Adapted Decimation on Finite Frames for Arbitrary Orders of Sigma-Delta Quantization

Kung-Ching Lin
Norbert Wiener Center
Department of Mathematics, University of Maryland
College Park, MD 20742 USA
Email: kclin@math.umd.edu

***Abstract*—In Analog-to-digital (A/D) conversion, signal decimation has been proven to greatly improve the efficiency of data storage while maintaining high accuracy. When one couples signal decimation with the $\Sigma\Delta$ quantization scheme, the reconstruction error decays exponentially with respect to the bit-rate. We build on our previous result, which extends signal decimation to finite frames up to the second order. In this study we introduce a new scheme called adapted decimation, which yields polynomial reconstruction error decay rate of arbitrary order with respect to the oversampling ratio, and exponential decay rate with respect to the bit-rate.**

I. Introduction

A. Signal Quantization

Analog-to-digital (A/D) conversion is a process where bandlimited signals, e.g., audio signals, are digitized for storage and transmission, which is feasible thanks to the classical sampling theorem. In particular, the theorem indicates that discrete sampling is sufficient to capture all features of a given bandlimited signal, provided that the sampling rate is higher than the Nyquist rate.

Given a function $f \in L^1(\mathbb{R})$, its Fourier transform $\hat{f}$ is defined as

$$\hat{f}(\gamma) = \int_{-\infty}^{\infty} f(t) e^{-2\pi \imath t \gamma}\, dt.$$

The Fourier transform can also be uniquely extended to $L^2(\mathbb{R})$ as a unitary transformation.

Definition I.1. Given $f \in L^2(\mathbb{R})$, $f \in PW_\Omega$ if its Fourier transform $\hat{f} \in L^2(\mathbb{R})$ is supported in $[-\Omega, \Omega]$.

An important component of A/D conversion is the following theorem:

Theorem I.2 (Classical Sampling Theorem)**.** *Given $f \in PW_{[-1/2,1/2]}$, for any $g \in L^2(\mathbb{R})$ satisfying*

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

with $\epsilon > 0$ and $T \in (0, 1 - 2\epsilon)$, $t \in \mathbb{R}$, one has

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

where the convergence is both uniform on compact sets of $\mathbb{R}$ and in $L^2(\mathbb{R})$.

However, the discrete nature of digital data storage makes it impossible to store exactly the samples $\{f(nT)\}_{n\in\mathbb{Z}}$. Instead, the quantized samples $\{q_n\}_{n\in\mathbb{Z}}$ chosen from a pre-determined finite alphabet $\mathscr{A}$ are stored. This results in the following reconstructed signal

$$\tilde{f}(t) = T \sum q_n g(t - nT).$$

As for the choice of the quantized samples $\{q_n\}_n$, we shall discuss the following two schemes.

- Pulse Code Modulation (PCM):
  Quantized samples are taken as the direct-roundoff of the current sample, i.e.,
  $$q_n = Q_0(f(nT)) := \arg\min_{q \in \mathscr{A}} |q - f(nT)|. \tag{2}$$
- $\Sigma\Delta$ Quantization:
  A sequence of auxiliary variables $\{u_n\}_{n\in\mathbb{Z}}$ is introduced for this scheme. $\{q_n\}_{n\in\mathbb{Z}}$ is defined recursively as
  $$q_n = Q_0(u_{n-1} + f(nT)),$$
  $$u_n = u_{n-1} + f(nT) - q_n.$$

$\Sigma\Delta$ quantization was introduced in 1963 [13] and is still widely used, due to some of its advantages over PCM. Specifically, $\Sigma\Delta$ quantization is robust against hardware imperfection [8], a decisive weakness for PCM.

As its direct generalization, given $r \in \mathbb{N}$, one can consider an *r-th order $\Sigma\Delta$ quantization scheme*:

$$f(nT) - q_n = (\Delta^r u)_n,$$

where, given $\{v_n\}_{n\in\mathbb{Z}}$, $(\Delta v)_n = v_n - v_{n-1}$. Higher order $\Sigma\Delta$ quantization has been known for a long time [6], [11], and it improves the error decay rate from linear

to polynomial degree r while preserving the advantages of a first order $\Sigma\Delta$ quantization scheme.

