# Distributed Noise Shaping of Signal Quantization

Kung-Ching Lin

Norbert Wiener Center
Department of Mathematics
University of Maryland, College Park

September 18, 2017

# Overview

1. Introduction
   - Pulse Coding Modulation (PCM) Quantization
   - Alternative Scheme: $\Sigma\Delta$ and Noise Shaping
2. Adaptation to Finite Dimensional Space
   - Quantization on Frame Setting
   - Modification on Dual Frame
3. Distributed Noise Shaping: Beta Dual
   - Main Result
   - Setting of Distributed Noise Shaping
   - Beta Dual for Unitarily Generated Frames
4. Reference

# A/D Conversion for Signals

## Theorem (Classical Sampling Theorem)

*Given $f \in PW_{[-1/2,1/2]}$, i.e. , $f, \hat{f} \in L^2(\mathbb{R})$, and supp$(\hat{f}) \subset [-1/2, 1/2]$. Then for any $g$ satisfying*

- $\hat{g}(\omega) = 1$ *on* $[-1/2, 1/2]$
- $\hat{g}(\omega) = 0$ *for* $|\omega| \geq 1/2 + \epsilon$,

*and for any $T \in (0, 1 - 2\epsilon)$, $t \in \mathbb{R}$,*

$$f(t) = T \sum_{n \in \mathbb{Z}} f(nT) g(t - nT) \tag{1}$$

*where the convergence is both uniform on compact sets and in $L^2$.*

## Remark

$f$ has a continuous representative, so it makes sense to evaluate $f$ at points.

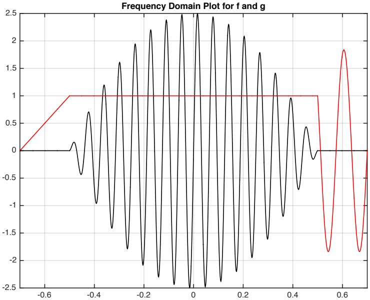


Figure: Black: Signal f, Red: Reconstruction Kernel g

- In particular, it is only necessary to store $\{f(nT)\}_{n\in\mathbb{Z}}$ to reconstruct $f$.
- Computers cannot store real numbers, so instead $\{q_n\}_{n\in\mathbb{Z}} \subset \mathscr{A}$ is considered where $\mathscr{A}$ is a finite subset of $\mathbb{R}$.

# Naive Approach: Pulse Coding Modulation (PCM)

Given a finite alphabet $\mathscr{A} \subset \mathbb{R}$, define $Q : \mathbb{R} \to \mathscr{A}$ by

$$Q(x) = \arg\min_{q \in \mathscr{A}} |x - q| \tag{2}$$

For a bandlimited function $f \in PW_{[-1/2,1/2]}$, its reconstructed function via PCM will be

$$\tilde{f}(t) = T \sum_{n \in \mathbb{Z}} Q(f(nT)) g(t - nT) \tag{3}$$

**Remark**

In practice, mid-rise uniform quantizer is often used. That is,

$$\mathscr{A} = \{(k + 1/2)\delta : k = -N, \ldots, N - 1\} \tag{4}$$

# Reconstruction Error Estimate of PCM

For the mid-rise uniform quantizer, the reconstruction error is

$$\begin{aligned} |f(t) - \tilde{f}(t)| &= T\left|\sum_{n\in\mathbb{Z}} \Big(f(nT) - Q(f(nT))\Big) g(t - nT)\right| \\ &\leq \delta \cdot T \sum_{n\in\mathbb{Z}} |g(t - nT)| \\ &= C_{g,T} \cdot \delta \end{aligned} \tag{5}$$

**Remark**

$C_{g,T} \to \|g\|_1$ as $T \to 0^+$, so oversampling doesn't improve the reconstruction significantly.

# Caveat of PCM: Imperfect Quantizers

Consider the following imperfect base quantizer

$$Q_1(x) = \begin{cases} 0 & \text{if} \quad x \leq \frac{1}{2} - \epsilon \\ 1 & \text{if} \quad x \geq \frac{1}{2} + \epsilon \\ 0 \text{ or } 1 & \text{if} \quad x \in (\frac{1}{2} - \epsilon, \frac{1}{2} + \epsilon) \end{cases} \tag{6}$$

For $x \in (0, 1)$,Let

$$Q^k(x) = \sum_{n=1}^{k} \frac{Q_n(x)}{2^n} \tag{7}$$

where $Q_n(x) = Q_1\left(2^{n-1}(x - \sum_{s=1}^{n-1} \frac{Q_s(x)}{2^s})\right)$.

