ABSTRACT

Title of dissertation: SIGMA-DELTA QUANTIZATION:
NUMBER THEORETIC ASPECTS
OF REFINING QUANTIZATION ERROR

Aram Tangboondouangjit
Doctor of Philosophy, 2006

Dissertation directed by: Professor John J. Benedetto
Department of Mathematics

The linear reconstruction phase of analog-to-digital (A/D) conversion in signal processing is analyzed in quantizing finite frame expansions for $\mathbb{R}^d$. The specific setting is a K-level first order Sigma-Delta ($\Sigma\Delta$) quantization with step size δ. Based on basic analysis, the d-dimensional Euclidean 2-norm of quantization error of $\Sigma\Delta$ quantization with input of elements in $\mathbb{R}^d$ decays like $\mathcal{O}(1/N)$ as the frame size N approaches infinity; while the L^∞ norm of quantization error of $\Sigma\Delta$ quantization with input of bandlimited functions decays like $\mathcal{O}(T)$ as the sampling ratio T approaches zero. It has been, however, observed via numerical simulation that, with input of bandlimited functions, the mean square error norm of quantization error seems to decay like $\mathcal{O}(T^{3/2})$ as T approaches zero. Since the frame size N can be taken to correspond to the reciprocal of the sampling ratio T, this belief suggests that the corresponding behavior of quantization error, namely $\mathcal{O}(1/N^{3/2})$, holds in the setting of finite frame expansions in $\mathbb{R}^d$ as well. A number theoretic technique involving uniform distribution of sequences of real numbers and approxi-

mation of exponential sums is introduced to derive a better quantization error than $\mathcal{O}(1/N)$, $N \to \infty$. This estimate is signal dependent.

SIGMA-DELTA QUANTIZATION: NUMBER THEORETIC ASPECTS OF REFINING QUANTIZATION ERROR

by

Aram Tangboondouangjit

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2006

Advisory Commmittee:

Professor John J. Benedetto, Chair/Advisor
Professor Rama Chellappa
Professor Rebecca A. Herb
Professor Raymond L. Johnson
Professor Lawrence C. Washington

TABLE OF CONTENTS

Chapter 1

Introduction

1.1 Quantization of signals

In signal processing, transmitted signals in analog form need to be converted into digital form for storing, coding, and recovering purposes. This process of analog-to-digital (A/D) conversion consists of two main steps: *sampling* and *quantization*. In the sampling step, a given signal x is expressed as a linear combination over an at most countable dictionary $\{e_n\}_{n\in\Lambda}$ with real or complex coefficients, i.e., $x = \sum_{n\in\Lambda} x_n e_n$ ($x_n \in \mathbb{C}$ or $\mathbb{R}$). The expansion is said to be *redundant* if the choice of the coefficient sequence $\{x_n\}_{n\in\Lambda}$ is not unique. We shall refer to a coefficient sequence $\{x_n\}_{n\in\Lambda}$ as a sampling sequence. In order to be able to process the signal, one needs to reduce the continuous range of the sampling sequence consisting of real or complex numbers to a finite set. This step of signal processing is called *quantization*. More precisely, quantization is a mapping process with a map Q such that $Q : x \to \widetilde{x} = \sum_{n\in\Lambda} q_n e_n$, where, for each $n \in \Lambda$, q_n is an element from a finite set $\mathcal{A}$ called the *quantization alphabet*. The map Q is naturally called a *quantizer*. We see that Q replaces the sampling sequence $\{x_n\}_{n\in\Lambda}$ with $\{q_n\}_{n\in\Lambda}$ in a linear manner; so we refer to this manner of mapping as *linear reconstruction*. The natural question arises: how different is the new expansion $\widetilde{x} = \sum_{n\in\Lambda} q_n e_n$ from the signal x? This difference occurring in the quantization step is called *quantization error*, and it

is measured by computing $\|x - \widetilde{x}\|$, where $\|\cdot\|$ is a suitable norm in the space of signals. An optimal quantizer is the one that minimizes the quantization error norm. Nevertheless, finding a good quantizer has been proved to be a nontrivial, yet challenging, problem to the engineering community involved in signal processing.

For reasons of applicability, an audio signal f of interest is usually modelled as a bandlimited function. This means that f is an L^∞ function on $\mathbb{R}$ whose Fourier transform $\widehat{f}$ (as a distribution) is compactly supported. For each $0 < T < 1$, the function f can be reconstructed from the sampling sequence $\{f(nT)\}_{n\in\mathbb{Z}}$ as follows:

$$f(t) = T\sum_{n\in\mathbb{Z}} f(nT)g(t - nT), \tag{1.1.1}$$

where g is an appropriate smoothing kernel or sampling function. Applying a first order $\Sigma\Delta$ scheme on f yields a function $\widetilde{f}_T$ such that

$$\widetilde{f}_T(t) = T\sum_{n\in\mathbb{Z}} q_n^T g(t - nT), \tag{1.1.2}$$

where each $q_n^T \in \{-1, 1\}$. Standard analysis (see, e.g., [3],[7]) has shown that for some absolute constant $C > 0$,

$$\left\| f - \widetilde{f}_T \right\|_{L^\infty} \leq CT.$$

However, numerical experiments suggest a better bound than T. More precisely, it has been conjectured that there exists an absolute constant $C > 0$, independent of f, such that

$$\lim_{K\to\infty} \frac{1}{2K} \int_{|t|\leq K} \left| f(t) - \widetilde{f}_T(t) \right|^2 dt \leq CT^3. \tag{1.1.3}$$

This means the approximation error decays "on average" like $T^{3/2}$ [7]. We shall see later that the basic bound T corresponds to the basic bound $1/N$ in Euclidean norm

in the setting of finite frames for $\mathbb{R}^d$ where N is the frame size. This correspondence suggests that there should be a better bound for the setting of finite frames as well.

We shall assume that the signal of interest is an element of the Euclidean space $\mathbb{R}^d$, and that the sampling coefficients are real numbers. We shall also focus on structured dictionaries called *frames.*

1.2 Overview of the thesis and main results

We begin Chapter 2 by discussing material on frame theory. We discuss the definition of frames in Hilbert spaces and prove some properties of frames in this setting. Then we focus on finite frames for Euclidean space $\mathbb{R}^d$. Some interesting results dealing with finite unit norm tight frames are analyzed, based on the works by Benedetto and Fickus [14] and by Zimmermann [20]. We pay attention to a specific infinite family of frames called the harmonic frames. This family of frames provides substantive structure, and it is used in Chapter 5 to provide examples to illustrate the results on quantization error. The notion of the first order frame variation, $\sigma(F, p)$, is introduced, and it is generalized to define the nth order frame variation, $\sigma_n(F, p)$. We derive a general formula of $\sigma_n(F, p)$ for harmonic frames. We shall see that frame variation plays an important role in the basic quantization error as it relates the dependency of the $\Sigma\Delta$ scheme with the properties of frames.

In Chapter 3, we discuss a classic quantization scheme called Pulse Code Modulation (PCM) and derive quantization error estimate associated to this scheme for finite frames for $\mathbb{R}^d$. We then provide the setting of this thesis, viz., the first order

K-level $\Sigma\Delta$ scheme with step size δ. The quantizer map is defined algorithmically. However, this makes it inconvenient to program numerical experiments using MATLAB so we derive the general formula for this quantizer. Then we derive a basic quantization error estimate based on the $\Sigma\Delta$ scheme. This is done in [16], where it is proved that if F is a unit norm tight frame for $\mathbb{R}^d$ of cardinality $N \geq d$, then the K-level $\Sigma\Delta$ scheme with quantization step size δ gives quantization error

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

where x is a given signal, $\widetilde{x}$ is the quantized signal, and $\|\cdot\|$ is the d-dimensional Euclidean 2-norm.

In Chapter 4, we first provide the background material from the theory of uniform distribution of sequences of real numbers [17]. In particular, we define uniform distribution modulo 1, and state examples of real sequences with this property. We then discuss the notion of discrepancy of a finite sequence and prove some basic results on the bound of discrepancy. We provide two inequalities that improve the bound of discrepancy and emphasize one of them, viz., the Erdös-Turán Inequality which states the following: For any finite sequence $x_1, \ldots, x_N$ of real numbers and any positive integer m, we have

$$D_N \leq \frac{6}{m+1} + \frac{4}{\pi} \sum_{h=1}^{m} \left(\frac{1}{h} - \frac{1}{m+1} \right) \left| \frac{1}{N} \sum_{n=1}^{N} e^{2\pi i h x_n} \right|,$$

where D_N is the discrepancy of the sequence $x_1, \ldots, x_N$. This inequality plays an important role in our analysis of quantization error as it approximates discrepancy in terms of an exponential sum which will be approximated further by a theorem of van der Corput. This latter theorem states the following: If a and b are integers

with $a < b$, and if f is a twice differentiable function on $[a, b]$ with $f''(x) \geq \rho > 0$ for all $x \in [a, b]$ or $f''(x) \leq -\rho < 0$ for all $x \in [a, b]$, then

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

In Chapter 5, we collect all the ingredients to give the detailed proof of the theorem on improving the quantization error stated in [16]. We provide a new construction which corrects errors observed in the original proof. One ingredient we need in the proof is a result by Güntürk [7, 8]. This theorem allows us to construct an analytic function with certain properties such that the values at the natural numbers correspond to the terms of a given real sequence. We prove the special case of this theorem, and give an explicit bound for the inequality not given in the original theorem. The quantization error obtained in this chapter is an improvement from the basic error estimate obtained in Chapter 3. In fact, our improvement goes from order $1/N$ to one of order $1/N^{5/4-\epsilon}$ $(\epsilon > 0)$ for certain choices of frames, where N denotes the cardinality of frame. We show further that with a certain natural assumption, the order of the quantization error estimate can be improved to $1/N^{4/3}$. This order is better than the order obtained by Güntürk in [7] in the setting of bandlimited functions. There he obtained a bound of order $1/N^{4/3-\epsilon}$ for $\epsilon > 0$. On the other hand, the bounds we obtain for these improved estimates depend on the given signals. One of the main goals of our future research is to dispense with this restriction.

The last section of Chapter 5 is devoted to examples to justify the results of the theorems we have proved. We show various graphs of quantization error norms

which are plotted against the cardinality of the frames. We analyze some interesting phenomena concerning the periodic pattern occurring in the shapes of these graphs.

1.3 New results

In this section we specifically describe our own contributions.

- In Chapter 2, we generalize the notion of the first order frame variation $\sigma(F, p_N)$ to the nth order frame variation $\sigma_n(F, p_N)$, where F is a given frame and p_N is a permutation of the set $\{1, \ldots, N\}$. We then prove the explicit formulae of $\sigma_n(H_N^d, p)$ for the harmonic frame H_N^d with respect to the identity permutation p (Theorem 2.4.5). Such formulae can be used in refined quantization error estimates. We also prove a result (Theorem 2.4.6), which is a consequence of the proof of Theorem 2.4.5, which gives relatively sharp inequalities for some new trigonometric binomial sums.

- In Theorem 3.3.1 of Chapter 3, we prove the general formula of the quantizer associated with first order $\Sigma\Delta$ quantization. Theorem 3.3.1 is crucial in programming numerical experiments using MATLAB.

- In Chapters 2 and 4, we give details for difficult issues concerning frames, uniform distribution, and discrepancy, which are not readily available in the literature. For example see Proposition 2.3.6, Examples 4.1.2, 4.2.8, 4.4.9, Theorem 4.2.5. In particular, we proved in Example 4.1.2 that the following

sequence is u.d. mod 1:

$$\frac{0}{1}, \frac{0}{2}, \frac{1}{2}, \frac{0}{3}, \frac{1}{3}, \frac{2}{3}, \ldots, \frac{0}{n}, \frac{1}{n}, \ldots, \frac{n-1}{n}, \ldots.$$

- In Chapter 5, noting that there was a gap in the original proof of Güntürk's theorem in [7] we provided a complete proof for an important special case. Independently, Güntürk has given a complete proof in [8] and in a private communication, the latter after seeing our work. We also compute an explicit bound of the inequality occurring in the theorem, which will be useful in evaluating quantization error independent of signal. We also correct the proof of Theorem 5.1.1 from the original one by providing a new intricate construction. Finally, we have constructed a new class of examples of quantization error plots, showing and giving preliminary analysis of various periodic patterns of the shape of graphs of the quantization error as a function of the frame size.