B. Signal Decimation

Given an r-th order $\Sigma\Delta$ quantization scheme, there exist $\{q_n^T\}, \{u_n\}$ such that

$$f(nT) - q_n^T = (\Delta^r u)_n, \tag{3}$$

where $\|u\|_\infty < \infty$. Then, consider

$$\tilde{q}_n^{T_0} := (S_\rho^r q^T)_{(2\rho+1)n}, \tag{4}$$

a sub-sampled sequence of $S_\rho^r q^T$, where $(S_\rho h)_n := \frac{1}{2\rho+1}\sum_{m=-\rho}^{\rho} h_{n+m}$.

The process where we convert the quantized samples $\{q_n^T\}$ to $\{\tilde{q}_n^{T_0}\}$ is called signal decimation. See Figure 1 for an illustration of decimation.

Decimation has been known in the engineering community [3], and it was observed to result in exponential error decay with respect to the bit-rate, even though the observation remained a conjecture until 2015 [9], when Daubechies and Saab proved the following theorem:

Theorem I.3 (Signal Decimation for Bandlimited Functions, [9]). *Given $f \in PW_{1/2}$, $T < 1$, and $T_0 = (2\rho+1)T < 1$, there exists a function $\tilde{g}$ such that*

$$|f(t) - T_0 \sum \tilde{q}_n^{T_0} \tilde{g}(t - nT_0)| \le C^r \|u\|_\infty \left(\frac{T}{T_0}\right)^r =: \mathcal{D}, \tag{5}$$

where $\tilde{q}_n^{T_0}$ is defined in (4)*, and C depends on neither T nor T_0. Moreover, the number of bits needed for each unit interval is*

$$\frac{1}{T_0}\log_2((2\rho+1)^r + 1) \le \frac{1}{T_0}\log_2\left(2\left(\frac{T_0}{T}\right)^r\right) =: \mathcal{R}. \tag{6}$$

Consequently,

$$\mathcal{D}(\mathcal{R}) = 2C_{\Sigma\Delta} C^r 2^{-T_0\mathcal{R}}.$$

In [14], the author made an extension of decimation to finite frames, and the basic terminology of quantization for finite frames will be introduced below.

C. Quantization for Finite Frames

Fix a Euclidean space $\mathbb{C}^k$, a collection of vectors $E = \{e_j\}_{j=1}^m$ is a frame for $\mathbb{C}^k$ if for any $x \in \mathbb{C}^k$, there exists $A, B > 0$ such that $A\|x\|_2^2 \le \sum_j |<x, e_j>|^2 \le B\|x\|_2^2$. The largest possible A is the *lower frame bound* of E. Consider an m-by-k matrix with $\{e_j^*\}_j$ as its rows. With abuse of notation, we also denote it as E. Then, given $x \in \mathbb{C}^k$, the r-th order $\Sigma\Delta$ quantization satisfies

$$Ex - q = \Delta^r u$$

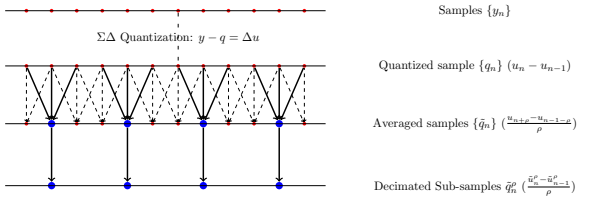

Figure 1: Illustration of the first order decimation scheme. After obtaining the quantized samples $\{q_n\}_n$, decimation takes the average of quantized samples within disjoint blocks. Such outputs are used as the decimated sub-samples $\{\tilde{q}_n^\rho\}$. The effect on the reconstruction (replacing q_n with $y_n - q_n$) is illustrated in parentheses.