Norbert Wiener Center for Harmonic Analysis and Applications

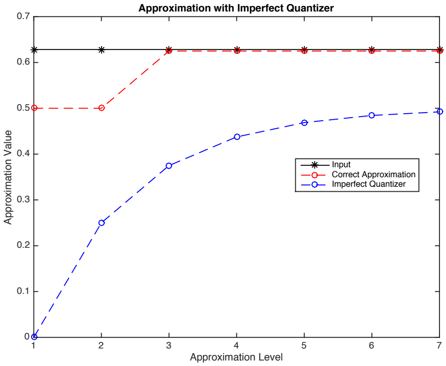


Figure: Illustration of Imperfect Quantization

## Remark

For $x \in (\frac{1}{2} - \epsilon, \frac{1}{2} + \epsilon)$, $|Q^k(x) - x| \geq \epsilon$ in worst case scenario, and we cannot improve such issue by adding more bits.

# Alternative Option: ΣΔ Quantization

Introduce auxiliary variable $\{u_n\}_{n\in\mathbb{Z}}$ and the recursive equation

$$u_{n+1} = u_n + f(nT) - q_n \tag{8}$$

where $q_n = Q(u_n + f(nT))$ for each $n$.

**Proposition (Uniform Boundedness of $\{u_n\}$)**

*With the choice of mid-rise uniform quantizer* $\mathscr{A} = \{(k + 1/2)\delta : k = -N, \dots, N-1\}$, $\|u\|_\infty < \delta$ *if* $\sup|f(nT)| \le (N-1)\delta$.
*We call such scheme a stable ΣΔ quantization scheme.*

Norbert Wiener Center for Harmonic Analysis and Applications

# Reconstruction Error for ΣΔ Quantization

Consider the reconstructed signal

$$\tilde{f}(t) = T \sum_{n\in\mathbb{Z}} q_n g(t - nT) \tag{9}$$

Then the reconstruction error is

$$\begin{aligned} |f(t) - \tilde{f}(t)| &= |T \sum_{n\in\mathbb{Z}} (f(nT) - q_n) g(t - nT)| \\ &= |T \sum_{n\in\mathbb{Z}} (u_{n+1} - u_n) g(t - nT)| \\ &= |T \sum_{n\in\mathbb{Z}} u_n (g(t - nT) - g(t - (n-1)T))| \\ &= |T \sum_{n\in\mathbb{Z}} u_n \int_{(n-1)T}^{nT} g'(t - u)\, du| \\ &\leq T\|u\|_\infty \|g'\|_1 \to 0 \quad \text{as } T \to 0 \end{aligned} \tag{10}$$

Note that $g$ is independent of $T$.

# Robustness of ΣΔ Against Imperfect Quantizer

- For ΣΔ quantization, the scheme is

$$u_{n+1} = u_n + f(nT) - q_n \tag{11}$$

- Imperfect quantizer $Q$ gives a larger sup-norm for $\{u_n\}$. In particular, it now changes to $\|u\|_\infty \leq \delta + \epsilon$.
- However, the scheme can still be stable, and the reconstruction error is still

$$\|f - \tilde{f}\|_\infty \leq T\|u\|_\infty\|g'\|_1 \tag{12}$$

# r-th Order ΣΔ Quantization

It is now a natural step to consider the following scheme:

$$y - q = \Delta^r u \tag{13}$$

Existence of a stable scheme of such kind is proven in [6, Daubechies & DeVore (2003)].

**Proposition (Error decay for high order ΣΔ)**

*Let $f$ and $\tilde{f}$ as before, except that $\{q_n\}$ now comes from (13). Suppose the kernel $g \in C^r$, then*

$$\|f - \tilde{f}\|_\infty \leq T^r \|u\|_\infty \|g^{(r)}\|_\infty \tag{14}$$

*Again, both $\|u\|_\infty$ and $\|g^{(r)}\|_\infty$ are independent of sampling period $T$.*

