$$

Let $x_N$, $\|x_N\| = 1$ be the point where $\sum_{n=1}^{N} |\langle x, e_n^N \rangle|$ attains its minimum. Then,

$$\alpha_{F_N} \to 0 \Rightarrow \lim_{N\to\infty} \frac{d}{N} \sum_{n=1}^{N} |\langle x_N, e_n^N \rangle| = 1.$$

Note that

$$\left| \frac{d}{N} \sum_{n=1}^{N} |\langle x_N, e_n^N \rangle| - \frac{d}{N} \sum_{n=1}^{N} |\langle x_0, e_n^N \rangle| \right| \leq d\|x_N - x_0\|. \qquad (5)$$

# Comparison of 1-bit PCM and 1-bit $\Sigma\Delta$

*Proof.*
By compactness, $(x_N)$ has a convergent subsequence. Without loss of generality, assume that $\lim_{N\to\infty} x_N = x_0$. Letting $N \to \infty$ in (5), we obtain

$$\lim_{N\to\infty} \frac{d}{N} \sum_{n=1}^{N} |\langle x_0, e_n^N \rangle| = 1.$$

Let

$$\begin{aligned} A_N^\varepsilon &= \{n = 1, \ldots, N : |\langle x_0, e_n^N \rangle| - |\langle x_0, e_n^N \rangle|^2 \leq \varepsilon\}, \\ B_N^\varepsilon &= \{n = 1, \ldots, N : |\langle x_0, e_n^N \rangle| - |\langle x_0, e_n^N \rangle|^2 > \varepsilon\}. \end{aligned}$$

Then,

$$\frac{d\varepsilon\ card B_N^\varepsilon}{N} \leq \frac{d}{N} \sum_{n=1}^{N} |\langle x_0, e_n^N \rangle| - 1 \Rightarrow \lim_{N\to\infty} \frac{card B_N^\varepsilon}{N} = 0 \Rightarrow \lim_{N\to\infty} \frac{card A_N^\varepsilon}{N} = 1.$$

# Comparison of 1-bit PCM and 1-bit $\Sigma\Delta$

## Definition

A function $e : [a, b] \to \mathbb{C}^d$ is of *bounded variation (BV)* if there is a $K > 0$ such that for every $a \leq t_1 < t_2 < \cdots < t_N \leq b$,

$$\sum_{n=1}^{N-1} \|e(t_n) - e(t_{n+1})\| \leq K.$$

The smallest such $K$ is denoted by $|e|_{BV}$, and defines a seminorm for the space of BV functions.

# Comparison of 1-bit PCM and 1-bit $\Sigma\Delta$

**Theorem 2**

Let $e : [0,1] \to \{x \in \mathbb{C}^d : \|x\| = 1\}$ be continuous function of bounded variation such that $F_N = (e(n/N))_{n=1}^N$ is a FUNTF for $\mathbb{C}^d$ for every $N$. Then,

$$\exists N_0 > 0 \text{ such that } \forall N \geq N_0 \text{ and } \forall 0 < \|x\| \leq 1$$

$$\mathtt{err}_{\Sigma\Delta}(x) \leq \mathtt{err}_{PCM}(x).$$

Moreover, a lower bound for $N_0$ is $d(1 + |e|_{BV})/(\sqrt{d} - 1)$.

# Comparison of 1-bit PCM and 1-bit $\Sigma\Delta$

*Proof.*
Let $e_n^N = e(n/N)$. If $\|x\| = 1$, then $\frac{d}{N}\sum_{n=1}^{N} |Re(\langle x, e_n^N\rangle)| + |Im(\langle x, e_n^N\rangle)| - 1 \geq \alpha_{F_N} > 0$. Second,

$$\begin{aligned}
&\lim_{N\to\infty} \frac{d}{N}\sum_{n=1}^{N} |Re(\langle x, e_n^N\rangle)| + |Im(\langle x, e_n^N\rangle)| - 1 \\
= &\lim_{N\to\infty} \left(\frac{d}{N}\sum_{n=1}^{N} |Re(\langle x, e(n/N)\rangle)| + |Im(\langle x, e(n/N)\rangle)|\right) \\
&- \lim_{N\to\infty} \left(\frac{d}{N}\sum_{n=1}^{N} |Re(\langle x, e(n/N)\rangle)|^2 + |Im(\langle x, e(n/N)\rangle)|^2\right) \\
= &\; d\int_0^1 \left(|Re(\langle x, e(t)\rangle)| + |Im(\langle x, e(t)\rangle)|\right. \\
&\left. -|Re(\langle x, e(t)\rangle)|^2 - |Im(\langle x, e(t)\rangle)|^2\right) dt. \qquad (6)
\end{aligned}$$

# Comparison of 1-bit PCM and 1-bit $\Sigma\Delta$

*Proof.*
The integrand in (6) cannot be equal to zero for every $t$, since otherwise we would have that

$$\forall t, \quad |Re(\langle x, e(t)\rangle)| = 0 \text{ or } 1 \text{ and } |Im(\langle x, e(t)\rangle)| = 0 \text{ or } 1.$$

But, because

$$1 \geq |\langle x, e(t)\rangle|^2 = |Re(\langle x, e(t)\rangle)|^2 + |Im(\langle x, e(t)\rangle)|^2,$$

either $|Re(\langle x, e(t)\rangle)| = 0$ or $|Im(\langle x, e(t)\rangle)| = 0$ or both. Hence,

$$|\langle x, e(t)\rangle| = 0 \text{ or } 1.$$

Since $x \neq 0$, there should exist a $t^*$ such that $|\langle x, e(t^*)\rangle| = 1$ which implies that there is a $|\lambda_0| = 1$ such that $x = \lambda_0 e(t^*)$, and that $\langle x, e(t)\rangle = 0$ for every $t$ such that $e(t) \neq \lambda e(t^*)$ for some $|\lambda| = 1$. But this contradicts the continuity of $e$. Therefore, the integrand in (6) is not zero at every point.