with $q, u \in \mathbb{C}^m$, and $\Delta \in \mathbb{C}^{m\times m}$ is the backward difference matrix, with 1 on the diagonal entries, and -1 on the sub-diagonal entries. Using the alternative decimation operator, it is proven that up to the second order sigma-delta quantization, results similar to Theorem I.3 can be achieved:

Definition I.4 (Alternative Decimation)**.** Given fixed $m, \rho \in \mathbb{N}$, the (r, m, ρ)-alternative decimation operator is defined to be $D_\rho S_\rho^r$, where

- $S_\rho = S_\rho^+ - S_\rho^- \in \mathbb{R}^{m\times m}$ is the integration operator satisfying

$$(S_\rho^+)_{l,j} = \begin{cases} \frac{1}{\rho} & \text{if} \quad l \ge \rho,\ l - (\rho-1) \le j \le l \\ 0 & \text{otherwise,} \end{cases}$$

$$(S_\rho^-)_{l,j} = \begin{cases} \frac{1}{\rho} & \text{if} \quad l \le \rho - 1,\ j \in [l+1, m-\rho+l] \\ 0 & \text{otherwise,} \end{cases}$$

and

- $D_\rho \in \mathbb{N}^{\eta\times m}$ is the sub-sampling operator satisfying

$$(D_\rho)_{l,j} = \begin{cases} 1 & \text{if} \quad j = \rho \cdot l \\ 0 & \text{otherwise,} \end{cases}$$

where $\eta = m/\rho$.

Definition I.5 (Unitarily Generated Frames (UGF))**.** Given a base vector $\phi_0 \in \mathbb{C}^k$ and a Hermitian matrix $\Omega \in \mathbb{R}^{k\times k}$, the unitarily generated frame $\Phi_{m,k} = \{\phi_j^{(m)}\}_j$ is

$$\phi_j^{(m)} = U_{j/m}\phi_0, \quad U_t := e^{2\pi\imath\Omega t}. \tag{7}$$

The eigenvalues and eigenvectors of Ω are represented as $\{\lambda_j\}_j$ and $\{v_j\}_j$.

Theorem I.6 (Alternative Decimation for Finite Frames up to the Second Order, [14])**.** *Given Ω, ϕ_0, $\{\lambda_j\}_j$, $\{v_j\}_j$, and $\Phi = \Phi_{m,k}$ as the generator, base vector, eigenvalues, eigenvectors, and the corresponding UGF, respectively, and $r = 1, 2$. Suppose*

- $\{\lambda_j\}_{j=1}^k \subset [-\eta/2, \eta/2] \cap \mathbb{Z}\backslash\{0\}$,

- $C_{\phi_0} = \min_s |<\phi_0, v_s>|^2 > 0$, *and*
- $\rho \mid m$,

then the dual frame $F = (D_\rho S_\rho^r \Phi_{m,k})^\dagger D_\rho S_\rho^r$ *combined with the* r*-th order* $\Sigma\Delta$ *quantization has reconstruction error* $\mathscr{E}_{m,\rho,r}$ *with polynomial decay rate of degree* r *with respect to the oversampling ratio* ρ*:*

$$\mathscr{E}_{m,\rho,r} \leq C\|u\|_\infty \frac{1}{\rho^r}.$$

Moreover, the total number of bits used to record the quantized samples is $\mathscr{R} = O(\log(m))$ *bits. Suppose* $\eta = m/\rho$ *is fixed as* $m \to \infty$*, then as a function of total number of bits used,* $\mathscr{E} = \mathscr{E}_{m,\rho}$ *satisfies*

$$\mathscr{E}(\mathscr{R}) \leq C\|u\|_\infty 2^{-\frac{1}{2\eta}\mathscr{R}}.$$

II. Perspective and Prior Works

1) Quantization for Bandlimited Functions: It was proven in [7] that the r-th order $\Sigma\Delta$ quantization has error decay of polynomial order r. Leveraging the different constants for this family of quantization schemes, sub-exponential decay can also be achieved. A different family of quantization schemes was shown [12] to have exponential error decay with small exponent ($c \approx 0.07$.) In [10], the exponent was improved to $c \approx 0.102$.