# Noise Shaping Feature of ΣΔ Quantization

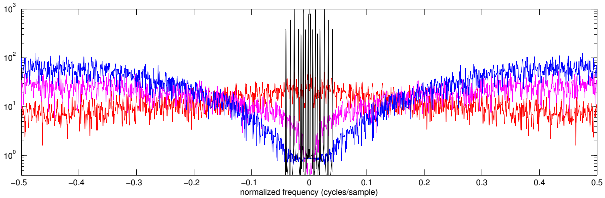


Figure: Classical noise shaping via ΣΔ modulation[5, Chou, Gunturk, Krahmer, Saab, Yilmaz (2015)]

- **Black** Fourier spectra of a bandlimited signal
- **Red** Quantization error signals using PCM
- **Pink** Error signal for 1st order ΣΔ quantization
- **Blue** Error signal for 2nd order ΣΔ quantization

# Generalization: Noise Shaping Quantization

Instead of difference operator $\Delta$, consider the following scheme:

$$y - q = h * u \tag{15}$$

where

1. $h = \{h_n\}_n \in \mathbb{N}$ has $h_0 = 1$, and
2. $(h * u)_n = \sum_{m=0}^{\infty} h_m u_{n-m}$

[8, Gunturk, 2003] constructed a family of $h$ to achieve exponential decay, with sub-optimal exponent.

Norbert Wiener Center for Harmonic Analysis and Applications

# Finite Frame and Quantization

For a given space $\mathbb{F}^k$ where $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$, suppose $\{e_n\}_{n=1}^m \subset \mathbb{F}^k$ is a spanning set and let the rows of $E \in \mathbb{F}^{m\times k}$ be $\{e_n^*\}$, the conjugate transpose of $\{e_n\}$.
Then for any dual $F \in \mathbb{F}^{k\times m}$, we have

$$FE = I_k \tag{16}$$

In particular, if $F = (f_1 \mid \cdots \mid f_m)$, then for any $x \in \mathbb{F}^k$,

$$x = \sum_{n=1}^{m} < x, e_n > f_n \tag{17}$$

where $f_n \in \mathbb{F}^k$. Then the quantized version $\tilde{x}$ shall be

$$\tilde{x} = \sum_{n=1}^{m} q_n f_n \tag{18}$$

# First Order ΣΔ Quantization for Finite Frames

Consider the following scheme:

$$y - q = \Delta u \tag{19}$$

Then the reconstruction error $\|x - \tilde{x}\|_2$ is

$$\begin{aligned}
\|x - \tilde{x}\|_2 &= \|\sum_{n=1}^{m}(< x, e_n > - q_n)f_n\|_2 \\
&= \|\sum_{n=1}^{m}((u_n - u_{n-1})f_n)\|_2 \\
&= \|\sum_{n=1}^{m} u_n(f_n - f_{n+1}) + u_m f_m\|_2 \\
&\leq \|u\|_\infty(\sum_{n=1}^{m}\|f_n - f_{n+1}\|_2 + \|f_m\|_2)
\end{aligned} \tag{20}$$

# Frame Analogy of ΣΔ Quantization[1]

## Definition (Frame Variation)

Let $E = \{e_n\}_{n=1}^{m}$ be a finite frame for $\mathbb{R}^k$, and $p$ a permutation of $\{1, 2, \ldots, N\}$. The variation of the frame $E$ with respect to $p$ is

$$\sigma(E, p) := \sum_{n=1}^{m-1} \|e_{p(n)} - e_{p(n+1)}\|_2 \tag{21}$$

## Theorem ([1], Benedetto, Powell, Yilmaz, 2006)

*Suppose $E = \{e_n\}_{n=1}^{m}$ is a zero-sum FUNTF with frame bound $m/k$. Then the reconstruction error $\|x - \tilde{x}\|_2$ satisfies*

$$\|x - \tilde{x}\|_2 \leq \begin{cases} \frac{\delta k}{2m}\sigma(E, p) & \text{if} \quad m \quad \text{is even} \\ \frac{\delta k}{2m}(\sigma(E, p) + 1) & \text{if} \quad m \quad \text{is odd} \end{cases} \tag{22}$$

# Noise Shaping Scheme

**Definition**

Let $\mathscr{B}$ be the unit ball centered around origin of $\mathbb{F}^m$.
A map $Q : \mathscr{B} \to \mathscr{A}^m$ is a stable noise shaping scheme if $\exists H$: lower triangular, $\{u_n\}$ uniformly bounded by $\delta$ such that

$$y - q = Hu \tag{23}$$

where $q = Q(y)$ and we use $\tilde{x} = Fq$ as the reconstruction vector for $x$.