1.4 Definitions and notation

We shall use the following definitions and notation.

- The Fourier transform is formally defined by

$$\widehat{f}(\gamma) = \int f(t) e^{-2\pi i t\gamma} dt.$$

- We denote the characteristic function of a set E by $\mathbb{1}_E$, i.e.,

$$\mathbb{1}_E(x) = \begin{cases} 1 & \text{if } x \in E, \\ 0 & \text{otherwise.} \end{cases}$$

- For $x \in \mathbb{R}$, we denote by $\lfloor x \rfloor$ the floor function of x, which is the largest integer that is not greater than x; and we denote by $\{x\}$ the fractional part $x - \lfloor x \rfloor \in [0, 1)$ of x. We also denote by $\lceil x \rceil$ the ceiling function of x, which is the smallest integer that is at least x.

Chapter 2

Frame Theory

2.1 Overview

The necessary condition for a sequence $\{e_n\}_{n=1}^{\infty}$ of unit norm vectors to be an orthonormal basis (ONB) for a Hilbert space $\mathcal{H}$ is that it satisfies Parseval's equation, that is,

$$\sum_{n=1}^{\infty} |\langle x, e_n \rangle|^2 = \|x\|^2 \text{ for all } x \in \mathcal{H}. \tag{2.1.1}$$

A relaxation of this condition (specified later) leads to a generalization of the notion of ONB, namely *frames*. If a sequence $\{e_n\}_{n=1}^{\infty}$ of vectors is a frame for a Hilbert space $\mathcal{H}$ then it spans $\mathcal{H}$ and yet it is not necessarily linearly independent. In other words, in the case of frames, for each $x \in \mathcal{H}$, there exists a sequence $\{x_n\}_{n=1}^{\infty}$ of real or complex numbers such that

$$x = \sum_{n=1}^{\infty} x_n e_n. \tag{2.1.2}$$

Because the frame elements are allowed to be linearly dependent, the coefficients $\{x_n\}_{n=1}^{\infty}$ are not necessarily unique. We usually referred to this property as the *redundancy* of frames and it is one of the main reasons why frames have been extensively used in signal processing. The notion of frames was introduced by Duffin and Schaeffer in their 1952 paper [18]. The main subject of their study is *non-harmonic Fourier series*, i.e., sequences of the type $\{e^{i\lambda_n x}\}_{n\in\mathbb{Z}}$, where $\{\lambda_n\}_{n\in\mathbb{Z}}$ is

a family of real or complex numbers satisfying a *uniform density* condition. However, the potential of frames was not realized until 34 years later during the era of wavelet theory, in a paper by Daubechies, Grossman and Meyer [19] (1986). Using frames, they expanded functions $f \in L^2(\mathbb{R})$ in a similar manner as using orthonormal bases. The mathematical framework of signal processing was set rigorously by assuming the signal of interest referred by the authors as "incoming information" to be an element of a Hilbert space $\mathcal{H}$, particularly of $\mathcal{H} = L^2(\mathbb{R})$. Parts of the following materials on frames in Hilbert spaces are adapted from Chapters 3 and 5 of Christensen's book [11].

2.2 Bessel sequences

Definition 2.2.1. A sequence $\{e_n\}_{n=1}^{\infty}$ in $\mathcal{H}$ is said to be a *Bessel sequence* if there exists a constant $B > 0$ such that

$$\forall x \in \mathcal{H}, \qquad \sum_{n=1}^{\infty} |\langle x, e_n \rangle|^2 \leq B \|x\|^2 . \tag{2.2.1}$$

A number B satisfying condition (2.2.1) is called a *Bessel bound* for $\{e_n\}_{n=1}^{\infty}$.

Lemma 2.2.2. *Let $\{e_n\}_{n=1}^{\infty}$ be a Bessel sequence in a Hilbert space $\mathcal{H}$. Define the associated* Bessel map $L : \mathcal{H} \to \ell^2(\mathbb{N})$ *by*

$$x \mapsto \{\langle x, e_n \rangle\}_{n=1}^{\infty} .$$

Then L is a bounded (continuous) linear operator. Moreover, the corresponding adjoint operator $L^* : \ell^2(\mathbb{N}) \to \mathcal{H}$ is given by

$$\{a_n\}_{n=1}^{\infty} \mapsto \sum_{n=1}^{\infty} a_n e_n .$$

Proof. We see that L is well defined since $\{e_n\}_{n=1}^{\infty}$ is a Bessel sequence. Let B be a Bessel bound for the sequence $\{e_n\}_{n=1}^{\infty}$. Then for each $x \in \mathcal{H}$,

$$\|Lx\|^2 = \sum_{n=1}^{\infty} |\langle x, e_n\rangle|^2 \leq B \|x\|^2 .$$

So $\|Lx\| \leq \sqrt{B}\|x\|$. This shows that L is bounded. Let $c \in \ell^2(\mathbb{N})$ and define $S_N = \sum_{n=1}^{N} c[n]e_n$ for each $N \in \mathbb{N}$. Then for all integers N, M with $N > M$,

$$\begin{aligned} \|S_N - S_M\|^2 &= \sup_{\|x\|=1} \Big| \sum_{n=M+1}^{N} c[n]\langle e_n, x\rangle \Big|^2 \\ &\leq \sup_{\|x\|=1} \Big(\sum_{n=M+1}^{N} |c[n]|^2 \Big) \Big(\sum_{n=M+1}^{N} |\langle e_n, x\rangle|^2 \Big) \\ &\leq B \sum_{n=M+1}^{N} |c[n]|^2 . \end{aligned}$$

The first equality is an equivalent way of expressing the norm in a Hilbert space, see Remark 2.2.3. The second inequality follows from the Hölder Inequality. Now, since $c \in \ell^2(\mathbb{N})$, it follows that the sequence $\left\{ \sum_{k=1}^{n} |c[k]|^2 \right\}_{n=1}^{\infty}$ is Cauchy. We therefore see from the above calculation that the sequence $\{S_n\}_{n=1}^{\infty}$ is Cauchy, and hence converges in $\mathcal{H}$. To find the formula for the adjoint operator L^*, we let $c \in \ell^2(\mathbb{N})$, and let $x \in \mathcal{H}$. Then

$$\begin{aligned} \langle x, L^*c\rangle = \langle Lx, c\rangle &= \sum_{n=1}^{\infty} (Lx)[n]\overline{c[n]} = \sum_{n=1}^{\infty} \langle x, e_n\rangle \overline{c[n]} \\ &= \sum_{n=1}^{\infty} \langle x, c[n]e_n\rangle = \langle x, \sum_{n=1}^{\infty} c[n]e_n\rangle. \end{aligned} \tag{2.2.2}$$

The last equality follows from the continuity of the inner product, see Remark 2.2.3. Since this is true for all $x \in \mathcal{H}$, it follows from the Hahn-Banach Theorem that

$$L^*c = \sum_{n=1}^{\infty} c[n]e_n. \qquad \square$$

Remark 2.2.3. In the proof of Lemma 2.2.2, we have used an equivalent way of expressing the norm in a Hilbert space. This can be shown as follows. Let $y \in \mathcal{H}$. Then the map ψ^y defined by $\psi^y x = \langle x, y \rangle$ is a bounded linear operator. From Cauchy-Schwarz Inequality we have $|\psi^y x| = |\langle x, y \rangle| \leq \|y\| \, \|x\|$. Since the equality holds if and only if $x = ay$ for some scalar a, we see that $\|\psi^y\| = \|y\|$. Since $\|\psi^y\| = \sup_{\|x\|=1} |\psi^y x| = \sup_{\|x\|=1} |\langle x, y \rangle|$, it follows that $\|y\| = \sup_{\|x\|=1} |\langle x, y \rangle|$. The fact that ψ^y is bounded, and therefore continuous, allows the final equality in (2.2.2).

We note that the proof of Lemma 2.2.2 remains the same if the order of the sequence $\{e_n\}_{n=1}^{\infty}$ has been changed. Hence we have the following corollary.

Corollary 2.2.4. *If $\{e_n\}_{n=1}^{\infty}$ is a Bessel sequence in $\mathcal{H}$, then $\sum_{n=1}^{\infty} c[n]e_n$ converges unconditionally for all $c \in \ell^2(\mathbb{N})$.*

By Corollary 2.2.4 we see that it does not matter what index set we use to index the series $\sum_{n=1}^{\infty} c[n]e_n$ since each reordering of the sequence $\{c[n]e_n\}_{n=1}^{\infty}$ will have the series converge to the same element. Hence we can use natural numbers as the standard index set.

2.3 Frames in Hilbert spaces

We are now in a position to state the definition of frames.

Definition 2.3.1. A sequence $\{e_n\}_{n=1}^{\infty}$ of elements in a Hilbert space $\mathcal{H}$ is said to

be a frame for $\mathcal{H}$ if there exist constants $0 < A \leq B < \infty$ such that

$$A \|x\|^2 \leq \sum_{n=1}^{\infty} |\langle x, e_n \rangle|^2 \leq B \|x\|^2 \quad \text{for all } x \in \mathcal{H}. \tag{2.3.1}$$

The numbers A and B are called *frame bounds*. The *optimal upper frame bound* is the infimum over all upper frame bounds and the *optimal lower frame bound* is the supremum over all lower frame bounds. We note that the optimal bounds are actually frame bounds. A frame is said to be *A-tight* if $A = B$, and is said to be *unit norm* if $\|e_n\| = 1$ for all n. A frame is said to be *exact* if it ceases to be a frame when one of the elements is removed from the sequence $\{e_n\}_{n=1}^{\infty}$.

Some basic examples of frames are as follows:

Example 2.3.2. Let $\{e_n\}_{n=1}^{\infty}$ be an orthonormal basis for $\mathcal{H}$.

(i) By repeating each element in $\{e_n\}_{n=1}^{\infty}$ twice, we obtain

$$\{f_n\}_{n=1}^{\infty} = \{e_1, e_1, e_2, e_2, \dots\}$$

which is a 2-tight frame. In fact, for each $x \in \mathcal{H}$ we have

$$\begin{aligned} \sum_{n=1}^{\infty} |\langle x, f_n \rangle|^2 &= \sum_{n=1}^{\infty} |\langle x, e_n \rangle|^2 + \sum_{n=1}^{\infty} |\langle x, e_n \rangle|^2 \\ &= \|x\|^2 + \|x\|^2 = 2 \|x\|^2 . \end{aligned}$$

(ii) By repeating only e_1, we obtain

$$\{f_n\}_{n=1}^{\infty} = \{e_1, e_1, e_2, e_3, \dots\}$$

which is a frame with frame bounds $A = 1$ and $B = 2$. In fact, for each $x \in \mathcal{H}$

we have

$$\begin{aligned}\|x\|^2 = \sum_{n=1}^{\infty} |\langle x, e_n\rangle|^2 &\leq |\langle x, e_1\rangle|^2 + \sum_{n=1}^{\infty} |\langle x, e_n\rangle|^2 \\ &= \sum_{n=1}^{\infty} |\langle x, f_n\rangle|^2 \\ &\leq \sum_{n=1}^{\infty} |\langle x, e_n\rangle|^2 + \sum_{n=1}^{\infty} |\langle x, e_n\rangle|^2 \\ &= \|x\|^2 + \|x\|^2 = 2\|x\|^2 .\end{aligned}$$

(iii) Let

$$\{f_n\}_{n=1}^{\infty} = \left\{e_1, \frac{1}{\sqrt{2}}e_2, \frac{1}{\sqrt{2}}e_2, \frac{1}{\sqrt{3}}e_3, \frac{1}{\sqrt{3}}e_3, \frac{1}{\sqrt{3}}e_3, \dots\right\}.$$

This is the sequence where each vector $\frac{1}{\sqrt{n}}e_n$ is repeated n times. As such, it is a 1-tight frame. In fact, for each $x \in \mathcal{H}$ we have

$$\sum_{n=1}^{\infty} |\langle x, f_n\rangle|^2 = \sum_{n=1}^{\infty} n \left|\langle x, \frac{1}{\sqrt{n}}e_n\rangle\right|^2 = \|x\|^2 .$$

(iv) Let $v_1 = (1, 0), v_2 = (-2/\sqrt{5}, 2), v_3 = (4/\sqrt{5}, 1)$. By a direct computation, one can show that $\{v_1, v_2, v_3\}$ is a 5-tight frame for $\mathbb{R}^2$. In fact, letting $v = (a, b)$ be a vector in $\mathbb{R}^2$, we have

$$\sum_{n=1}^{3} |\langle v, v_n\rangle|^2 = a^2 + (-\frac{2}{\sqrt{5}}a + 2b)^2 + (\frac{4}{\sqrt{5}}a + b)^2 = 5(a^2 + b^2) = 5\|v\|^2 .$$

Let $\{e_n\}_{n=1}^{\infty}$ be a frame for a Hilbert space $\mathcal{H}$. We define an operator $S : \mathcal{H} \to \mathcal{H}$ by

$$Sx = L^*Lx = \sum_{n=1}^{\infty} \langle x, e_n\rangle e_n.$$

We see that since $\{e_n\}_{n=1}^{\infty}$ is a Bessel sequence, Corollary 2.2.4 implies that S is a well-defined operator. The operator S is called the *frame operator* for $\{e_n\}_{n=1}^{\infty}$. We prove some properties of the frame operator S in the following lemma.