# Comparison of 1-bit PCM and 1-bit $\Sigma\Delta$

*Proof.*
Moreover, since the integrand is continuous,

$$\int_0^1 |Re(\langle x, e(t)\rangle)|+|Im(\langle x, e(t)\rangle)|-|Re(\langle x, e(t)\rangle)|^2-|Im(\langle x, e(t)\rangle)|^2\, dt > 0$$

for each $x$, $\|x\| = 1$. Then, by the compactness of the unit sphere,

$$\alpha := d \inf_{\|x\|=1} \int_0^1 |Re(\langle x, e(t)\rangle)|+|Im(\langle x, e(t)\rangle)|-|Re(\langle x, e(t)\rangle)|^2-|Im(\langle x, e(t)\rangle$$

Clearly, $\lim_{N\to\infty} \alpha_{F_N} = \alpha$. Therefore, we conclude that

$$\exists \beta > 0 \quad \text{such that} \quad \mathtt{err}_{PCM}(x) \geq \alpha_{F_N} + 1 - \|x\| \geq \beta + 1 - \|x\|$$

for every $0 < \|x\| \leq 1$, and for every $N > 0$.

# Comparison of 1-bit PCM and 1-bit $\Sigma\Delta$

*Proof.*
Third, $\sum_{n=1}^{N} \|e_n - e_{n+1}\| \leq |e|_{BV} =: M$. Therefore, by Theorem (BOT), we have

$$\mathtt{err}_{\Sigma\Delta}(x) \leq \frac{d}{N}(1+M)$$

for every $N$. Then, for $\varepsilon = \beta$,

$$\exists N_0 > 0 \text{ such that } \forall N \geq N_0 \text{ and } \forall 0 < \|x\| \leq 1$$

$$\mathtt{err}_{\Sigma\Delta}(x) \leq \frac{d}{N}(1+M) \leq \beta \leq \alpha_{F_N} + 1 - \|x\| \leq \mathtt{err}_{PCM}(x).$$

# Comparison of 1-bit PCM and 1-bit $\Sigma\Delta$

**Example (Roots of unity frames for $\mathbb{R}^2$)**

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

Here, $e(t) = (\cos(2\pi t), \sin(2\pi t))$,
$M = |e|_{BV} = 2\pi$, $lim\alpha_{F_N} = 2/\pi$.

**Example (Real Harmonic Frames for $\mathbb{R}^{2k}$)**

$e_n^N = \frac{1}{\sqrt{k}}(\cos(2\pi n/N), \sin(2\pi n/N), \ldots, \cos(2\pi kn/N), \sin(2\pi kn/N))$.

In this case, $e(t) = \frac{1}{\sqrt{k}}(\cos(2\pi t), \sin(2\pi t), \ldots, \cos(2\pi kt), \sin(2\pi kt))$,

$M = |e|_{BV} = 2\pi\sqrt{\frac{1}{d}\sum_{k=1}^{d} k^2}$.

# Comparison in multibit case

## Definition

For an integer $b \geq 1$, let $\delta = 2^{1-b}$ and $K = \delta^{-1}$. The *midrise quantization alphabet* is

$$\mathcal{A}_\delta = \{(m + \frac{1}{2})\delta + in\delta : m = -K, \ldots, K-1, \quad n = -K, \ldots, K\},$$

and the associated *scalar uniform quantizer with step size* $\delta$ is given by

$$Q_\delta(a + ib) = \delta\left(\frac{1}{2} + \left\lfloor \frac{a}{\delta} \right\rfloor + i\left\lfloor \frac{b}{\delta} \right\rfloor\right).$$

- Multibit PCM uses the simple rule $q_n = Q(\langle x, e_n \rangle)$
- Multibit 1st order $\Sigma\Delta$ uses the iterative sequence:

$$\begin{aligned} u_n &= u_{n-1} + \langle x, e_n \rangle - q_n \\ q_n &= Q(\langle x, e_n \rangle + u_{n-1}) \end{aligned}$$

# Comparison in multibit case

## Theorem

Let $\{F_N = (e_n^N)_{n=1}^N\}$ be a family of FUNTFs for $\mathbb{C}^d$ such that

$$\exists M > 0 \ \textit{such that} \ \forall N, \quad \sum_{n=1}^{N-1} ||e_n^N - e_{n+1}^N|| \leq M.$$

Then,

$$\forall \varepsilon > 0 \ \exists N_0 \ \textit{such that} \ \forall N \geq N_0 \ \textit{and} \ \forall\, 0 < ||x|| \leq \frac{\delta}{2} - \varepsilon$$

$$\mathtt{err}_{\Sigma\Delta}(x) \leq \mathtt{err}_{PCM}(x).$$

at the same bit rate.

*Proof.*
$0 < \|x\| \leq \delta/2 \Rightarrow \forall n, \ |\langle x, e_n\rangle| \leq \delta/2 \Rightarrow Q(\langle x, e_n\rangle) = \pm\delta/2 \pm i\delta$. It is not hard to show that 1-bit quantized coefficients of $\delta^{-1}x$ are $\delta^{-1}Q_\delta(\langle x, e_n\rangle)$. The result follows from the 1-bit case Theorem 1.

# Comparison in multibit case

**Theorem**

Let $e : [0,1] \to \{x : ||x|| = 1\}$ be continuous and of BV such that $F_N = (e(n/N))_{n=1}^N$ is a FUNTF for $\mathbb{R}^d$ for every $N$. Then,

$$\exists N_0 \ \textit{such that} \ \forall\, 0 < ||x|| \leq \frac{\delta}{2} \ \textit{and} \ \forall N \geq N_0$$