- Stability: [4, Chou, Gunturk, 2016] gives a sufficient condition for such scheme to work
- Recursiveness: Requires $H$ to be lower triangular

# Reconstruction Frame/ Kernel $F$

Many choices. Depends on $H$ in quantization scheme
$(\|x - Fq\| = \|F(y - q)\| = \|FHu\|)$

- Canonical dual: $E^\dagger$
- V-dual: $(VE)^\dagger V$ such that $VE$ is still a frame.
  1. Sobolev dual [2]: $(\Delta^{-1}E)^\dagger\Delta^{-1}$ achieves minimum 2-norm for $F\Delta$.
  2. Alternative dual: $(H^{-1}E)^\dagger H^{-1}$
  3. Beta dual: $(V_\beta E)^\dagger V_\beta$, $V_\beta$ to be specified later.

How does the choice of $V$ affect our reconstruction?

# De-Noising Aspect of Dual Frame

$$
\begin{tikzcd}
 & im(H^{-1}E) & & im(VE) \\
\mathbb{F}^{k} \arrow[r, "H^{-1}E"] & \mathbb{F}^{m} \arrow[u, "\text{Alt. Dual}"'] \arrow[r, "H"] & \mathbb{F}^{m} \arrow[r, "V"] & \mathbb{F}^{p} \arrow[u, "\text{Noise Shaping}"'] \\
 & u \arrow[u, "\text{Noise}"'] & &
\end{tikzcd}
$$

Figure: Error Cutoff Procedures for Different Quantization Schemes

For $M \in \mathbb{F}^{\ell \times k}$ injective, $ker(M^{\dagger}) = (M(\mathbb{F}^{k}))^{\perp}$.

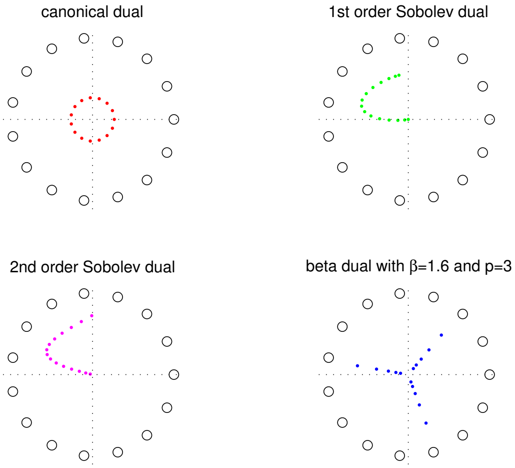


Figure: Different alternative duals for the 15th roots-of-unity frame in $\mathbb{R}^2$ [5]

## Theorem (Chou, Gunturk)

*Given a unitarily generated frame $\Phi$ with generator $\Omega$, a $k \times k$ Hermitian matrix with $\{v_s\}_{s=1}^{k}$ being a basis of orthonormal eigenvectors and $\phi_0 \in \mathbb{F}^k$. Suppose the eigenvalues are all distinct modulo $l$ where $l \geq k$, then we have*

$$\|x - F_V q\|_2 < 7e\left(\frac{m}{l} + 1\right)c(\phi_0) \cdot \begin{cases} \sqrt{2}\lfloor\sqrt{L}\rfloor^{-m/l} & \text{if} \quad \mathbb{F} = \mathbb{C} \\ L^{-m/l} & \text{if} \quad \mathbb{F} = \mathbb{R} \end{cases} \tag{24}$$

*where*

$$c(\phi_0) := \left(\min_{1 \leq s \leq k} |<\phi_0, v_s>|\right)^{-1} \tag{25}$$

# Setup of Distributed Noise Shaping (DNS)

**Definition (V-dual)**

Let $E \in \mathbb{R}^{m \times k}$ be a frame, $m > k$. $F_V \in \mathbb{R}^{k \times m}$ is a V-dual of $E$ if

$$F_V = (VE)^{\dagger} V \tag{26}$$

where $V \in \mathbb{R}^{p \times m}$ such that $VE$ is still a frame.