Lemma 2.3.3. *Let $\{e_n\}_{n=1}^{\infty}$ be a frame with frame bounds A, B. Then the following hold:*

(i) The frame operator S is bounded, invertible, self-adjoint, and positive.

(ii) The sequence $\{S^{-1}e_n\}_{n=1}^{\infty}$ is a frame with bounds B^{-1} and A^{-1}; if A, B are the optimal bounds for $\{e_n\}_{n=1}^{\infty}$, then the bounds B^{-1}, A^{-1} are optimal for $\{S^{-1}e_n\}_{n=1}^{\infty}$. The frame operator for $\{S^{-1}e_n\}_{n=1}^{\infty}$ is S^{-1}.

The sequence $\{S^{-1}e_n\}_{n=1}^{\infty}$ is called the *(canonical) dual frame* of $\{e_n\}_{n=1}^{\infty}$. Before proving the lemma, we state two classical results from operator theory.

Lemma 2.3.4 (Neumann Theorem). *Let X be a Banach space and let $U : X \to X$ be a bounded operator. If $\|I - U\| < 1$, then U is invertible with*

$$U^{-1} = \sum_{n=1}^{\infty} (I - U)^n.$$

Furthermore, $\|U^{-1}\| \leq (1 - \|I - U\|)^{-1}$.

Lemma 2.3.5. *Let $\mathcal{H}$ be a Hilbert space and let $U_j : H \to H\,(j = 1, 2, 3)$ be self-adjoint operators with $U_3 \geq 0$. If $U_1 \leq U_2$ and U_3 commutes with U_1 and U_2, then $U_1U_3 \leq U_2U_3$. (By definition, two self-adjoint operators $U \leq W$ if $\langle Ux, x\rangle \leq \langle Wx, x\rangle$ for all $x \in \mathcal{H}$.)*

We are now ready to prove Lemma 2.3.3.

Proof of Lemma 2.3.3. (i) Since L and L^* are bounded operator, the frame operator S being the composition of these two operators is also bounded. Now since $S^* = (L^*L)^* = L^*(L^*)^* = L^*L = S$, the operator S is self-adjoint. By direct calculation

we see that for each $x \in \mathcal{H}, \langle Sx, x\rangle = \sum_{n=1}^{\infty} |\langle x, e_n\rangle|^2$. So we can rewrite the frame condition (2.3.1) in terms of S as

$$AI \leq S \leq BI \tag{2.3.2}$$

This shows that for each $x \in \mathcal{H}, \langle Sx, x\rangle \geq A\|x\|^2 \geq 0$. So S is positive. By subtracting BI and multiplying by B^{-1} through the inequality (2.3.2), we obtain that $0 \leq I - B^{-1}S \leq \frac{B-A}{B} I$. Therefore

$$\left\|I - B^{-1}S\right\| = \sup_{\|x\|=1} \left|\langle (I - B^{-1}S)x, x\rangle\right| \leq \frac{B-A}{B} < 1,$$

which, by Lemma 2.3.4, shows that S is invertible.

(ii) We first show that $\{S^{-1}e_n\}_{n=1}^{\infty}$ is a Bessel sequence. Indeed, for each $x \in \mathcal{H}$,

$$\sum_{n=1}^{\infty} \left|\langle x, S^{-1}e_n\rangle\right|^2 = \sum_{n=1}^{\infty} \left|\langle S^{-1}x, e_n\rangle\right|^2 \leq B\left\|S^{-1}x\right\|^2 \leq B\left\|S^{-1}\right\|^2 \|x\|^2.$$

Hence the frame operator for $\{S^{-1}e_n\}_{n=1}^{\infty}$ is well defined. This frame operator acts on $x \in \mathcal{H}$ by

$$\sum_{n=1}^{\infty} \langle x, S^{-1}e_n\rangle S^{-1}e_n = S^{-1}\sum_{n=1}^{\infty} \langle S^{-1}x, e_n\rangle e_n = S^{-1}SS^{-1}x = S^{-1}x.$$

This shows that the frame operator of $\{S^{-1}e_n\}_{n=1}^{\infty}$ is S^{-1}. Now since the operator S^{-1} commutes with both S and I, we can apply Lemma 2.3.5 and obtain that, upon multiplying the inequality (2.3.2) with S^{-1},

$$B^{-1}I \leq S^{-1} \leq A^{-1}I.$$

This means for all $x \in \mathcal{H}$,

$$B^{-1}\|x\|^2 \leq \langle S^{-1}x, x\rangle = \sum_{n=1}^{\infty} \left|\langle x, S^{-1}e_n\rangle\right|^2 \leq A^{-1}\|x\|^2.$$

Thus $\{S^{-1}e_n\}_{n=1}^{\infty}$ is a frame for $\mathcal{H}$ with frame bounds B^{-1} and A^{-1}. Now suppose that A, B are optimal bounds for the frame $\{e_n\}_{n=1}^{\infty}$. Let C be the optimal upper bound for the frame $\{S^{-1}e_n\}_{n=1}^{\infty}$ and assume that $C < 1/A$. Then since S^{-1} is the frame operator for $\{S^{-1}e_n\}_{n=1}^{\infty}$ it follows that the frame $\{(S^{-1})^{-1}S^{-1}e_n\}_{n=1}^{\infty} = \{e_n\}_{n=1}^{\infty}$ has lower bound $1/C > A$. This is a contradiction since A is the optimal lower bound for $\{e_n\}_{n=1}^{\infty}$. Hence $C = 1/A$. We can show similarly that the optimal lower bound for $\{S^{-1}e_n\}_{n=1}^{\infty}$ is $1/B$. □

Proposition 2.3.6. *Let $\{e_n\}_{n=1}^{\infty}$ be a frame for a Hilbert space $\mathcal{H}$ with frame bounds A, B and with frame operator S. Then the following inequalities hold:*

$$A\|x\| \leq \|Sx\| \leq B\|x\| \text{ for all } x \in \mathcal{H}.$$

Proof. Let $x \in \mathcal{H}$. We shall prove the leftmost inequality first. By definition of operator S, we have $\langle Sx, x\rangle = \sum_{n=1}^{\infty} |\langle x, e_n\rangle|^2$. It follows from the Cauchy-Schwarz Inequality and the frame condition (2.3.1) that

$$\Big(\sum_{n=1}^{\infty} |\langle x, e_n\rangle|^2\Big)^2 = \langle Sx, x\rangle^2 \leq \|Sx\|^2 \|x\|^2 \leq \|Sx\|^2 \frac{1}{A} \sum_{n=1}^{\infty} |\langle x, e_n\rangle|^2.$$

This implies

$$A\sum_{n=1}^{\infty} |\langle x, e_n\rangle|^2 \leq \|Sx\|^2.$$

From the frame condition (2.3.1) we have $\sum_{n=1}^{\infty} |\langle x, e_n\rangle|^2 \geq A\|x\|^2$, and so

$$A^2\|x\|^2 \leq \|Sx\|^2.$$

Hence the leftmost inequality follows. Now we show the rightmost inequality. Let $c \in \ell^2(\mathbb{N})$ and recall that $L^*c = \sum_{n=1}^{\infty} c[n]e_n$. By an equivalent definition of norm in

a Hilbert space, see Remark 2.2.3 and the Hölder Inequality, it follows that

$$\begin{aligned}
\|L^*c\| &= \sup_{\|y\|=1} |\langle L^*c, y\rangle| = \sup_{\|y\|=1} \left| \langle \sum_{n=1}^{\infty} c[n]e_n, y\rangle \right| \\
&= \sup_{\|y\|=1} \left| \sum_{n=1}^{\infty} c[n]\langle e_n, y\rangle \right| \leq \sup_{\|y\|=1} \sum_{n=1}^{\infty} |c[n]\langle e_n, y\rangle| \\
&\leq \Big(\sum_{n=1}^{\infty} |c[n]|^2 \Big)^{1/2} \sup_{\|y\|=1} \Big(\sum_{n=1}^{\infty} |\langle e_n, y\rangle|^2 \Big)^{1/2} \\
&\leq \sqrt{B}\, \|c\| .
\end{aligned}$$

Hence $\|L^*\| \leq \sqrt{B}$. Now since $S = L^*L$ it follows from a property of the adjoint operator that

$$\|S\| = \|L^*L\| = \|L^*\|^2 \leq B.$$

Thus

$$\|Sx\| \leq \|S\|\, \|x\| \leq B\, \|x\| ,$$

which is the rightmost inequality and hence the proof is complete. □

Now we arrive at the main elementary theorem in frame theory. All applications of frames start with this so-called *frame decomposition* which shows that every element in a Hilbert space can be represented as an infinite linear combination of the frame elements.

Theorem 2.3.7 (Frame Decomposition). *Let $\{e_n\}_{n=1}^{\infty}$ be a frame for a Hilbert space $\mathcal{H}$ with corresponding frame operator S. Then*

$$x = \sum_{n=1}^{\infty} \langle x, S^{-1}e_n\rangle e_n = \sum_{n=1}^{\infty} \langle x, e_n\rangle S^{-1}e_n \quad \text{for all } x \in \mathcal{H}. \tag{2.3.3}$$

Both of the series converge unconditionally for all $x \in \mathcal{H}$.

Proof. Let $x \in \mathcal{H}$. Then we have from properties of the frame operator in Lemma 2.3.3 that

$$x = SS^{-1}x = \sum_{n=1}^{\infty} \langle S^{-1}x, e_n \rangle e_n = \sum_{n=1}^{\infty} \langle x, S^{-1}e_n \rangle e_n.$$

The last equality follows from the fact that S^{-1} is self-adjoint. Now since $\{e_n\}_{n=1}^{\infty}$ is a Bessel sequence and $\{\langle x, S^{-1}e_n \rangle\}_{n=1}^{\infty} \in \ell^2(\mathbb{N})$, it follows from Corollary 2.2.4 that the series converges unconditionally. Similarly, by composing S^{-1} with S we have another way to represent the element x, that is,

$$x = S^{-1}Sx = \sum_{n=1}^{\infty} \langle Sx, S^{-1}e_n \rangle S^{-1}e_n = \sum_{n=1}^{\infty} \langle x, SS^{-1}e_n \rangle S^{-1}e_n = \sum_{n=1}^{\infty} \langle x, e_n \rangle S^{-1}e_n.$$

The penultimate equality follows from the fact that S is self-adjoint. Since $\{S^{-1}e_n\}_{n=1}^{\infty}$ is a Bessel sequence and $\{\langle x, e_n \rangle\}_{n=1}^{\infty} \in \ell^2(\mathbb{N})$, it follows from Corollary 2.2.4 that the series converges unconditionally. Hence the proof is complete. □

2.4 Harmonic frames for $\mathbb{R}^d$

The atomic decompositions in (2.3.3) are the first step towards a digital representation. If the frame is tight with frame bound A, then from (2.3.2) we have the frame operator $S = AI$, and therefore we see that both of the frame expansions in (2.3.3) are equivalent, i.e., for each $x \in \mathcal{H}$,

$$x = \frac{1}{A} \sum_{n=1}^{\infty} \langle x, e_n \rangle e_n.$$

For convenience, we let $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$. When the Hilbert space $\mathcal{H}$ is $\mathbb{K}^d$ and the cardinality of frame is finite, the frame is referred to as a *finite frame* for $\mathcal{H}$. In this case, there is a systematic method to check whether an arbitrary finite set

of vectors is a tight frame. Let $\{v_n\}_{n=1}^N$ be a set of N vectors in $\mathbb{K}^d$ We define the associated matrix L to be the $N \times d$ matrix whose rows are the $\overline{v}_n$. The following lemma, found in [20], allows us to determine whether $\{v_n\}_{n=1}^N$ forms a tight frame for $\mathbb{K}^d$.