2) Finite Frames: It was proven [1] that for any family of finite frames with bounded frame variation, the reconstruction error decays linearly with respect to the oversampling ratio. With different choices of dual frames, [2] proved that the so-called Sobolev dual achieves minimum induced matrix 2-norm for reconstructions, and [5] proved that using a β-dual for random frames results in exponential decay of near-optimal exponent and with high probability.

3) Decimation: In [3], using the assumption that the noise in $\Sigma\Delta$ quantization is random along with numerical experiments, it was asserted that decimation greatly reduces the number of bits needed while maintaining the reconstruction accuracy. In [9], a rigorous proof was given to show that such an assertion is indeed valid, and the reduction of bits used turns the linear decay into exponential decay with respect to the bit-rate.

4) Beta Dual of Distributed Noise Shaping: Chou and Günturk [5], [4] proposed a distributed noise shaping quantization scheme with beta dual. The definition of a beta dual is as follows:

Definition II.1 (Beta Dual)**.** Let $E \in \mathbb{C}^{m\times k}$ be an analysis operator and suppose $k \mid m$. Given $\beta > 1$, the β-dual $F_V = (VE)^\dagger V$ has $V = V_{\beta,m}$, a k-by-m block matrix such that each block is $v = [\beta^{-1}, \beta^{-2}, \ldots, \beta^{-m/k}] \in \mathbb{R}^{1\times m/k}$.

In this case, the noise shaping scheme is $y - q = Hu$, where H is an m-by-m block matrix with each block h as an m/k-by-m/k matrix with unit diagonal entries and $-\beta$ as sub-diagonal entries. In this setting, it is proven that the reconstruction error decays exponentially.

III. Main Results

We have seen in Theorem I.6 that alternative decimation is only useful up to the second order. Thus, we extend our results to arbitrary orders, and the solution we present here is called the adapted decimation.

Definition III.1 (Adapted Decimation)**.** Given $r, m, \rho \in \mathbb{N}$, the (r, m, ρ)-adapted decimation operator is defined to be

$$A_r = \frac{1}{\rho^r} D_\rho \bar{\Delta}_\rho^r \Delta^{-r},$$

where $\bar{\Delta}_\rho \in \mathbb{R}^{m\times m}$ satisfies $(\bar{\Delta}_\rho)_{l,s} = \delta(l-s) - \delta(l+\rho-s) + \delta(s-m)\delta(l-\rho)$, and $D_\rho \in \mathbb{N}^{m/\rho\times m}$ has $(D_\rho)_{l,s} = \delta(s - l\rho)$.

Theorem III.2. *Given* Ω*,* ϕ_0*,* $\{\lambda_j\}_j$*,* $\{v_j\}_j$*, and* $\Phi = \Phi_{m,k}$ *as the generator, base vector, eigenvalues, eigenvectors, and the corresponding UGF, respectively, and* $r \in \mathbb{N}$ *fixed. Suppose*

- $\eta \geq 3rk$,
- $\{\lambda_j\}_{j=1}^k \subset [-\eta/2, \eta/2] \cap \mathbb{Z}\backslash\{0\}$,
- $C_{\phi_0} = \min_s |<\phi_0, v_s>|^2 > 0$, *and*
- $\rho \mid m$,

where $\eta = m/\rho$*. Then the following statements are true.*

(a) **Recursivity:** *For all* $s \in \{1, \ldots, \eta\}$*, there exists* $\{c_j^s\}_{j=1}^{s\rho}$ *such that* $(A_r q)_s = \sum_{j=1}^{s\rho} c_j^s q_j$.