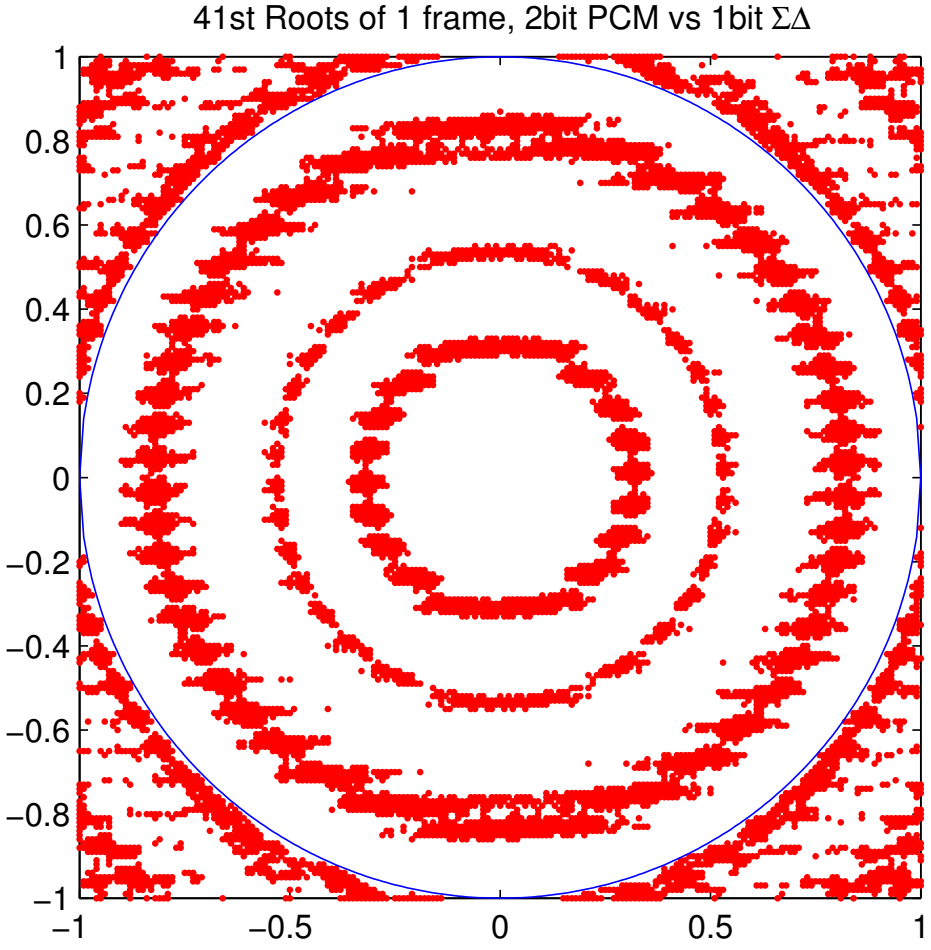
$$\mathtt{err}_{\Sigma\Delta}(x) \leq \mathtt{err}_{PCM}(x).$$

*Proof.*

$$0 < \|x\| \leq \delta/2 \Rightarrow \forall n,\ |\langle x, e_n\rangle| \leq \delta/2 \Rightarrow Q(\langle x, e_n\rangle) = \pm\delta/2 \pm i\delta.$$

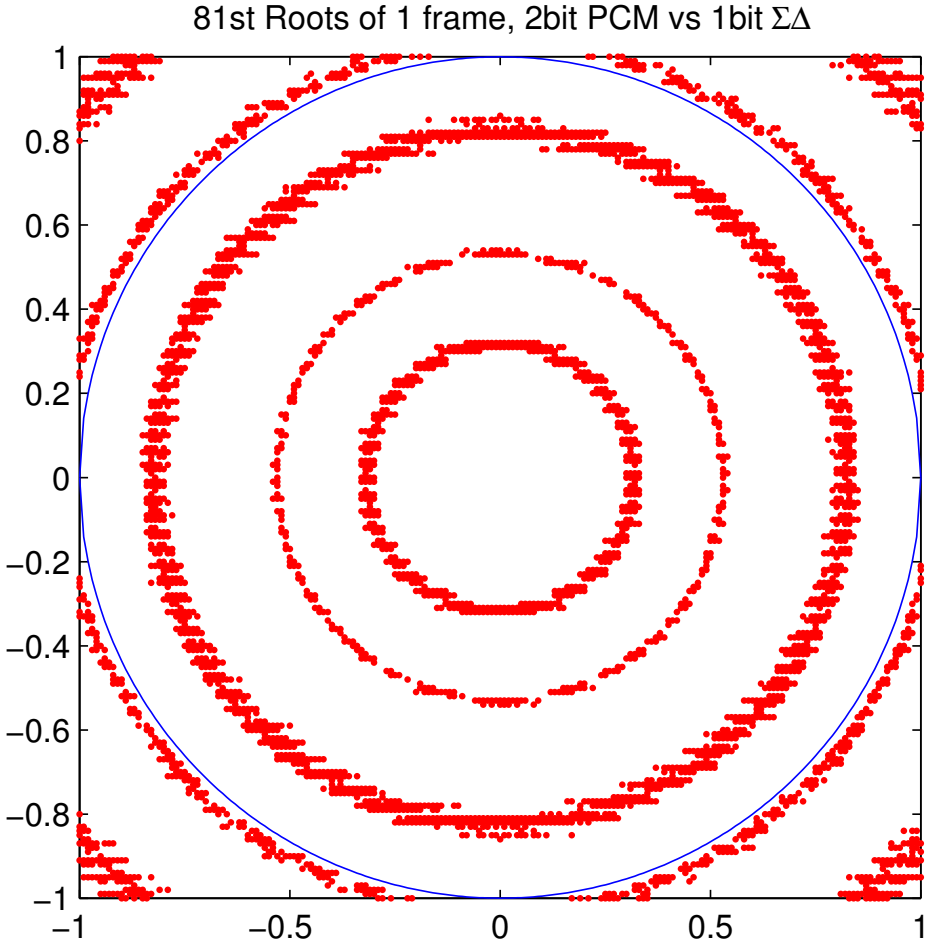
It is not hard to show that 1-bit quantized coefficients of $\delta^{-1}x$ are $\delta^{-1}Q_\delta(\langle x, e_n\rangle)$. The result follows from the 1-bit case Theorem 2.