Lemma 2.4.1. *A set of vectors $\{v_n\}_{n=1}^N$ in $\mathbb{K}^d$ is a tight frame with frame bound A if and only if its associated matrix L satisfies*

$$L^*L = AI_d,$$

where L^* is the conjugate transpose of L, and I_d is the $d \times d$ identity matrix. Moreover the frame $\{v_n\}_{n=1}^N$ is unit norm if and only if the diagonal of LL^* equals (1,. . . ,1).

Proof. Let $x = (x_1, \ldots, x_d) \in \mathbb{K}^d$. Then by a straightforward calculation, we obtain

$$Lx = (\langle x, v_1\rangle, \ldots, \langle x, v_N\rangle) \tag{2.4.1}$$

From (2.4.1) we obtain

$$\sum_{n=1}^{N} |\langle x, v_n\rangle|^2 = (Lx)^* \cdot (Lx) = x^*(L^*L)x. \tag{2.4.2}$$

A set $\{v_n\}_{n=1}^N$ is an A-tight frame for $\mathbb{K}^d$ if and only if $\sum_{n=1}^N |\langle x, v_n\rangle|^2 = A\|x\|^2 = x^*(AI_d)x$ for all $x \in \mathbb{K}^d$. From (2.4.2) this is true if and only if $x^*(AI_d)x = x^*(L^*L)x$ for all $x \in \mathbb{K}^d$; and this in turn is true if and only if $AI_d = L^*L$. To prove the second part we observe that the frame $\{v_n\}_{n=1}^N$ is unit norm if and only if $1 = \|v_n\|^2 = \langle v_n, v_n\rangle = \sum_{j=1}^d v_n(j)\overline{v_n(j)}$ for each $1 \le n \le N$. We see that the last sum is exactly the nth diagonal element of the matrix LL^* for each $1 \le n \le N$. Hence the result follows. □

The following lemma determines the frame bound for a finite unit norm tight frame in $\mathbb{K}^d$. The first proof can be found in [20] where the author uses matrix properties, and the second in [14] where the authors use the definition of an orthonormal basis.

Lemma 2.4.2. *A unit norm tight frame for $\mathbb{K}^d$ with N elements has frame bound $A = N/d$.*

First proof. We denote the trace of a matrix M by $\mathrm{Tr}(M)$. It is straightforward to show that $\mathrm{Tr}(MN) = \mathrm{Tr}(NM)$ for all matrices M, N that can be multiplied. Using this property and Lemma 2.4.1, we have

$$A = \frac{1}{d}\mathrm{Tr}(L^*L) = \frac{1}{d}\mathrm{Tr}(LL^*) = \frac{1}{d}N.$$

Second proof. Let $\{v_n\}_{n=1}^N$ be a unit norm tight frame for $\mathbb{K}^d$ with frame bound A. Let $\{e_j\}_{j=1}^d$ be an orthonormal basis for $\mathbb{K}^d$. Then

$$Ad = \sum_{j=1}^{d} A\left\|e_n\right\|^2 = \sum_{j=1}^{d}\sum_{n=1}^{N}\left|\langle e_j, v_n\rangle\right|^2 = \sum_{n=1}^{N}\sum_{j=1}^{d}\left|\langle e_j, v_n\rangle\right|^2 = \sum_{n=1}^{N}\left\|v_n\right\|^2 = N. \quad \square$$

Now we introduce harmonic frames for $\mathbb{R}^d$. This family of frames has a Fourier-based structure, and it provides good examples that we shall use later in Chapter 5. The definition of the harmonic frame $H_N^d = \{e_n\}_{n=0}^{N-1}$, $N > d$, depends on whether the dimension d is even or odd.

If $d \geq 2$ is even, let

$$e_n = \sqrt{\frac{2}{d}}\left[\cos\frac{2\pi n}{N}, \sin\frac{2\pi n}{N}, \cos\frac{2\pi 2n}{N}, \sin\frac{2\pi 2n}{N}, \ldots, \cos\frac{2\pi\frac{d}{2}n}{N}, \sin\frac{2\pi\frac{d}{2}n}{N}\right] \quad (2.4.3)$$

for $n = 0, 1, \ldots, N-1$.

If $d > 1$ is odd, let

$$e_n = \sqrt{\frac{2}{d}}\left[\frac{1}{\sqrt{2}}, \cos\frac{2\pi n}{N}, \sin\frac{2\pi n}{N}, \cos\frac{2\pi 2n}{N}, \sin\frac{2\pi 2n}{N}, \ldots, \cos\frac{2\pi\frac{d-1}{2}n}{N}, \sin\frac{2\pi\frac{d-1}{2}n}{N}\right] \tag{2.4.4}$$

for $n = 0, 1, \ldots, N-1$.

We shall now show that H_N^d, as defined above, is a unit norm tight frame for $\mathbb{R}^d$. From the identity $\cos^2\theta + \sin^2\theta = 1$, it follows immediately that e_n is unit norm for each $0 \le n \le N-1$. To verify that H_N^d is a tight frame, we have options either to apply Lemma 2.4.1 or to verify the definition directly. In this case it turns out that the latter option is easier. We verify only the case when d is even. The case when d is odd will be similar. So let d be even, let $N > d$, and take $x = (a_1, b_1, \ldots, a_{\frac{d}{2}}, b_{\frac{d}{2}}) \in \mathbb{R}^d$. We want to show that

$$\sum_{n=0}^{N-1} |\langle x, e_n\rangle|^2 = \frac{N}{d}\|x\|^2. \tag{2.4.5}$$

We have

$$\begin{aligned}
\frac{d}{2}\sum_{n=0}^{N-1}|\langle x, e_n\rangle|^2 &= \sum_{n=0}^{N-1}\left(\sum_{j=1}^{d/2} a_j\cos\frac{2\pi nj}{N} + b_j\sin\frac{2\pi nj}{N}\right)^2 \\
&= \sum_{n=0}^{N-1}\left(\sum_{j=1}^{d/2}\sqrt{a_j^2+b_j^2}\sin\left(\frac{2\pi nj}{N}+\phi_j\right)\right)^2 \\
&= \sum_{n=0}^{N-1}\sum_{j=1}^{d/2}(a_j^2+b_j^2)\sin^2\left(\frac{2\pi nj}{N}+\phi_j\right) \\
&\quad + \sum_{n=0}^{N-1}\sum_{j\neq k}\sqrt{a_j^2+b_j^2}\sqrt{a_k^2+b_k^2}\sin\left(\frac{2\pi nj}{N}+\phi_j\right)\sin\left(\frac{2\pi nk}{N}+\phi_k\right).
\end{aligned}$$

Now by using the identities

$$\sin^2\theta = (1-\cos 2\theta)/2 \quad \text{and} \quad 2\sin\theta\sin\psi = \cos(\theta-\psi) - \cos(\theta+\psi)$$

and interchanging the sums, the right-hand side quantity equals

$$\sum_{j=1}^{d/2}(a_j^2+b_j^2)\sum_{n=0}^{N-1}\left(\frac{1}{2}-\frac{1}{2}\cos\left(\frac{2\pi 2nj}{N}+2\phi_j\right)\right)$$
$$+\frac{1}{2}\sum_{j\neq k}\sqrt{a_j^2+b_j^2}\sqrt{a_k^2+b_k^2}\sum_{n=0}^{N-1}\cos\left(\frac{2\pi 2n}{N}(j-k)+\phi_j-\phi_k\right)$$
$$-\frac{1}{2}\sum_{j\neq k}\sqrt{a_j^2+b_j^2}\sqrt{a_k^2+b_k^2}\sum_{n=0}^{N-1}\cos\left(\frac{2\pi 2n}{N}(j+k)+\phi_j+\phi_k\right).$$

By using the identity

$$\sum_{n=0}^{N-1}\cos\left(\frac{2\pi nj}{N}+\alpha\right)=0,$$

which holds for each integer j that is not divisible by N and for each $\alpha\in\mathbb{R}$, the sums above simplify to

$$\frac{N}{2}\sum_{j=1}^{d/2}(a_j^2+b_j^2)=\frac{N}{2}\|x\|^2.$$

Since this is equal to $\frac{d}{2}\sum_{n=0}^{N-1}|\langle x,e_n\rangle|^2$, we obtain (2.4.5).

Definition 2.4.3. Let k,d, and N be integers such that $1\leq k<N$ and $2\leq d<N$. Let $F_N=\{e_n\}_{n=1}^N$ be a frame for $\mathbb{R}^d$. Let p_N be a permutation of $\{1,2,\ldots,N\}$. We define the variation of order k of the frame F_N with respect to p_N as

$$\sigma_k(F_N,p_N):=\sum_{n=1}^{N-k}\|\Delta^k e_{p_N(n)}\|,$$

where Δ^k denotes the kth order difference defined recursively by $\Delta e_{p_N(n)}=e_{p_N(n)}-e_{p_N(n+1)}$ and $\Delta^k e_{p_N(n)}=\Delta(\Delta^{k-1}e_{p_N(n)})$ for all $k\geq 2$.

Frame variation is the quantity that reflects the "interdependencies" among frame elements. More precisely, if a frame F has low variation with respect to a permutation p, then the frame elements will not oscillate too much in that ordering

[15]. We shall see in Chapter 3 that the notion of frame variation plays an important role in refining quantization error. Families of frames that have bounded frame variation will result in a better quantization error. Harmonic frames are an example of such a family of frames. In fact, one can compute the frame variation of harmonic frames explicitly.

Lemma 2.4.4. *Let a sequence* $\{e_n\}_{n=1}^{\infty}$ *of vectors in* $\mathbb{R}^2$ *be defined by* $e_n = (\cos n\theta, \sin n\theta)$ *for some* $\theta \in [0, 2\pi]$. *We have, for each integer* $k \geq 1$,

$$\|\Delta^k e_n\| = \left(2 \sin \frac{\theta}{2}\right)^k.$$

Proof. By induction one can show that the kth order difference is equivalent to the following:

$$\forall k = 1, 2, \ldots, \quad \Delta^k e_n = \sum_{j=0}^{k} (-1)^j \binom{k}{j} e_{n+j}. \tag{2.4.6}$$

With another application of induction, one can show that

$$\|\Delta^k e_n\|^2 = 2 \sum_{j=0}^{k} (-1)^j \binom{2k}{k+j} \cos j\theta - \binom{2k}{k}. \tag{2.4.7}$$

Now we shall use induction to show that the last step is equal to $2^k(1 - \cos\theta)^k$.