(b) **Signal reconstruction:** $A_r\Phi_{m,k}$ *is a frame.*

(c) **Error estimate:** *For the dual frame* $F = (A_r\Phi_{m,k})^\dagger A_r$*, the reconstruction error* $\mathscr{E}_{m,\rho}$ *satisfies*

$$\mathscr{E}_{m,\rho} \leq \left(\frac{4k}{\eta} C_{\phi_0} \big(\frac{8\eta}{\pi}\big)^r\right) \|u\|_\infty \frac{1}{\rho^r}. \tag{8}$$

(d) **Efficient data storage:** *Suppose the length of the quantization alphabet is* $2L$*, then the total number of bits used to record the quantized samples* $A_r q$ *is* $\mathscr{R} = 2\eta r \log(2m) + 2\eta \log(2L)$ *bits. Furthermore, suppose* $\eta = m/\rho$ *is fixed as* $m \to \infty$*, then as a function of total number of bits used,* $\mathscr{E} = \mathscr{E}_{m,\rho}$ *satisfies*

$$\mathscr{E}(\mathscr{R}) \leq C_{k,\eta,\phi_0,L} \|u\|_\infty 2^{-\frac{1}{2\eta}\mathscr{R}}, \tag{9}$$

where $C_{k,\eta,\phi_0,L} = \frac{8kL}{\eta} C_{\phi_0} \big(\frac{16\eta^2}{\pi}\big)^r$*, independent of* ρ*.*

IV. Proof of Theorem III.2

Due to the constraint of space, we only give a sketch of proof, omitting the details. Interested readers can refer to the full manuscript [15].

Lemma IV.1. *Given $r, m, \rho \in \mathbb{N}$ with $\eta = m/\rho \in \mathbb{N}$,*

$$D_\rho \bar{\Delta}_\rho^r = (\Delta^{(\eta)})^r D_\rho,$$

where $\Delta^{(\eta)}$ is the η-dimensional backward difference matrix.

Proposition IV.2. *Suppose $\eta \geq 3rk$, then $A_r\Phi_{m,k}$ is a frame with lower frame bound larger than $kC_{\phi_0}(\frac{2}{\pi})^r$.*

Lemma IV.3.

$$\|(A_r\Phi_{m,k})^*\Delta^r\|_{\infty,2} \leq 2^{2r+2}\eta^{r-1},$$

where the operator norm $\|S\|_{\infty,2} := \sup_{\|x\|_\infty=1}\|Sx\|_2$.

Proof. of Theorem III.2:

By Lemma IV.1,

$$\rho^r A_r q = D_\rho \bar{\Delta}_\rho^r \Delta^{-r} q = \Delta^r D_\rho(\Delta^{-r}q).$$

Since Δ and Δ^{-1} are lower-triangular, we see that, for any $1 \leq s \leq \eta$, there exist $\{a_j^s\}_{j=1}^s$ and $\{b_j^l\}_{j,l}$ such that

$$\begin{aligned}(A_r q)_s &= \sum_{j=1}^{s} a_j^s (D_\rho \Delta^{-r} q)_j \\ &= \sum_{j=1}^{s} a_j^s (\Delta^{-r} q)_{j\rho} \\ &= \sum_{j=1}^{s} a_j^s \sum_{l=1}^{j\rho} b_l^j q_l = \sum_{\xi=1}^{s\rho} c_\xi q_\xi,\end{aligned}$$

proving the first claim. The second assertion follows from Proposition IV.2.

Given $\Phi = \Phi_{m,k}$, $A = A_r = \frac{1}{\rho^r} D_\rho \bar{\Delta}_\rho^r \Delta^{-r}$, and $\mathcal{S} = (A\Phi)^* A\Phi$, the reconstruction error can be estimated as follows:

$$\begin{aligned}\mathscr{E} &= \|\mathcal{S}^{-1}(A\Phi)^* A q - x\|_2 \\ &= \|\mathcal{S}^{-1}(A\Phi)^* A \Delta^r u\|_2 \\ &= \frac{1}{\rho^r}\|\mathcal{S}^{-1}(A\Phi)^* \Delta^r D_\rho u\|_2 \\ &\leq \frac{1}{\rho^r}\|\mathcal{S}^{-1}\|_2 \|(A\Phi)^*\Delta^r\|_{\infty,2}\|D_\rho u\|_\infty \\ &\leq \left(\frac{4k}{\eta} C_{\phi_0}(\frac{8\eta}{\pi})^r\right)\|u\|_\infty \frac{1}{\rho^r},\end{aligned}$$

where the second inequality comes from Proposition IV.2 and Lemma IV.3.