# Comparison of 2-bit PCM and 1-bit ΣΔ



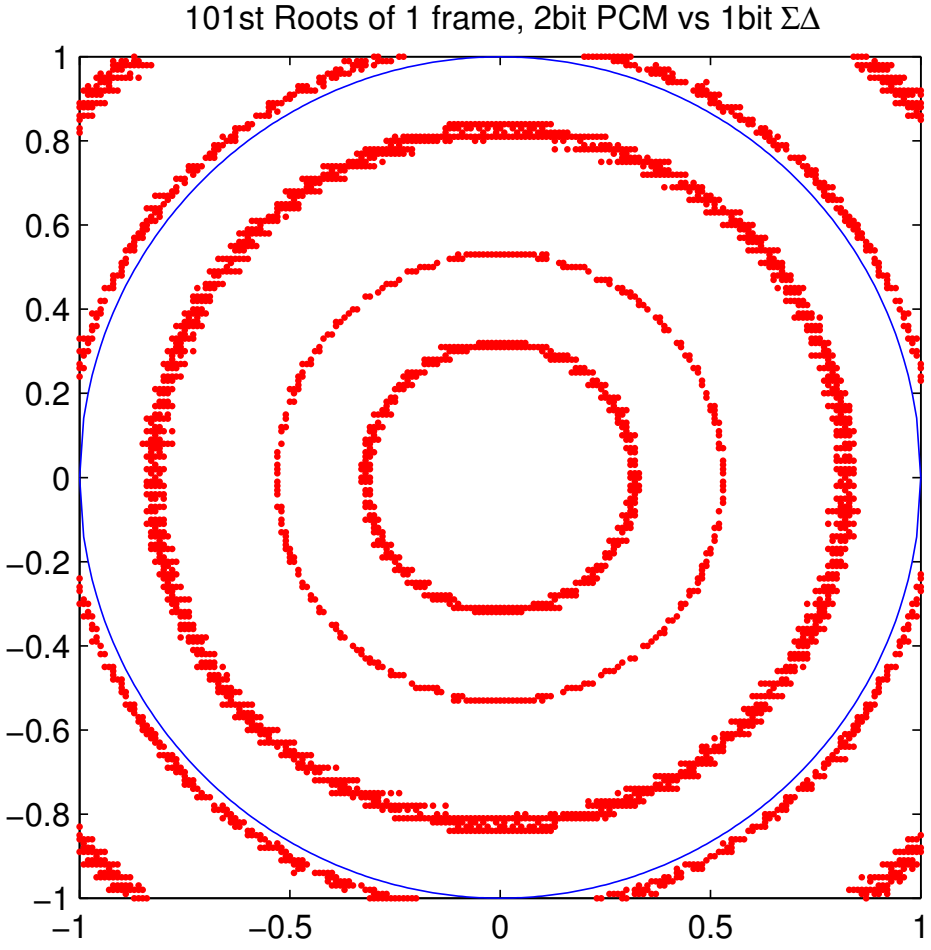

Red: $\mathtt{err}_{PCM}(x) < \mathtt{err}_{\Sigma\Delta}(x)$, Green: $\mathtt{err}_{PCM}(x) = \mathtt{err}_{\Sigma\Delta}(x)$.

# Comparison of 2-bit PCM and 1-bit ΣΔ



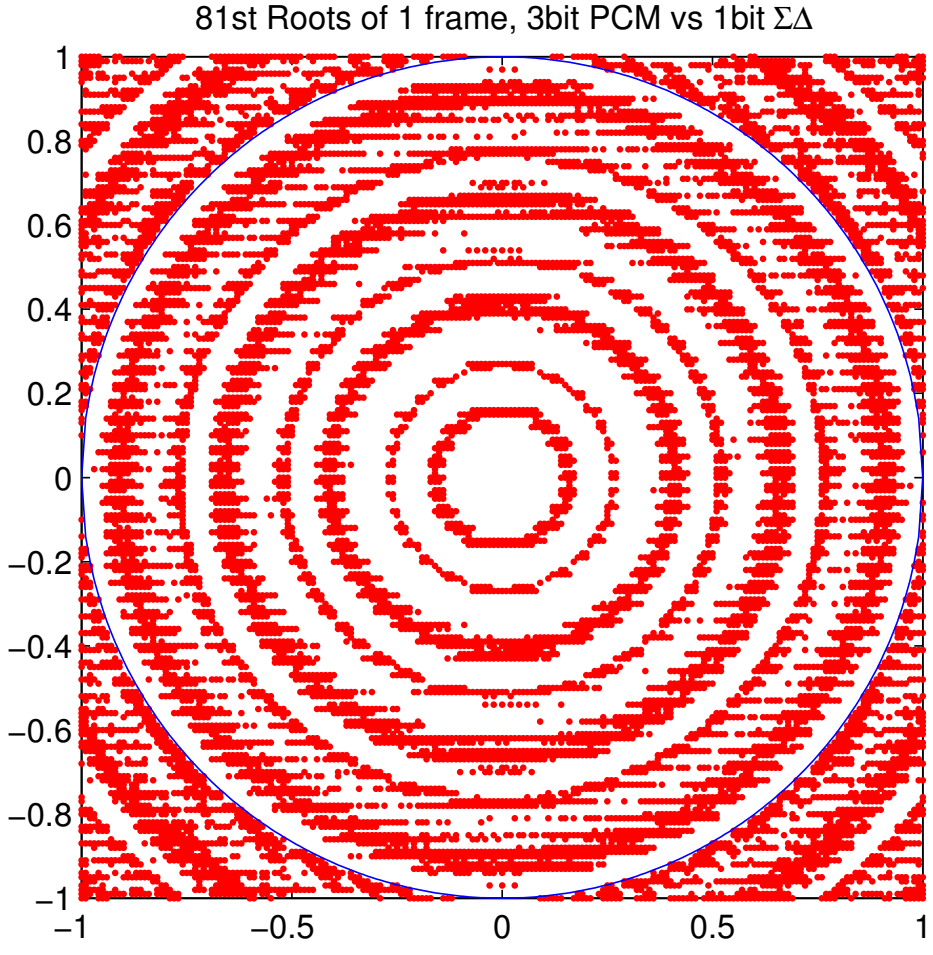

Red: $\mathrm{err}_{PCM}(x) < \mathrm{err}_{\Sigma\Delta}(x)$, Green: $\mathrm{err}_{PCM}(x) = \mathrm{err}_{\Sigma\Delta}(x)$.