It is easy to verify the formula for $k = 1$. Assume the formula holds for some $k > 1$, i.e.,

$$2 \sum_{j=0}^{k} (-1)^j \binom{2k}{k+j} \cos j\theta - \binom{2k}{k} = 2^k (1 - \cos\theta)^k. \tag{2.4.8}$$

We want to show that the formula holds for $k + 1$, i.e.,

$$2 \sum_{j=0}^{k+1} (-1)^j \binom{2k+2}{k+1+j} \cos j\theta - \binom{2k+2}{k+1} = 2^{k+1} (1 - \cos\theta)^{k+1}. \tag{2.4.9}$$

The left side of (2.4.9) is

$$
\begin{aligned}
&2\sum_{j=0}^{k+1}(-1)^j\left[\binom{2k+1}{k+j}+\binom{2k+1}{k+j+1}\right]\cos j\theta-\binom{2k+1}{k+1}-\binom{2k+1}{k}\\
&\quad=2\sum_{j=0}^{k+1}(-1)^j\left[\binom{2k}{k+j}+\binom{2k}{k+j-1}+\binom{2k}{k+j+1}+\binom{2k}{k+j}\right]\cos j\theta\\
&\qquad-\binom{2k}{k}-\binom{2k}{k+1}-\binom{2k}{k}-\binom{2k}{k-1}\\
&\quad=2^2\sum_{j=0}^{k+1}(-1)^j\binom{2k}{k+j}\cos j\theta-2\binom{2k}{k}+2\sum_{j=0}^{k+1}(-1)^j\binom{2k}{k+j-1}\cos j\theta\\
&\qquad+2\sum_{j=0}^{k+1}(-1)^j\binom{2k}{k+j+1}\cos j\theta-\binom{2k}{k+1}-\binom{2k}{k-1}\\
&\quad=2^{k+1}(1-\cos\theta)^k-2\sum_{j=-1}^{k}(-1)^j\binom{2k}{k+j}\cos(j+1)\theta\\
&\qquad-2\sum_{j=1}^{k+2}(-1)^j\binom{2k}{k+j}\cos(j-1)\theta-2\binom{2k}{k+1}\\
&\quad=2^{k+1}(1-\cos\theta)^k-2\cos\theta\sum_{j=0}^{k}(-1)^j\binom{2k}{k+j}\cos j\theta\\
&\qquad+2\sin\theta\sum_{j=0}^{k}(-1)^j\binom{2k}{k+j}\sin j\theta-2\cos\theta\sum_{j=1}^{k+2}(-1)^j\binom{2k}{k+j}\cos j\theta\\
&\qquad\quad-2\sin\theta\sum_{j=1}^{k+2}(-1)^j\binom{2k}{k+j}\sin j\theta\\
&\quad=2^{k+1}(1-\cos\theta)^k-S_k\cos\theta-S_k\cos\theta+2\cos\theta\binom{2k}{k}\\
&\qquad\left(\text{where }\ S_k=2\sum_{j=0}^{k}(-1)^j\binom{2k}{k+j}\cos j\theta\right)\\
&\quad=2^{k+1}(1-\cos\theta)^k-2\cos\theta\left(S_k-\binom{2k}{k}\right)\\
&\quad=2^{k+1}(1-\cos\theta)^k-2\cos\theta\cdot 2^k(1-\cos\theta)^k\\
&\quad=2^{k+1}(1-\cos\theta)^{k+1}
\end{aligned}
$$

and this is the right side of (2.4.9). Now using the identity that $1-\cos\theta=$

$2\sin^2(\theta/2)$, we have the result. □

Theorem 2.4.5. *For each integer $k \geq 1$, and each integer $N > 2$, if p is the identity permutation of $\{1, \ldots, N\}$ then*

$$\begin{aligned}
\sigma_k(H_N^2, p) &= (N-k)2^k \sin^k\Big(\frac{\pi}{N}\Big), \\
\sigma_k(H_N^d, p) &= \begin{cases} (N-k)\Big(\frac{2}{d}\Big)^{1/2}\Big(A(k,N,d) - \frac{d+1}{2}\binom{2k}{k}\Big)^{1/2} & \text{if } d \text{ is even,} \\ (N-k)\Big(\frac{2}{d}\Big)^{1/2}\Big(\frac{1}{2} + A(k,N,d-1) - \frac{d}{2}\binom{2k}{k}\Big)^{1/2} & \text{if } d \text{ is odd,} \end{cases}
\end{aligned}$$

where

$$A(k,N,d) = \sum_{j=0}^{k}(-1)^j\binom{2k}{k+j}\sin\Big(\frac{(d+1)\pi j}{N}\Big)\csc\Big(\frac{\pi j}{N}\Big).$$

Proof. From Lemma 2.4.4 we have

$$\sigma_k(H_N^2, p) = \sum_{n=0}^{N-k-1}\|\Delta^k e_n\| = \sum_{n=0}^{N-k-1}\left(2\sin\Big(\frac{1}{2}\cdot\frac{2\pi}{N}\Big)\right)^k = (N-k)2^k\sin^k\Big(\frac{\pi}{N}\Big).$$

To prove the second formula for d even, we let e_n be defined as in (2.4.3). Then from formulae (2.4.6)–(2.4.8), we obtain

$$\begin{aligned}
\frac{d}{2}\left\|\Delta^k e_n\right\|^2 &= \sum_{\ell=1}^{d/2}\left[\left(\sum_{j=0}^{k}(-1)^j\binom{k}{j}\cos(n+j)\frac{2\pi\ell}{N}\right)^2 + \left(\sum_{j=0}^{k}(-1)^j\binom{k}{j}\sin(n+j)\frac{2\pi\ell}{N}\right)^2\right] \\
&= \sum_{\ell=1}^{d/2}2^{2k}\sin^{2k}\Big(\frac{\pi\ell}{N}\Big) = 2^{2k}\sum_{\ell=1}^{d/2}\sin^{2k}\Big(\frac{\pi\ell}{N}\Big) \\
&= \sum_{\ell=1}^{d/2}\left[2\sum_{j=0}^{k}(-1)^j\binom{2k}{k+j}\cos\frac{2\pi j\ell}{N} - \binom{2k}{k}\right] \\
&= 2\sum_{j=0}^{k}(-1)^j\binom{2k}{k+j}\sum_{\ell=1}^{d/2}\cos\frac{2\pi j\ell}{N} - \frac{d}{2}\binom{2k}{k} \\
&= A(k,N,d) - \sum_{j=0}^{k}(-1)^j\binom{2k}{k+j} - \frac{d}{2}\binom{2k}{k} \\
&= A(k,N,d) - \frac{d+1}{2}\binom{2k}{k}. \qquad (2.4.10)
\end{aligned}$$

The last two equalities follow from two well-known summing identities. Combining equation (2.4.10) with the definition of $\sigma_k(H_N^d, p)$, we have the result. For the case when d is odd we use the definition of e_n defined in (2.4.4) and proceed similarly as in the above argument. □

Also, by using the same method as in the proof of Theorem 2.4.5, we have the following theorem.

Theorem 2.4.6. *Let θ be a real number and let N, d be positive integers. Then we have the following:*

(i)

$$\sum_{n=1}^{N}(-1)^{n+1}\binom{2N}{N+n}\frac{\sin(d+\frac{1}{2})n\theta}{\sin\frac{n\theta}{2}} \leq \Big(d+\frac{1}{2}\Big)\binom{2N}{N},$$

(ii)

$$\begin{aligned}\sum_{n=1}^{N}(-1)^{n+1}\binom{2N}{N+n}\sin(nd\theta)\cot\left(\frac{n\theta}{2}\right) &\leq d\binom{2N}{N} + 2^{2N-1}\sin^{2N}\left(\frac{d\theta}{2}\right)\\ &\leq d\binom{2N}{N} + \frac{(d\theta)^{2N}}{2},\end{aligned}$$

(iii)

$$(2d+1)\sum_{n=0}^{\lfloor N/2\rfloor}\binom{2N}{N+2n}+(-1)^{d+1}\sum_{n=0}^{\lfloor N/2\rfloor}\binom{2N}{N+2n+1} = 4^N\left\lceil\frac{d}{2}\right\rceil+\Big(d+\frac{1}{2}\Big)\binom{2N}{N},$$

(iv)

$$\sum_{n=0}^{\lfloor N/2\rfloor}\binom{2N}{N+2n} = 4^{N-1}+\frac{1}{2}\binom{2N}{N},$$

(v)

$$\sum_{n=0}^{\lfloor N/2\rfloor}\binom{2N}{N+2n+1} = 4^{N-1},$$

(vi)

$$\sum_{n=0}^{N}\binom{2N}{N+n}=\frac{1}{2}\left[4^N+\binom{2N}{N}\right].$$

Proof. By retracing the steps of proof of Theorem 2.4.5 beginning from the third equality of (2.4.10), we have

$$\begin{aligned}
2^{2N}\sum_{\ell=1}^{d}\sin^{2N}\left(\frac{\ell\theta}{2}\right) &= \sum_{\ell=1}^{d}\left[2\sum_{n=0}^{N}(-1)^n\binom{2N}{N+n}\cos(n\ell\theta)-\binom{2N}{N}\right]\\
&= 2\sum_{n=0}^{N}(-1)^n\binom{2N}{N+n}\sum_{\ell=1}^{d}\cos(n\ell\theta)-d\binom{2N}{N}\\
&= \sum_{n=0}^{N}(-1)^n\binom{2N}{N+n}\left[\frac{\sin(d+\frac{1}{2})n\theta}{\sin\frac{n\theta}{2}}-1\right]-d\binom{2N}{N}\\
&= \sum_{n=1}^{N}(-1)^n\binom{2N}{N+n}\frac{\sin(d+\frac{1}{2})n\theta}{\sin\frac{n\theta}{2}}+(2d+1)\binom{2N}{N}\\
&\quad -\left(d+\frac{1}{2}\right)\binom{2N}{N}\\
&= \sum_{n=1}^{N}(-1)^n\binom{2N}{N+n}\frac{\sin(d+\frac{1}{2})n\theta}{\sin\frac{n\theta}{2}}+\left(d+\frac{1}{2}\right)\binom{2N}{N}. \quad (2.4.11)
\end{aligned}$$

By letting $\theta=\pi$ and noting that

$$\sum_{\ell=1}^{d}\sin^{2N}\frac{\pi\ell}{2}=\left\lceil\frac{d}{2}\right\rceil,$$

we obtain (iii). We observe that

$$\frac{d}{4d+2}\le\frac{\lceil d/2\rceil}{2d+1}\le\frac{d+2}{4d+2}.$$

Thus,

$$\lim_{d\to\infty}\frac{\lceil d/2\rceil}{2d+1}=\frac{1}{4}.$$

Dividing both sides of (iii) by $2d+1$ and letting $d\to\infty$ we obtain (iv). To obtain (v), we substitute (iv) back in (iii) and solve for (v). The equality (vi) is obtained

by adding (iv) and (v). Inequality (i) follows by noting that the left-hand side of (2.4.11) is nonnegative. We obtain inequality (ii) by expanding

$$\sin\left(d+\frac{1}{2}\right)n\theta = \sin(dn\theta)\cos\frac{n\theta}{2} + \cos(nd\theta)\sin\frac{n\theta}{2},$$

and using 2.4.8. □

Remark 2.4.7. (a) We remark that the inequalities (i) and (ii) in Theorem 2.4.6 are quite sharp for certain choices of θ, d, and N. For example, for $\theta = 2\pi/7$, $d = 1$, and $N = 6$, the left side of inequality (i) is about 1385.817677, while the right side is 1386. With the same values of θ, d, and N the left side of inequality (ii) is about 923.9088384, while the right side is about 924.0911613.

(b) The combinatoric sum identities (iii)-(vi) also have a direct proof using some properties of binomial coefficients.

Chapter 3

Sigma-Delta ($\Sigma\Delta$) Quantization

3.1 Overview

In this chapter, we shall discuss two schemes of quantization. The first one, called Pulse Code Modulation or PCM, is considered perhaps the most basic scheme of quantization. This scheme quantizes a signal of interest by replacing each coefficient of the signal expansion with the element of a given discrete set (alphabet) that is closest in distance to the coefficient. We shall discuss the PCM of finite frame expansions of signals in $\mathbb{R}^d$ and shall derive a quantization error associated to this technique [16]. The second scheme of quantization, called Sigma-Delta ($\Sigma\Delta$) quantization, was introduced by Inose, Yasuda and Murakami in 1962 [23]. This scheme is widely used in quantizing signals because of its robustness against circuit imperfections, and it can provide high accuracy A/D conversion [3, 9, 25, 26]. We shall see that this scheme uses feedback loops in the sense that the elements of a quantized sequence keep being fed back into the scheme to produce new quantized coefficients. This exploitation of feedback loops "generates a quantized signal that oscillates between levels, keeping its average equal to the average input" [24]. We shall use basic analysis to derive the quantization error associated to the $\Sigma\Delta$ quantization. As such, we shall see that, in the setting of redundant signal expansions, this quantization scheme outperforms PCM with respect to faster decay in quan-

tization error. However, if the signal is expanded over an orthonormal basis, then PCM turns out to be the optimal quantizer since it minimizes the Euclidean norm of quantization error. More precisely, let x be a signal of interest in $\mathbb{R}^d$, and let $\{e_n\}_{n=1}^N$ be an orthonormal basis for $\mathbb{R}^d$. (It follows necessarily that $d = N$.) Then there exist unique $c_1, \ldots, c_N \in \mathbb{R}$ such that

$$x = \sum_{n=1}^{N} c_n e_n.$$

Let $q_1, \ldots, q_N$ be the quantized coefficients obtained from PCM algorithms and let $\widetilde{x} = \sum_{n=1}^N q_n e_n$. Then by a property of orthonomal bases, we have

$$\|x - \widetilde{x}\|^2 = \sum_{n=1}^{N} (c_n - q_n)^2.$$

Since PCM determines q_n, an element of the given alphabet, in such a way that $|c_n - q_n|$ is the minimum for each $1 \leq n \leq N$, we see that $\|x - \widetilde{x}\|$ is the minimum as well.