□

V. ACKNOWLEDGEMENT

The author would like to thank the support from ARO Grant W911NF-17-1-0014 and J. Benedetto for all the thoughtful advice and insights.

REFERENCES

[1] John J Benedetto, Alexander M Powell, and Ozgur Yilmaz, *Sigma-delta quantization and finite frames*, IEEE Transactions on Information Theory **52** (2006), no. 5, 1990–2005.

[2] James Blum, Mark Lammers, Alexander M Powell, and Özgür Yılmaz, *Sobolev duals in frame theory and sigma-delta quantization*, Journal of Fourier Analysis and Applications **16** (2010), no. 3, 365–381.

[3] James Candy, *Decimation for sigma delta modulation*, vol. 34, IEEE transactions on communications, 1986.

[4] Evan Chou and C. Sinan Güntürk, *Distributed noise-shaping quantization: Ii. classical frames*, Excursions in Harmonic Analysis, Volume 5: The February Fourier Talks at the Norbert Wiener Center (2017), no. 179-198.

[5] Evan Chou and C.Sinan Güntürk, *Distributed noise-shaping quantization: I. beta duals of finite frames and near-optimal quantization of random measurements.*, Constructive Approximation **44** (2016), no. 1, 1–22.

[6] Wu Chou, Ping Wah Wong, and Robert M Gray, *Multistage sigma-delta modulation*, IEEE Transactions on Information theory **35** (1989), no. 4, 784–796.

[7] Ingrid Daubechies and Ron DeVore, *Approximating a bandlimited function using very coarsely quantized data: A family of stable sigma-delta modulators of arbitrary order*, Annals of mathematics **158** (2003), no. 2, 679–710.

[8] Ingrid Daubechies, Ronald A DeVore, C Sinan Gunturk, and Vinay A Vaishampayan, *A/d conversion with imperfect quantizers*, IEEE Transactions on Information Theory **52** (2006), no. 3, 874–885.

[9] Ingrid Daubechies and Rayan Saab, *A deterministic analysis of decimation for sigma-delta quantization of bandlimited functions*, IEEE Signal Processing Letters **22** (2015), no. 11, 2093–2096.

[10] Percy Deift, Felix Krahmer, and C Sınan Güntürk, *An optimal family of exponentially accurate one-bit sigma-delta quantization schemes*, Communications on Pure and Applied Mathematics **64** (2011), 883–919.

[11] PF Ferguson, A Ganesan, and RW Adams, *One bit higher order sigma-delta a/d converters*, IEEE International Symposium on Circuits and Systems (1990), 890–893.

[12] C Sinan Güntürk, *One-bit sigma-delta quantization with exponential accuracy*, Communications on Pure and Applied Mathematics **56** (2003), no. 11, 1608–1630.

[13] Hiroshi Inose and Yasuhiko Yasuda, *A unity bit coding method by negative feedback*, Proceedings of the IEEE **51** (1963), 1524–1535.

[14] Kung-Ching Lin, *Analysis of decimation on finite frames with sigma-delta quantization*, arXiv preprint arXiv:1803.02921 (2018).

[15] Kung-Ching Lin, *Adapted decimation on finite frames for arbitrary orders of sigma-delta quantization*, arXiv preprint arXiv:1902.05976 (2019).

[16] S Tewksbury and RW Hallock, *Oversampled, linear predictive and noise-shaping coders of order n¿ 1*, IEEE Transactions on Circuits and Systems **25** (1978), no. 7, 436–447.