# Comparison of 2-bit PCM and 1-bit ΣΔ



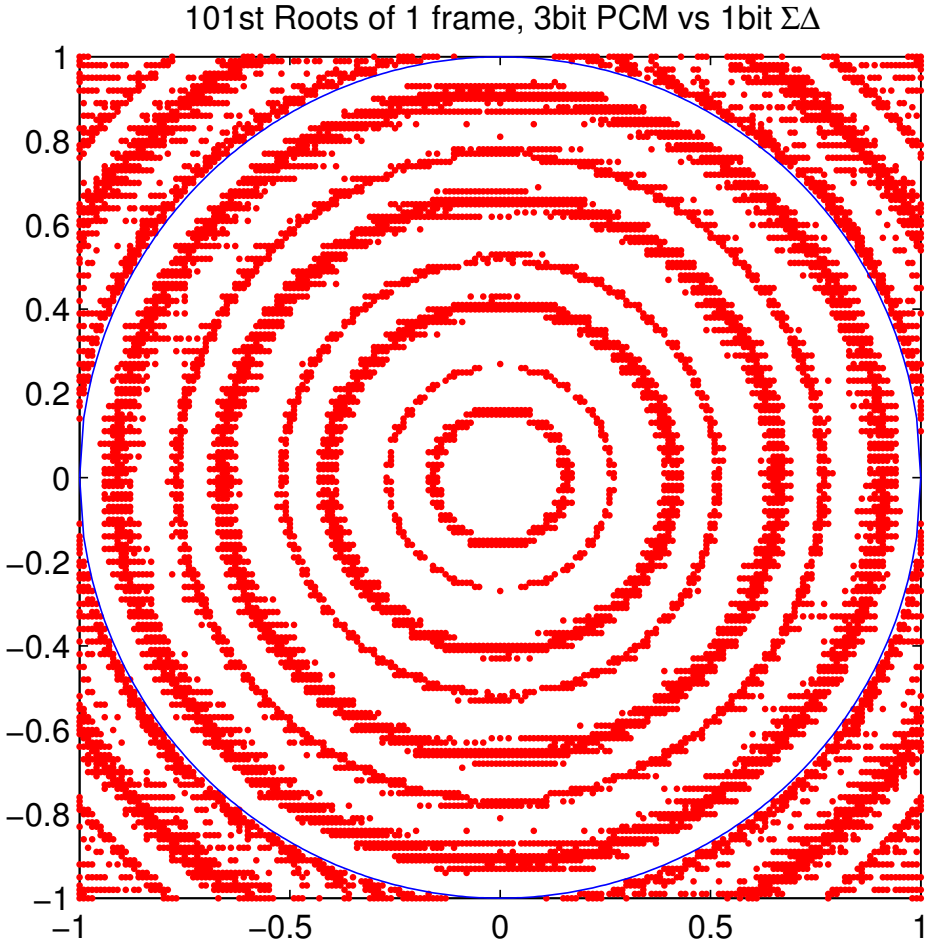

Red: $\mathrm{err}_{PCM}(x) < \mathrm{err}_{\Sigma\Delta}(x)$, Green: $\mathrm{err}_{PCM}(x) = \mathrm{err}_{\Sigma\Delta}(x)$.

# Comparison of 3-bit PCM and 1-bit $\Sigma\Delta$



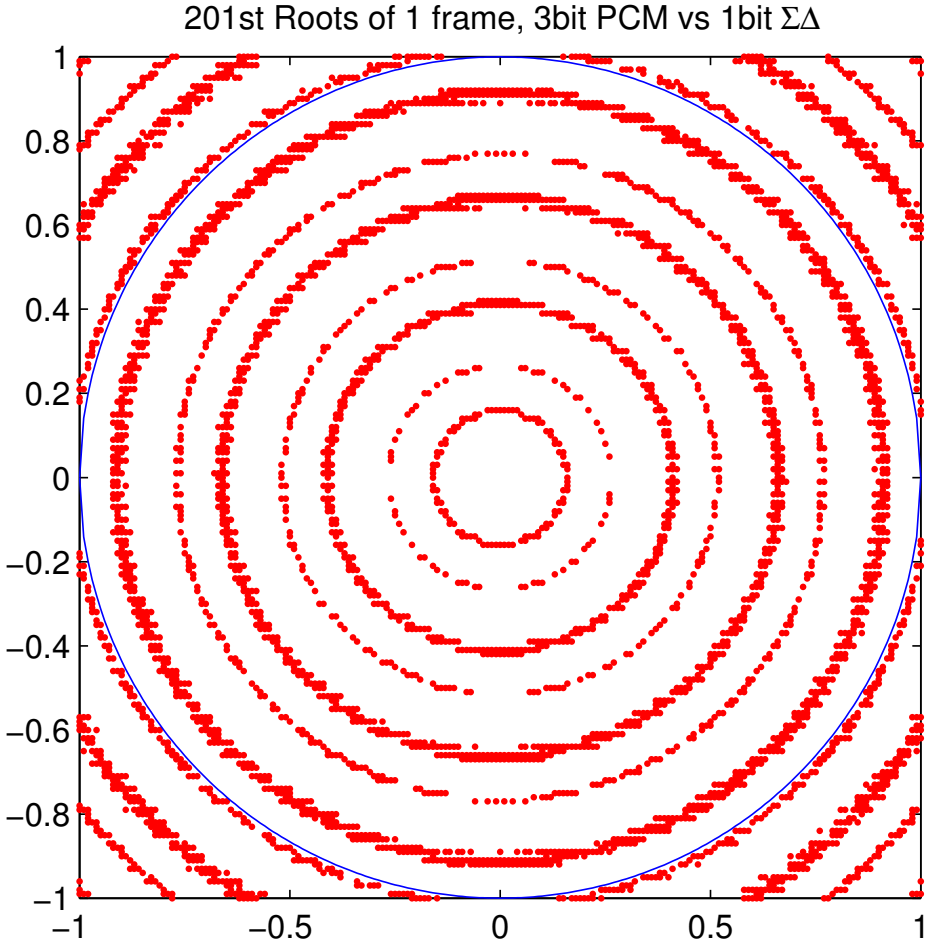

Red: $\mathrm{err}_{PCM}(x) < \mathrm{err}_{\Sigma\Delta}(x)$, Green: $\mathrm{err}_{PCM}(x) = \mathrm{err}_{\Sigma\Delta}(x)$.