3.2 Pulse Code Modulation (PCM)

Let $\{e_n\}_{n=1}^N$ be a unit norm tight frame for $\mathbb{R}^d$. Then from Chapter 2, we have the expansion for each $x \in \mathbb{R}^d$ by

$$x = \frac{d}{N} \sum_{n=1}^{N} x_n e_n, \quad x_n = \langle x, e_n \rangle. \tag{3.2.1}$$

Definition 3.2.1. Let $\delta > 0$. The $2\lceil 1/\delta \rceil$-level PCM quantizer with step size δ

replaces each $x_n \in \mathbb{R}$ in the frame expansion (3.2.1) with

$$q_n = q_n(x) = \begin{cases} \delta\big(\lceil x_n/\delta \rceil - 1/2\big) & \text{if } |x_n| < 1, \\ \delta\big(\lceil 1/\delta \rceil - 1/2\big) & \text{if } x_n \geq 1, \\ -\delta\big(\lceil 1/\delta \rceil - 1/2\big) & \text{if } x_n \leq -1. \end{cases} \tag{3.2.2}$$

Proposition 3.2.2. *Let $\delta > 0$, and let $\|\cdot\|$ be the d-dimensional Euclidean 2-norm. Let $x \in \mathbb{R}^d$ and let $\widetilde{x}$ be the quantized expansion given by $2\lceil 1/\delta \rceil$-level PCM. If $\|x\| \leq 1$ then the quantization error $\|x - \widetilde{x}\|$ satisfies*

$$\|x - \widetilde{x}\| \leq \Big(\frac{d}{2}\Big)\delta.$$

Proof. First we write

$$\widetilde{x} = \frac{d}{N} \sum_{n=1}^{N} q_n e_n, \tag{3.2.3}$$

where q_n is obtained from PCM for each $1 \leq n \leq N$. Then, from the Cauchy-Schwarz Inequality, we have for each $1 \leq n \leq N$, that

$$|x_n| = |\langle x, e_n \rangle| \leq \|x\| \, \|e_n\| \leq 1.$$

For each $1 \leq n \leq N$ we have $x_n/\delta \leq \lceil x_n/\delta \rceil < x_n/\delta + 1$, so that

$$-\frac{\delta}{2} = x_n - \delta\Big(\frac{x_n}{\delta} + 1\Big) + \frac{\delta}{2} < x_n - q_n = x_n - \delta\Big\lceil \frac{x_n}{\delta} \Big\rceil + \frac{\delta}{2} \leq x_n - \delta \cdot \frac{x_n}{\delta} + \frac{\delta}{2} = \frac{\delta}{2}.$$

Hence, for each $1 \leq n \leq N$,

$$|x_n - q_n| \leq \frac{\delta}{2}. \tag{3.2.4}$$

From (3.2.3) and (3.2.4) we have

$$\|x - \widetilde{x}\| = \frac{d}{N} \left\| \sum_{n=1}^{N} (x_n - q_n) e_n \right\| \leq \Big(\frac{\delta}{2}\Big)\Big(\frac{d}{N}\Big) \sum_{n=1}^{N} \|e_n\| = \Big(\frac{d}{2}\Big)\delta.$$

Thus, the proof is complete. □

3.3 Sigma-Delta ($\Sigma\Delta$) quantization

When first introduced, Sigma-Delta ($\Sigma\Delta$) Quantization was used to quantize oversampled bandlimited functions; so, before we define the definition of this quantization scheme in the setting of finite frame expansion, we should understand the definition in its original setting [9].

Let f be a bandlimited function on $\mathbb{R}$ with bandwidth $\Omega > 0$ and assume that f takes value in the interval $[1, 2]$. We recall from Chapter 1 that this means that f is an L^∞ function on $\mathbb{R}$ whose Fourier transform $\widehat{f}$ (as a distribution) vanishes outside $[-\Omega, \Omega]$. Then from the classical sampling theorem [3, 27], for each $0 < T < 1$, the function f can be reconstructed from the sampling sequence $\{f(nT)\}_{n\in\mathbb{Z}}$ as follows:

$$f(t) = T\sum_{n\in\mathbb{Z}} f(nT)g(t - nT), \tag{3.3.1}$$

where g is an appropriate smoothing kernel or sampling function, that is, $\widehat{g} \in C^\infty$ and

$$\widehat{g}(\xi) = \begin{cases} 1 & \text{for } |\xi| \leq \Omega, \\ \\ 0 & \text{for } |\xi| > \Omega/T. \end{cases}$$

Then the first order $\Sigma\Delta$ modulator uses $\{f(nT)\}_{n\in\mathbb{Z}}$ as inputs to generate the se-

quence $\{q_T(n)\}_{n\in\mathbb{Z}}$ as follows:

$$F_T(n) = \begin{cases} \sum_{k=1}^{n} f(kT) & \text{for } n \geq 1, \\ 0 & \text{for } n = 0, \\ -\sum_{k=n+1}^{0} f(kT) & \text{for } n < 0, \end{cases} \tag{3.3.2}$$

$$Q_T(n) = \lfloor F_T(n) \rfloor, \tag{3.3.3}$$

$$q_T(n) = Q_T(n) - Q_T(n-1), \tag{3.3.4}$$

From (3.3.2) we see that $F_T(n) - F_T(n-1) = f(nT)$ for all $n \in \mathbb{Z}$. Since f takes value in the interval $[1, 2]$ we can check that $q_T(n)$ takes either the value 1 or 2. In fact, for each $n \in \mathbb{Z}$, we have $F_T(n) - 1 < \lfloor F_T(n) \rfloor \leq F_T(n)$ and similarly $-F_T(n-1) \leq -\lfloor F_T(n-1) \rfloor < -F_T(n-1) + 1$. Adding these inequalities yields

$$\begin{aligned} 0 \leq f(nT) - 1 = F_T(n) - F_T(n-1) - 1 \\ < q_T(n) < F_T(n) - F_T(n-1) + 1 = f(nT) + 1 \leq 3. \end{aligned}$$

Since $q_T(n)$ is an integer, from the above chain of inequalities we have either $q_T(n) = 1$ or $q_T(n) = 2$. We note that the equations (3.3.2) and (3.3.4) correspond to "Σ" and "Δ", respectively; hence the name of the modulator. Since F_T and Q_T will accumulate into large numbers as time elapses, neither can be calculated in a circuit. Thus one introduces the auxiliary variable $u_T = F_T - Q_T = \{F_T\} \in [0, 1)$. Then u_T satisfies the recursive relation:

$$u_T(n) - u_T(n-1) = F(nT) - q_T(n). \tag{3.3.5}$$

Since $u_T(n) \in [0, 1)$, from (3.3.5) we have the relation

$$q_T(n) = \lfloor f(nT) + u_T(n-1) \rfloor. \tag{3.3.6}$$

To see this, we observe from (3.3.5) that $q_T(n)+u_T(n) = f(nT)+u_T(n-1)$, so that $q_T(n) + \lfloor u_T(n) \rfloor = \lfloor q_T(n) + u_T(n) \rfloor = \lfloor f(nT) + u_T(n-1) \rfloor$. Since $u_T(n) \in [0, 1)$, it follows that $\lfloor u_T(n) \rfloor = 0$ and hence (3.3.6) follows. Using the auxiliary variable u_T, now we can translate the procedure in (3.3.2)–(3.3.4) into the following equivalent procedure:

$$u_T(n) = u_T(n-1) + F(nT) - q_T(n) \quad \text{with } u_T(0) = 0, \tag{3.3.7}$$

$$q_T(n) = \begin{cases} 1 & \text{if } f(nT) + u_T(n-1) < 2, \\ 2 & \text{if } f(nT) + u_T(n-1) \geq 2. \end{cases} \tag{3.3.8}$$

Formulae (3.3.7) and (3.3.8) motivate the definition of the $\Sigma\Delta$ quantization for finite frame expansions in $\mathbb{R}^d$. Let $K \in \mathbb{N}$ and $\delta > 0$. We define the *midrise quantization alphabet* $\mathcal{A}_K^\delta$ to be the set of $2K$ numbers in arithmetic progression with common difference δ, and the first number $(-K + 1/2)\delta$. Thus,

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

We define the $2K$-*level midrise uniform scalar quantizer* with stepsize δ by

$$Q(u) = \arg\min_{q \in \mathcal{A}_\mathcal{K}^\delta} |u - q|.$$

In words, $Q(u)$ denotes the element q in $\mathcal{A}_\mathcal{K}^\delta$ which is closest in distance to the element u. By convention, if there are two elements in $\mathcal{A}_\mathcal{K}^\delta$ which are equally closest to u, then $Q(u)$ will be chosen to be the larger of these two elements. In order to be able to do numerical simulation with $\Sigma\Delta$ quantization, we need an explicit formula for the quantizer Q. The following theorem will let us do just that.

Theorem 3.3.1. *The quantizer Q as defined above has the following formula:*

$$Q(u) = \begin{cases} \operatorname{sgn}(u)\big(K - \frac{1}{2}\big)\delta & \text{if } |u| \geq K\delta, \\ \big(\frac{1}{2} + \big\lfloor \frac{u}{\delta} \big\rfloor\big)\delta & \text{if } |u| < K\delta. \end{cases} \tag{3.3.9}$$

Proof. Let a real number u be given. If $|u| \geq K\delta$ then the formula is easily verified.

If

$$(K - \frac{1}{2})\delta \leq u < K\delta$$

then by definition $Q(u) = (K - 1/2)\delta$. Since

$$K - \frac{1}{2} \leq \frac{u}{\delta} < K,$$

we have $\lfloor u/\delta \rfloor = K - 1$. So

$$\Big(\frac{1}{2} + \Big\lfloor \frac{u}{\delta} \Big\rfloor\Big)\delta = \Big(\frac{1}{2} + K - 1\Big)\delta = \Big(K - \frac{1}{2}\Big)\delta = Q(u).$$

Similarly for the case

$$-K\delta < u \leq \Big(- K + \frac{1}{2}\Big)\delta.$$

Now assume that

$$\Big(- K + \frac{1}{2}\Big)\delta < u < \Big(K - \frac{1}{2}\Big)\delta$$

and that for some $m = 0, \ldots, 2K - 1$

$$Q(u) = \Big(- K + \frac{1}{2}\Big)\delta + m\delta.$$

Then by definition of $Q(u)$ we see that

$$-\frac{\delta}{2} \leq u - Q(u) < \frac{\delta}{2}.$$

Replacing $Q(u)$ with $(-K+1/2)\delta + m\delta$, we have

$$0 \leq \frac{u}{\delta} + K - m < 1.$$

Hence

$$\left\lfloor \frac{u}{\delta} + K - m \right\rfloor = 0.$$

Now using the property that $\lfloor x+n \rfloor = \lfloor x \rfloor + n$ for all integers n and real numbers x, we obtain that $m = \lfloor u/\delta \rfloor + K$. Replacing this value of m in $Q(u)$, we get

$$Q(u) = \Big(-K + \frac{1}{2} \Big)\delta + \Big(\Big\lfloor \frac{u}{\delta} \Big\rfloor + K \Big)\delta = \Big(\frac{1}{2} + \Big\lfloor \frac{u}{\delta} \Big\rfloor \Big)\delta. \qquad \square$$

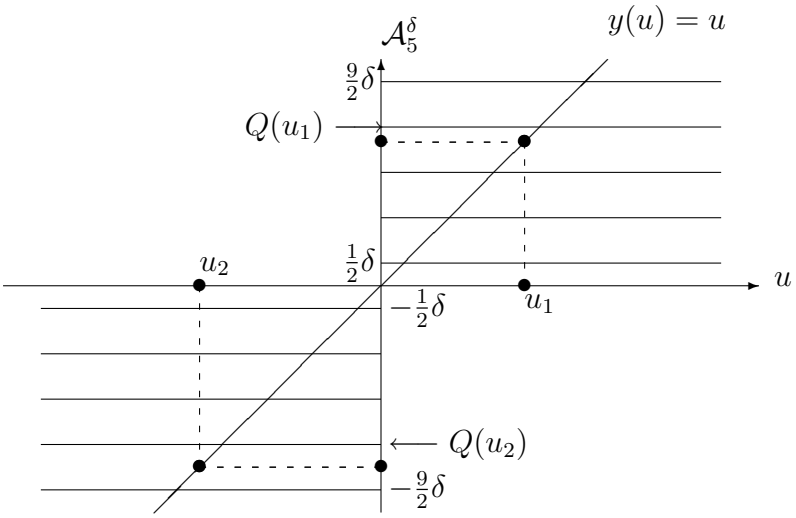

Figure 3.1: Picture diagram of the quantizer Q.