Recall that a stable noise shaping scheme has

$$y - q = Hu \tag{27}$$

# Setup of DNS (Cont'd)

In the setting of DNS, $V \in \mathbb{R}^{p\times m}$ and $H \in \mathbb{R}^{m\times m}$ are block matrices

$$V = \begin{pmatrix} V_1 & & & & \\ & V_2 & & & \\ & & V_3 & & \\ & & & \ddots & \\ & & & & V_l \end{pmatrix}, \quad H = \begin{pmatrix} H_1 & & & & \\ & H_2 & & & \\ & & H_3 & & \\ & & & \ddots & \\ & & & & H_l \end{pmatrix} \tag{28}$$

where $V_i \in \mathbb{R}^{p_i\times m_i}$, $H_i \in \mathbb{R}^{m_i\times m_i}$ with $\sum p_i = p, \sum m_i = m$.

Norbert Wiener Center for Harmonic Analysis and Applications

# $\beta$-Dual

**Definition ($\beta$-dual)**

A $\beta$-dual $F_V = (VE)^\dagger V$ has $V = V_{\boldsymbol{\beta},\boldsymbol{m}}$, where $\boldsymbol{\beta} = [\beta_1, \dots, \beta_l]^t$ and $\boldsymbol{m} = [m_1, \dots, m_l]^t$, is a $l$-by-$m$ block matrix such that $V_i = [\beta_i^{-1}, \beta_i^{-2}, \dots, \beta_i^{-m_i}] \in \mathbb{R}^{1\times m_i}$, i.e. $l = p$.

- $\{\beta_i\}$ satisfies $\beta_i > 1$ for every $1 \le i \le l$, whose choice is limited by a technical lemma.
- Under this setting, each $H_i$ is chosen to be a $m_i \times m_i$ matrix with unit diagonal entries and $-\beta_i$ sub-diagonal entries.

# $\beta$-Dual (Cont'd)

$$V_i H_i = [\beta_i^{-1} \beta_i^{-2} \cdots \beta_i^{-m_i}] \begin{pmatrix} 1 & & & & \\ -\beta_i & 1 & & & \\ 0 & -\beta_i & 1 & & \\ 0 & 0 & \ddots & \ddots & \\ & & & -\beta_i & 1 \end{pmatrix} \tag{29}$$

$$= \begin{pmatrix} 0 & 0 & \dots & 0 & \beta_i^{-m_i} \end{pmatrix}$$

where $\beta_i > 1$, so it effectively reduced the size of error.

Norbert Wiener Center for Harmonic Analysis and Applications

# Reconstruction Error Estimate for $\beta$-Dual

**Lemma (Chou, Gunturk, 2016)**

*Given a $\beta$-dual $V$, suppose $VE$ is a frame, then the reconstruction error is*

$$
\begin{aligned}
\|x - F_V q\|_2 = \|F_V H u\|_2 &\leq \|F_V H\|_{\infty\to 2}\|u\|_\infty \\
&\leq \|u\|_\infty \frac{1}{\sigma_{min}(VE)}\|VH\|_{\infty\to 2} \\
&\leq \frac{\sqrt{l}}{\sigma_{min}(VE)}\delta\beta^{-\lfloor m/l\rfloor} \\
&< \frac{\|E\|_{2\to\infty} e(1+\lfloor m/l\rfloor)\sqrt{l}}{\sigma_{min}(VE)} L^{-(\lfloor m/l\rfloor+1)}
\end{aligned}
\tag{30}
$$

# Proof of Theorem[3]

- Assume that $m/l \in \mathbb{N}$ for simplicity.
- Assumption: Eigenvalues $\{\lambda_s\}$ are distinct modulo $l$.
- $\sigma_{\min}(VE)$ can be controlled uniformly for all $m$ in such setting. $\|E\|_{2\to\infty}$ is easily seen to be uniformly bounded.

Given a $k \times k$ Hermitian matrix $\Omega$ and $\phi_0 \in \mathbb{F}^k$, consider

$$U_t := e^{2\pi\imath\Omega t}, \quad \phi_n = U_{\frac{n}{m}}\phi_0, \quad n = 0, \dots, m-1 \tag{31}$$

Set $E$ to be the collection of such elements, that is,

$$E = \begin{pmatrix} \phi_0^* \\ \vdots \\ \phi_{m-1}^* \end{pmatrix} = \begin{pmatrix} E_1 \\ \vdots \\ E_l \end{pmatrix} \tag{32}$$

# Proof (Cont'd)

Then,

$$VE = \begin{pmatrix} V_1 E_1 \\ \vdots \\ V_l E_l \end{pmatrix} \tag{33}$$

where

$$(V_j E_j)^* = \sum_{n=1}^{m/l} \beta^{-n} \phi_{(j-1)m/l+n} \in \mathbb{F}^k \tag{34}$$