# Comparison of 3-bit PCM and 1-bit ΣΔ



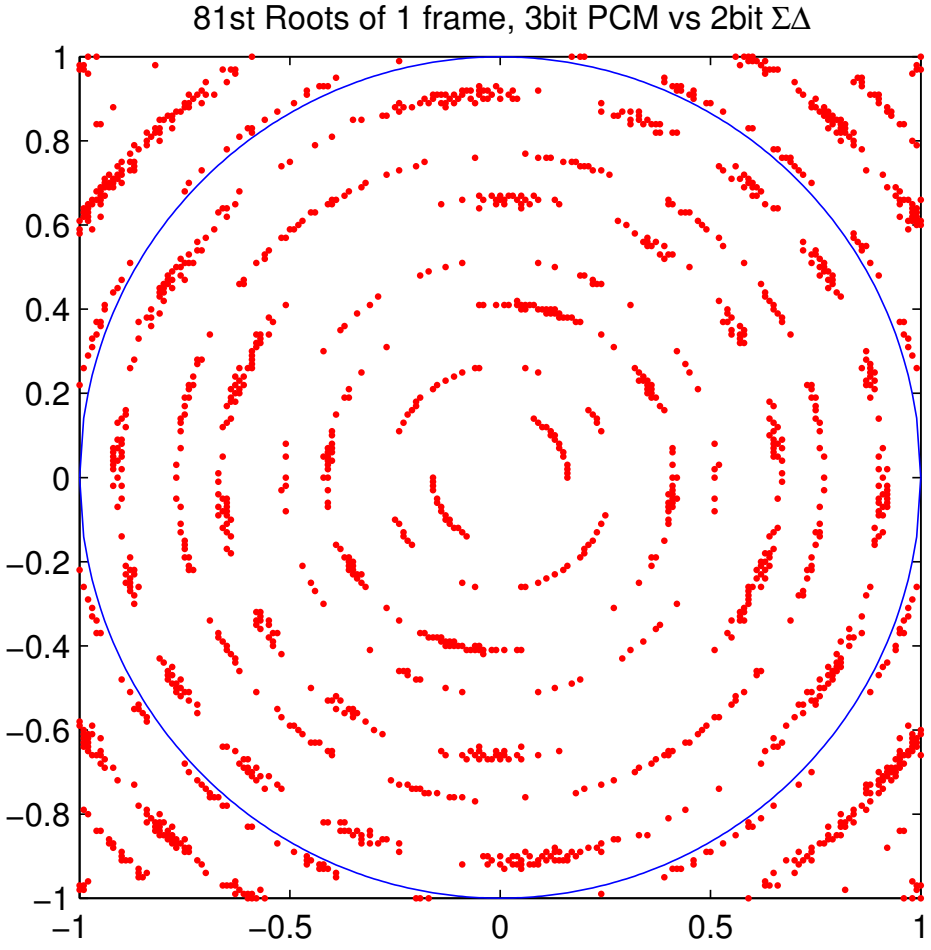

Red: $\mathrm{err}_{PCM}(x) < \mathrm{err}_{\Sigma\Delta}(x)$, Green: $\mathrm{err}_{PCM}(x) = \mathrm{err}_{\Sigma\Delta}(x)$.

# Comparison of 3-bit PCM and 1-bit ΣΔ



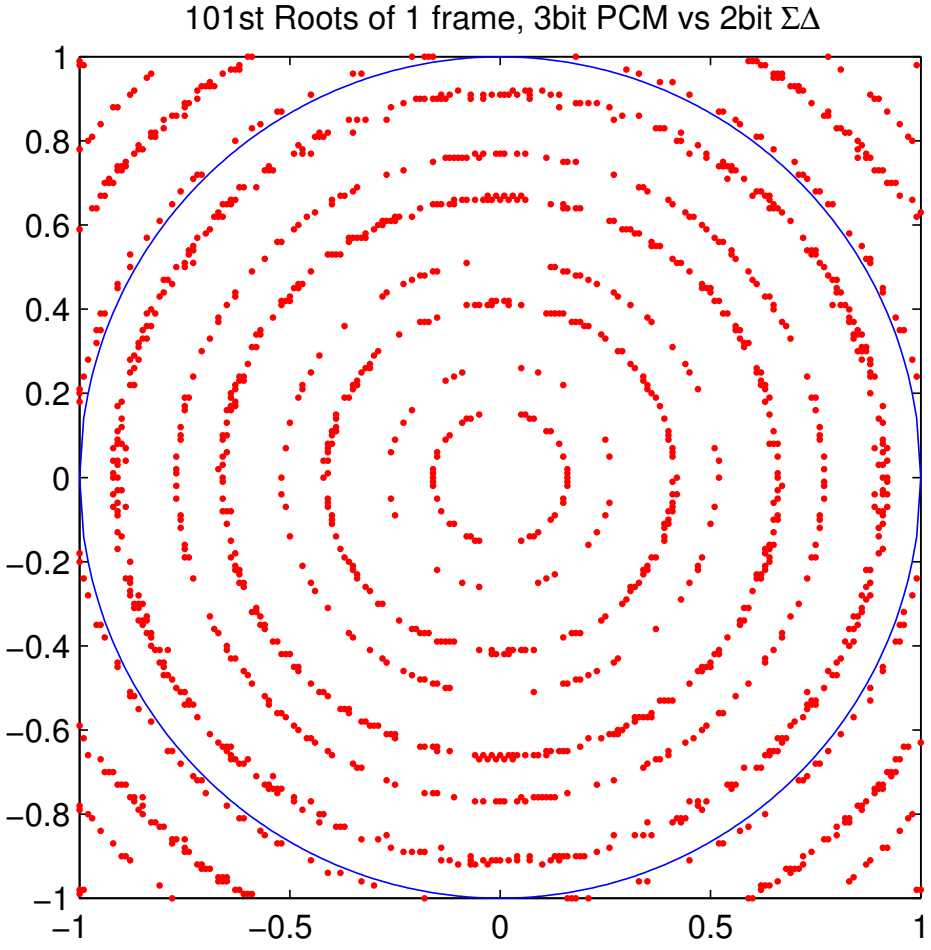

Red: $\mathrm{err}_{PCM}(x) < \mathrm{err}_{\Sigma\Delta}(x)$, Green: $\mathrm{err}_{PCM}(x) = \mathrm{err}_{\Sigma\Delta}(x)$.