Definition 3.3.2. Let $\{x_n\}_{n=1}^N \subseteq \mathbb{R}$, and let p be a permutation of the set $\{1, 2, \ldots, N\}$. Then the $2K-$*level first-order* $\Sigma\Delta$ *quantizer with step size* δ is defined recursively by

$$u_n = u_{n-1} + x_{p(n)} - q_n, \tag{3.3.10}$$

$$q_n = Q(u_{n-1} + x_{p(n)}), \tag{3.3.11}$$

where u_0 is a specified constant.

We usually refer to this definition as the first-order $\Sigma\Delta$ quantizer for short. We see that $\Sigma\Delta$ quantizer produces two sequences: $\{u_n\}_{n=0}^N$ and $\{q_n\}_{n=1}^N$. We shall refer to $\{q_n\}_{n=1}^N$ as the *quantized sequence* and refer to $\{u_n\}_{n=0}^N$ as the *auxiliary sequence of state variables*. We shall refer to the permutation p as the *quantization order*. The following proposition shows that the $\Sigma\Delta$ quantizer defined above is stable, that is, the auxiliary sequence $\{u_n\}_{n=0}^N$ is uniformly bounded provided that the input sequence $\{x_n\}_{n=1}^N$ is appropriately uniformly bounded.

Proposition 3.3.3. *Let K be a positive integer, let $\delta > 0$, and consider the $\Sigma\Delta$ quantizer defined by (3.3.10)-(3.3.11). If $|u_0| \leq \delta/2$ and for all integers $1 \leq n \leq N$,*

$$|x_n| \leq \Big(K - \frac{1}{2}\Big)\delta,$$

then for all integers $0 \leq n \leq N$,

$$|u_n| \leq \frac{\delta}{2}.$$

Proof. We may assume without loss of generality that p is the identity permutation. We shall proceed by induction. The base step, $|u_0| \leq \delta/2$, holds by assumption. Next, we suppose that $|u_{j-1}| \leq \delta/2$ for some $2 \leq j \leq N$. We want to show that $|u_j| \leq \delta/2$. We have $|u_{j-1} + x_j| \leq |u_{j-1}| + |x_j| \leq K\delta$. We also note from the definition of Q that if $|u| \leq K\delta$, then $0 \leq Q(u) - u \leq \delta/2$. Combining this with (3.3.10)-(3.3.11), we have

$$|u_j| = |(u_{j-1} + x_j) - Q(u_{j-1} + x_j)| \leq \frac{\delta}{2}. \qquad \square$$

The following Theorem is one of the main results found in [15]. It states the basic quantization error estimate associated to the first-order $\Sigma\Delta$ quantization. We begin with the setup. Let $K \in \mathbb{N}$ and $\delta > 0$. Let $F = \{e_n\}_{n=1}^N$ be a frame for $\mathbb{R}^d$, let p be a permutation of $\{1, \ldots, N\}$, and let $x \in \mathbb{R}^d$ be the input. We form the quantized expansion

$$\widetilde{x} = \sum_{n=1}^{N} q_n S^{-1} e_{p(n)} \tag{3.3.12}$$

from the frame expansion

$$x = \sum_{n=1}^{N} x_{p(n)} S^{-1} e_{p(n)}, \quad x_{p(n)} = \langle x, e_{p(n)} \rangle. \tag{3.3.13}$$

Here, $\{q_n\}_{n=1}^N$ is the quantized sequence which is calculated using the recurrence relations (3.3.10)-(3.3.11). We want to calculate how well (3.3.12) approximates (3.3.13).

Theorem 3.3.4. *Given the $\Sigma\Delta$ quantization as above. Let $F = \{e_n\}_{n=1}^N$ be a finite unit norm frame for $\mathbb{R}^d$, p a permutation of $\{1, 2, \ldots, N\}$, $|u_0| \leq \delta/2$. If $x \in \mathbb{R}^d$ satisfies $\|x\| \leq (K - 1/2)\delta$, then we have the following quantization error:*

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

where S^{-1} is the inverse frame operator for F and $\sigma(F, p)$, the frame variation with respect to p, is defined by (see Chapter 2, Section 2.4)

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

Proof.

$$\begin{aligned} x - \widetilde{x} &= \sum_{n=1}^{N} (x_{p(n)} - q_n) S^{-1} e_{p(n)} \\ &= \sum_{n=1}^{N} (u_n - u_{n-1}) S^{-1} e_{p(n)} \\ &= \sum_{n=1}^{N-1} u_n S^{-1}(e_{p(n)} - e_{p(n+1)}) + u_N S^{-1} e_{p(N)} - u_0 S^{-1} e_{p(1)}. \end{aligned}$$

Since $\|x\| \leq (K - 1/2)\delta$ it follows from Cauchy-Schwarz Inequality that

$$\forall\, 1 \leq n \leq N, \quad |x_n| = |\langle x, e_n \rangle| \leq \|x\| \, \|e_n\| \leq (K - 1/2)\delta.$$

Thus, combining with the stability result of the sequence $\{u_n\}_{n=0}^{N}$ from Proposition 3.3.3,

$$\begin{aligned} \|x - \widetilde{x}\| &\leq \sum_{n=1}^{N-1} \frac{\delta}{2} \left\|S^{-1}\right\| \left\|e_{p(n)} - e_{p(n+1)}\right\| + |u_N| \left\|S^{-1}\right\| + |u_0| \left\|S^{-1}\right\| \\ &= \left\|S^{-1}\right\| \left(\sigma(F, p)\frac{\delta}{2} + |u_0| + |u_N|\right). \end{aligned}$$

□

For the purpose of applicability we usually use unit norm tight frame for $\mathbb{R}^d$ to expand a signal of interest. In this case the bound of the quantization error derived in Theorem 3.3.4 can be adjusted according to the properties of the frame operator S. More precisely, from Lemma 2.4.2 and the condition (2.3.2), we have

$$S = \frac{N}{d} I, \tag{3.3.14}$$

so that $\|S^{-1}\| = d/N$ and we have the following corollary.

Corollary 3.3.5. *Given the $\Sigma\Delta$ scheme of Definition 3.3.2. Let $F = \{e_n\}_{n=1}^{N}$ be a unit norm tight frame for $\mathbb{R}^d$ with frame bound $A = N/d$, let p be a permutation*

of $\{1, 2, \ldots, N\}$*, let* $|u_0| \leq \delta/2$*, and let* $x \in \mathbb{R}^d$ *satisfy* $\|x\| \leq (K - 1/2)\delta$*. Then the quantization error* $\|x - \widetilde{x}\|$ *satisfies*

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

If we apply the stability result of the auxiliary sequence $\{u_n\}_{n=1}^N$ *from Proposition 3.3.3 then we obtain*

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

Since from Definition 3.3.2 the only restriction of the initial variable u_0 is that $|u_0| \leq \delta/2$, to lower the bound of quantization error, we set the initial variable u_0 to be 0. Moreover the bound of quantization error can be improved if one knows more information about the variable u_N. An example of frame that allows us to characterize the variable u_N based on the parity of the cardinality of frame is the *zero sum* frame which is a type of frame for which the sum of all frame elements is equal to 0.

Theorem 3.3.6. *Given the* $\Sigma\Delta$ *scheme of Definition 3.3.2. Let* $F = \{e_n\}_{n=1}^N$ *be a unit norm tight frame for* $\mathbb{R}^d$ *with frame bound* $A = N/d$*, and assume that* F *satisfies the zero sum condition*

$$\sum_{n=1}^{N} e_n = 0. \tag{3.3.15}$$

Additionally, set $u_0 = 0$*. Then*

$$|u_N| = \begin{cases} 0 & \text{if } N \text{ is even,} \\ \delta/2 & \text{if } N \text{ is odd.} \end{cases} \tag{3.3.16}$$

Proof. From (3.3.10) we have $u_n - u_{n-1} = x_{p(n)} - q_n$, so that

$$u_N = u_N - u_0 = \sum_{n=1}^{N} u_n - u_{n-1} = \sum_{n=1}^{N} x_{p(n)} - \sum_{n=1}^{N} q_n = \sum_{n=1}^{N} x_n - \sum_{n=1}^{N} q_n.$$

Since $\sum_{n=1}^{N} e_n = 0$, we have

$$\sum_{n=1}^{N} x_n = \sum_{n=1}^{N} \langle x, e_n \rangle = \langle x, \sum_{n=1}^{N} e_n \rangle = \langle x, 0 \rangle = 0.$$

Therefore

$$\sum_{n=1}^{N} q_n = -u_N.$$

Since for each $1 \leq n \leq N$, q_n is an odd integer multiple of $\delta/2$, we consider two cases. The first case is that N is an odd integer. Then $\sum_{n=1}^{N} q_n$ is an odd integer multiple of $\delta/2$ and since from stability result $|u_N| \leq \delta/2$, it follows that

$$|u_N| = \frac{\delta}{2}.$$

The second case is that N is an even integer. Then $\sum_{n=1}^{N} q_n$ is an integer multiple of δ and since from stability result $|u_N| \leq \delta/2$, it follows that

$$|u_N| = 0.$$ □

Combining this theorem with Corollary 3.3.5, we have the following corollary.

Corollary 3.3.7. *Given the $\Sigma\Delta$ scheme of Definition 3.3.2. Let $F = \{e_n\}_{n=1}^{N}$ be a unit norm tight frame for $\mathbb{R}^d$ with frame bound $A = N/d$, and assume that F satisfies the zero sum condition (3.3.15). Let p be a permutation of $\{1, 2, \ldots, N\}$, let $|u_0| \leq \delta/2$, and let $x \in \mathbb{R}^d$ satisfy $\|x\| \leq (K - 1/2)\delta$. Then the quantization error*

$\|x - \widetilde{x}\|$ *satisfies*

$$\|x - \widetilde{x}\| \leq \begin{cases} \frac{\delta d}{2N}\sigma(F,p) & \text{if } N \text{ is even,} \\ \\ \frac{\delta d}{2N}\big(\sigma(F,p)+1\big) & \text{if } N \text{ is odd.} \end{cases}$$

Chapter 4

Uniform Distribution and Discrepancy

In this chapter, we develop the theory of uniform distribution of sequences of real numbers. Some classic examples are discussed including the sequence $\{n\theta\}_{n=1}^{\infty}$, where θ is irrational. The second part of this chapter deals with the question of how well a given sequence is *distributed* over an interval of finite length. The notion used to measure the distribution of a sequence is called discrepancy. The larger the discrepancy the worse the sequence is distributed. One tries to approximate discrepancy rather than compute it directly. Erdös-Turán Inequality is one of the major tools that are used to approximate discrepancies. We shall see that this inequality approximate discrepancies in terms of exponential sum, that is, sum of the form $\sum_{n=a}^{b} e^{2\pi i f(n)}$, for some real valued function f. The following materials on the theory of uniform distribution and discrepancy are adapted from Chapters 1 and 2 of [17].

4.1 Uniform distribution mod 1

We begin by setting up some notations. Let $I = [0, 1)$ be the unit interval. We recall that the fractional part of each real number lies in I. Let $\omega = \{x_n\}_{n=1}^{\infty}$ be a given sequence of real numbers. For a positive integer N and a subset E of I, let the counting function $A(E; N; \omega)$ be defined as the number of terms x_n $(1 \leq n \leq N)$

for which $\{x_n\} \in E$. We shall sometimes write $A(E;N)$ instead of $A(E;N;\omega)$ if the sequence ω is understood from the context.

Definition 4.1.1. The sequence $\omega = \{x_n\}_{n=1}^{\infty}$ of real numbers is said to be *uniformly distributed modulo 1* (abbreviated u.d. mod 1) if for every pair a, b of real numbers with $0 \leq a < b \leq 1$ we have

$$\lim_{N\to\infty} \frac{A([a,b);N;\omega)}{N} = b - a. \tag{4.1.1}$$

Descriptively the definition says that a real sequence is uniformly distributed modulo 1 if the share of the fractional parts of the terms of the sequence with the terms in each half open subinterval of I is finally equal to the length of each such subinterval.