Now, let $\{v_s\}$ be an ONB of eigenvectors with respect to $\Omega$ with eigenvalue $\{\lambda_s\}$. Then

$$\begin{aligned} < (V_j E_j)^*, v_s > &= \sum_{n=1}^{m/l} \beta^{-n} < U_{\frac{(j-1)m/l+n}{m}} \phi_0, v_s > \\ &= \sum_{n=1}^{m/l} \beta^{-n} < \phi_0, U_{\frac{-(j-1)m/l-n}{m}} v_s > \\ &= \sum_{n=1}^{m/l} \beta^{-n} e^{-2\pi\imath \frac{(j-1)m/l+n}{m} \lambda_s} < \phi_0, v_s > \\ &= e^{-2\pi\imath \frac{(j-1)}{l} \lambda_s} w_s < \phi_0, v_s > \end{aligned} \tag{35}$$

where

$$w_s = \sum_{n=1}^{m/l} \left(\beta^{-1} e^{2\pi\imath\lambda_s/m}\right)^n \tag{36}$$

## Remark

$$|w_s| = \Big|\sum_{n=1}^{m/l} \left(\beta^{-1} e^{2\pi\imath\lambda_s/m}\right)^n\Big| = \left|\frac{1-\beta^{-m/l}e^{2\pi\imath\lambda_s/l}}{1-\beta^{-1}e^{2\pi\imath\lambda_s/m}}\right| \geq \frac{1-\beta^{-1}}{1+\beta^{-1}} \tag{37}$$

Then, for any $x \in \mathbb{F}^k$,

$$\begin{aligned}
\|VEx\|_2^2 &= \sum_{j=1}^{l} |V_j E_j x|^2 \\
&= \sum_{j=1}^{l} \Big| \sum_{s=1}^{k} < (V_j E_j)^*, v_s >< x, v_s > \Big|^2 \\
&= \sum_{s=1}^{k} \sum_{t=1}^{k} < x, v_s >< v_t, x >< \phi_0, v_s >< v_s, \phi_0 > w_s \bar{w}_t \sum_{j=1}^{l} e^{2\pi\imath(j-1)(\lambda_t - \lambda_s)/l} \\
&= l \sum_{s=1}^{k} | < x, v_s > |^2 | < \phi_0, v_s > |^2 |w_s|^2 \\
&\geq l \left( \frac{1-\beta^{-1}}{1+\beta^{-1}} \right)^2 \min_{1 \leq s \leq k} | < \phi_0, v_s > |^2 \|x\|_2^2
\end{aligned} \tag{38}$$

if $\lambda_s - \lambda_t$ are integers and nonzero modulo $l$.

# Reference I

J. J. Benedetto, A. M. Powell, and O. Yilmaz.
Sigma-delta quantization and finite frames.
*IEEE Transactions on Information Theory*, 52(5):1990–2005, 2006.

J. Blum, M. Lammers, A. M. Powell, and Ö. Yılmaz.
Sobolev duals in frame theory and sigma-delta quantization.
*Journal of Fourier Analysis and Applications*, 16(3):365–381, 2010.

E. Chou and C. S. Güntürk.
Distributed noise-shaping quantization: Ii. classical frames.
*Not yet published*, .

E. Chou and C. S. Güntürk.
Distributed noise-shaping quantization: I. beta duals of finite frames and near-optimal quantization of random measurements.
*Constructive Approximation*, 44(1):1–22, 2016.

E. Chou, C. S. Güntürk, F. Krahmer, R. Saab, and Ö. Yılmaz.
*Noise-shaping quantization methods for frame-based and compressive sampling systems*.
Number 157–184. Springer, 2015.

I. Daubechies and R. DeVore.
Approximating a bandlimited function using very coarsely quantized data: A family of stable sigma-delta modulators of arbitrary order.
*Annals of mathematics*, 158(2):679–710, 2003.

Norbert Wiener Center
for Harmonic Analysis and Applications

# Reference II

I. Daubechies, R. A. DeVore, C. S. Gunturk, and V. A. Vaishampayan.
A/d conversion with imperfect quantizers.
*IEEE Transactions on Information Theory*, 52(3):874–885, 2006.

C. S. Güntürk.
One-bit sigma-delta quantization with exponential accuracy.
*Communications on Pure and Applied Mathematics*, 56(11):1608–1630, 2003.