# Comparison of 3-bit PCM and 2-bit ΣΔ



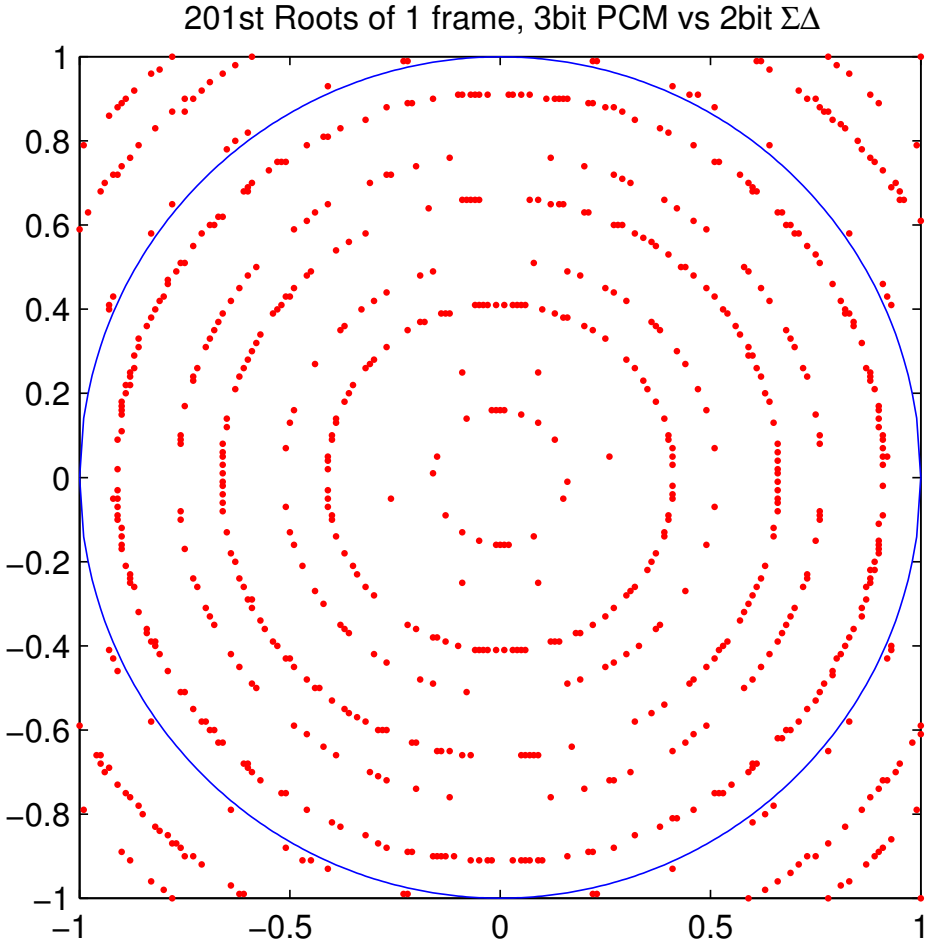

Red: $\mathrm{err}_{PCM}(x) < \mathrm{err}_{\Sigma\Delta}(x)$, Green: $\mathrm{err}_{PCM}(x) = \mathrm{err}_{\Sigma\Delta}(x)$.

# Comparison of 3-bit PCM and 2-bit $\Sigma\Delta$




Red: $\mathrm{err}_{PCM}(x) < \mathrm{err}_{\Sigma\Delta}(x)$, Green: $\mathrm{err}_{PCM}(x) = \mathrm{err}_{\Sigma\Delta}(x)$.

# Comparison of 3-bit PCM and 2-bit ΣΔ




Red: $\mathrm{err}_{PCM}(x) < \mathrm{err}_{\Sigma\Delta}(x)$, Green: $\mathrm{err}_{PCM}(x) = \mathrm{err}_{\Sigma\Delta}(x)$.

# Even – odd

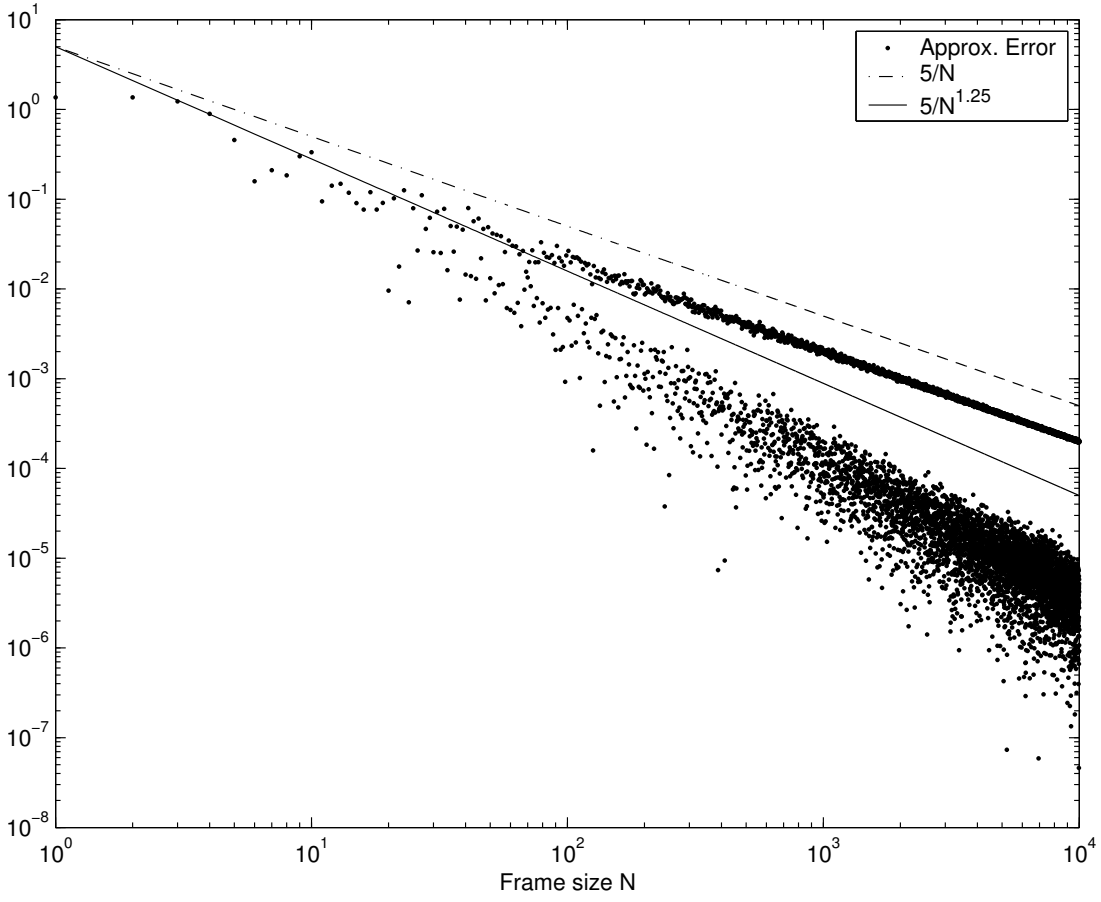



**Figure:** log-log plot of $\|x - \widetilde{x}_N\|$.

# Even – odd

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

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

Let $\widetilde{x}_N$ be the approximation given by the 1st order $\Sigma\Delta$ quantizer with alphabet $\{-1, 1\}$ and natural ordering.

# Improved estimates

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

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

using 1st order $\Sigma\Delta$ scheme with alphabet $\mathcal{A}_K^\delta$.

**Theorem**

If $N$ is even and large then $\|x - \widetilde{x}\| \leq B_x \frac{\delta \log N}{N^{5/4}}$.

If $N$ is odd and large then $A_x \frac{\delta}{N} \leq \|x - \widetilde{x}\| \leq B_x \frac{(2\pi+1)d}{N} \frac{\delta}{2}$.

- The proof uses a theorem of Güntürk (from complex or harmonic analysis); and Koksma and Erdös-Turán inequalities and van der Corput lemma (from analytic number theory).
- The Theorem is true for harmonic frames for $\mathbb{R}^d$.

# Proof of Improved Estimates theorem

- If $N$ is even and large then $\|x - \widetilde{x}\| \leq B_x \frac{\delta \log N}{N^{5/4}}$.
  If $N$ is odd and large then $A_x \frac{\delta}{N} \leq \|x - \widetilde{x}\| \leq B_x \frac{(2\pi+1)d}{N} \frac{\delta}{2}$.
- $\forall N, \{e_n^N\}_{n=1}^N$ is a FUNTF.

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

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

- To bound $v_n^N$.

# Koksma Inequality

## Definition

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

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

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

## Theorem(Koksma Inequality)

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

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

With $g(t) = t$ and $\omega_j = \widetilde{u}_j^N$,

$$|v_n^N| \le n\delta \mathrm{Disc}\left( \{\widetilde{u}_j^N\}_{j=1}^{n} \right).$$

# Erdös-Turán Inequality

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

To approximate the exponential sum.

# Approximation of Exponential Sum

## Güntürk's Proposition (1)

$\forall N, \exists X_N \in \mathcal{B}_{\Omega/N}$ such that , $\forall n = 0, \ldots, N$

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

and, $\forall t$,

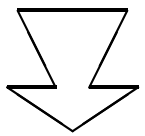
$$\left| X_N'(t) - h\left(\frac{t}{N}\right) \right| \leq B \frac{1}{N}$$

## Bernstein's Inequality (2)

If $x \in \mathcal{B}_\Omega$, then $\|x^{(r)}\|_\infty \leq \Omega^r \|x\|_\infty$

# Approximation of Exponential Sum

(1)+(2)



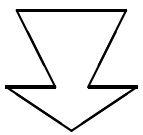
$$\forall t,\ \left| X_N''(t) - \tfrac{1}{N} h'\left(\tfrac{t}{N}\right) \right| \leq B \tfrac{1}{N^2}$$

- $\widehat{\mathcal{B}}_\Omega = \{ T \in A'(\widehat{\mathbb{R}}) \ : \ \operatorname{supp} T \subseteq [-\Omega, \Omega] \}$
- $\mathcal{M}_\Omega = \{ h \in \mathcal{B}_\Omega \ : \ h' \in L^\infty(\mathbb{R})$ and all zeros of $h'$ on $[0,1]$ 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 $f$ satisfy $f'' \geq \rho > 0$ on $[a, b]$ or $f'' \leq -\rho < 0$ on $[a, b]$. Then

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



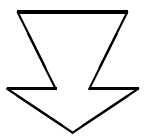
- $\forall 0 < \alpha < 1, \exists N_\alpha$ such that $\forall N \geq N_\alpha$,

$$\Big|\sum_{n=1}^{j} e^{2\pi i k \widetilde{u}_n^N}\Big| \leq B_x N^\alpha + B_x \frac{\sqrt{k} N^{1-\frac{\alpha}{2}}}{\sqrt{\delta}} + B_x \frac{k}{\delta}.$$

# Choosing appropriate $\alpha$ and $K$

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

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



**Conclusion**

$$\forall n = 1, \ldots, N, \ |v_n^N| \leq B_x \delta N^{\frac{3}{4}} \log N$$


Norbert Wiener Center
for Harmonic Analysis and Applications

Norbert Wiener Center
for Harmonic Analysis and Applications