Example 4.1.2. The following sequence is u.d. mod 1:

$$\left\{\frac{0}{1}, \frac{0}{2}, \frac{1}{2}, \frac{0}{3}, \frac{1}{3}, \frac{2}{3}, \ldots, \frac{0}{n}, \frac{1}{n}, \ldots, \frac{n-1}{n}, \ldots\right\}.$$

Proof. Let $N \in \mathbb{N}$ and $0 \leq a < b \leq 1$. We try to compute $A([a,b);N)$. We observe that there exists a unique integer k_N such that

$$\frac{1}{2}k_N(k_N+1) \leq N < \frac{1}{2}(k_N+1)(k_N+2). \tag{4.1.2}$$

We write the first N terms of the sequence in question in terms of k_N as follows:

$$\frac{0}{1}, \frac{0}{2}, \frac{1}{2}, \frac{0}{3}, \frac{1}{3}, \frac{2}{3}, \ldots, \frac{0}{k_N}, \frac{1}{k_N}, \ldots, \frac{k_N-1}{k_N}, \frac{0}{k_N+1}, \ldots, \frac{N-k_N(k_N+1)/2}{k_N+1}.$$

Now we partition this finite sequence into $k_N + 1$ "blocks" by letting the jth block $(1 \leq j \leq k_N)$ consist of

$$\frac{0}{j}, \frac{1}{j}, \frac{2}{j}, \ldots, \frac{j-1}{j}$$

and letting the (k_N+1)th block consist of

$$\frac{0}{k_N+1}, \frac{1}{k_N+1}, \frac{2}{k_N+1}, \ldots, \frac{N-k_N(k_N+1)/2}{k_N+1}.$$

We now count the number of elements in the jth block ($1 \le j \le k_N$) that are in $[a,b)$, i.e., we count the integers m in $[0, j-1)$ such that $a \le m/j < b$ or $ja \le m < jb$. We see that this number equals $j(b-a)+\theta_j$, where $|\theta_j| < 1$. We also see that the number of elements in the (k_N+1)th block that are in $[a,b)$ is not greater than k_N+1 and is at least 0. Using these ingredients we can approximate $A([a,b);N)$ as follows:

$$\sum_{j=1}^{k_N} j(b-a)+\theta_j \le A([a,b);N) \le \sum_{j=1}^{k_N} j(b-a)+\theta_j + k_N + 1.$$

Since $|\theta_j| < 1$ and $\sum_{j=1}^{k_N} j = k_N(k_N+1)/2$, we can approximate further that

$$\frac{1}{2}(b-a)k_N(k_N+1) - k_N < A([a,b);N) < \frac{1}{2}k_N(k_N+1) + k_N + k_N + 1.$$

By (4.1.2) we have

$$(b-a)(N-(k_N+1)) - k_N < A([a,b);N) < (b-a)N + 2k_N + 1$$

or

$$(b-a) - \frac{b-a}{N} - \frac{k_N}{N}(1+b-a) < \frac{1}{N}A([a,b);N) < b-a+\frac{2k_N+1}{N}. \qquad (4.1.3)$$

Since by (4.1.2), $k_N(k_N+1)/2 \le N$, we have $(k_N/N)(k_N/N+1/N)/2 \le 1/N$ so that $\lim_{N\to\infty} k_N/N = 0$. Hence inequalities (4.1.3) imply that

$$\lim_{N\to\infty} \frac{A([a,b);N)}{N} = b-a.$$

Since a, b were arbitrary in I we have that the sequence in question is u.d. mod 1. $\square$

Example 4.1.3. Let $\omega_1 = \{x_n\}_{n=1}^{\infty}$ and $\omega_2 = \{y_n\}_{n=1}^{\infty}$ be u.d. mod 1. Then the sequence $\omega_3 = \{x_1, y_1, x_2, y_2, \dots, x_n, y_n, \dots\}$ is u.d. mod 1.

Proof. Let $N \in \mathbb{N}$ and $0 \leq a < b \leq 1$. We want to compute $A([a,b); N; \omega_3)$. If N is even then we list the first N terms of the sequence ω_3 as follows:

$$x_1, y_1, x_2, y_2, \dots, x_{N/2}, y_{N/2}.$$

So

$$A([a,b); N; \omega_3) = A([a,b); N/2; \omega_1) + A([a,b); N/2; \omega_2).$$

And we have

$$\begin{aligned} \lim_{N\to\infty} \frac{A([a,b); N; \omega_3)}{N} &= \frac{1}{2} \lim_{N\to\infty} \frac{A([a,b); N/2; \omega_1)}{N/2} + \frac{1}{2} \lim_{N\to\infty} \frac{A([a,b); N/2; \omega_2)}{N/2} \\ &= \frac{1}{2} \lim_{K\to\infty} \frac{A([a,b); K; \omega_1)}{K} + \frac{1}{2} \lim_{K\to\infty} \frac{A([a,b); K; \omega_2)}{K} \\ &= \frac{b-a}{2} + \frac{b-a}{2} = b - a. \end{aligned}$$

If N is odd then we list the first N terms of the sequence ω_3 as follows:

$$x_1, y_1, x_2, y_2, \dots, x_{(N-1)/2}, y_{(N-1)/2}, x_{(N+1)/2}.$$

So

$$A([a,b); N; \omega_3) = A([a,b); (N+1)/2; \omega_1) + A([a,b); (N-1)/2; \omega_2).$$

And we have

$$\begin{aligned} \lim_{N\to\infty} \frac{A([a,b); N; \omega_3)}{N} &= \lim_{N\to\infty} \frac{N+1}{2N} \cdot \frac{A([a,b); (N+1)/2; \omega_1)}{(N+1)/2} \\ &\quad + \lim_{N\to\infty} \frac{N-1}{2N} \cdot \frac{A([a,b); (N-1)/2; \omega_2)}{(N-1)/2} \\ &= \frac{b-a}{2} + \frac{b-a}{2} = b - a. \end{aligned}$$

Hence by definition the sequence ω_3 is u.d. mod 1. □

Proposition 4.1.4. *If a sequence $\{x_n\}_{n=1}^{\infty}$ is u.d. mod 1, then the sequence $\{\{x_n\}\}_{n=1}^{\infty}$ is everywhere dense in $\overline{I}$.*

Proof. Assume that there exists an interval $[a, b) \subset \overline{I}$ such that $\{x_n\} \notin [a, b)$ for all $n \in \mathbb{N}$. Then by definition of uniform distribution mod 1 with ϵ equal to $(b-a)/2$ we have that there exists an integer N such that

$$\left| \frac{A([a, b); N)}{N} - (b - a) \right| < \frac{b-a}{2}.$$

Since $\{x_n\} \notin [a, b)$ for all $n \in \mathbb{N}$, we have that $A([a, b); N) = 0$. Consequently $b - a < (b - a)/2$ which is absurd. This shows that the sequence $\{\{x_n\}\}_{n=1}^{\infty}$ is everywhere dense in $\overline{I}$. □

Example 4.1.5. If r is a rational number, then the sequence $\{nr\}_{n=1}^{\infty}$ is not u.d. mod 1. We shall see later on that if r is instead an irrational number then the sequence $\{nr\}$ is u.d. mod 1.

Proof. Let $r = p/q$ be a rational number with an integer p and a positive integer q. From Euclidean Algorithm, for each $n \in \mathbb{N}$, there exist integers s_n and t_n with $0 \leq t_n \leq q-1$ such that $rn = pn/q = s_n + t_n/q$. It follows that the fractional part of each term of the sequence $\{rn\}_{n=1}^{\infty}$ is an element of the set $\{0/q, 1/q, \ldots, (q-1)/q\}$ which is a finite set and is therefore not dense in $\overline{I}$. Thus the sequence $\{\{rn\}\}_{n=1}^{\infty}$ is also not dense in $\overline{I}$. Hence by Proposition 4.1.4, the sequence $\{nr\}_{n=1}^{\infty}$ is not u.d. mod 1. □

We note that the condition (4.1.1) can be stated in terms of characteristic

function by

$$\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}\mathbb{1}_{[a,b)}(\{x_n\})=\int_0^1\mathbb{1}_{[a,b)}(x)\,dx. \tag{4.1.4}$$

This observation leads to the following theorem which is a criterion to determine whether a real sequence is u.d. mod 1.

Theorem 4.1.6. *The sequence $\{x_n\}_{n=1}^{\infty}$ of real numbers is u.d. mod 1 if and only if for every real-valued continuous function f defined on the closed unit interval $\overline{I}=[0,1]$ we have*

$$\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}f(\{x_n\})=\int_0^1 f(x)\,dx. \tag{4.1.5}$$

Proof. Let $\{x_n\}_{n=1}^{\infty}$ be u.d. mod 1 and let $f(x)=\sum_{j=0}^{k-1}d_j\mathbb{1}_{[a_j,a_{j+1})}(x)$ be a step function on $\overline{I}$, where $0=a_0<a_1<\cdots<a_k=1$. Then it follows from (4.1.4) that for every such function f equation (4.1.5) holds. We assume now that f is a real-valued continuous function defined on $\overline{I}$. Let $\epsilon>0$. Then there exists, by definition of the Riemann integral, two step functions f_1 and f_2 such that $f_1(x)\le f(x)\le f_2(x)$ for all $x\in\overline{I}$ and $\int_0^1(f_2(x)-f_1(x))\,dx\le\epsilon$. Then we have the following chain of inequalities:

$$\begin{aligned}\int_0^1 f(x)\,dx-\epsilon &\le \int_0^1 f_1(x)\,dx=\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}f_1(\{x_n\})\\ &\le \liminf_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}f(\{x_n\})\le\limsup_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}f(\{x_n\})\\ &\le \lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}f_2(\{x_n\})=\int_0^1 f_2(x)\,dx\le\int_0^1 f(x)\,dx+\epsilon.\end{aligned}$$

Therefore in the case of a continuous function f the relation (4.1.5) holds. Conversely, let a sequence $\{x_n\}_{n=1}^{\infty}$ be given, and suppose that (4.1.5) holds for every real-valued continuous function f on $\overline{I}$. Let $[a,b)$ be an arbitrary subinterval

of $\overline{I}$. Let $\epsilon > 0$. Then there exist two continuous functions g_1 and g_2 such that $g_1(x) \le \mathbb{1}_{[a,b)}(x) \le g_2(x)$ for all $x \in \overline{I}$ and $\int_0^1 (g_2(x) - g_1(x))\,dx \le \epsilon$. Then we have

$$\begin{aligned} b - a - \epsilon &\le \int_0^1 g_2(x)\,dx - \epsilon \le \int_0^1 g_1(x)\,dx = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} g_1(\{x_n\}) \\ &\le \liminf_{N\to\infty} \frac{A([a,b);N)}{N} \le \limsup_{N\to\infty} \frac{A([a,b);N)}{N} \le \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} g_2(\{x_n\}) \\ &= \int_0^1 g_2(x)\,dx \le \int_0^1 g_1(x)\,dx + \epsilon \le b - a + \epsilon. \end{aligned}$$

Since ϵ was arbitrary, we have condition (4.1.1). □

Corollary 4.1.7. *The sequence $\{x_n\}$ is u.d. mod 1 if and only if for every Riemann-integrable function f on $\overline{I}$ equation (4.1.5) holds.*

Proof. We note that since every real-valued continuous function is Riemann-integrable, the sufficient statement holds. The necessary statement follows by the same argument as in the first part of the proof of Theorem 4.1.6. □

Corollary 4.1.8. *The sequence $\{x_n\}_{n=1}^{\infty}$ is u.d. mod 1 if and only if for every complex-valued continuous function f on $\mathbb{R}$ with period 1 we have*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} f(x_n) = \int_0^1 f(x)\,dx. \tag{4.1.6}$$

Proof. For the necessity part, by applying Theorem 4.1.6 to the real and imaginary part of f, one shows first that (4.1.5) also holds for complex-valued function f. However, the periodicity condition implies $f(\{x_n\}) = f(x_n)$, and so we arrive at (4.1.6). As to the sufficiency of (4.1.6), we need only note that in the second part of the proof of Theorem 4.1.6 the functions g_1 and g_2 can be chosen in such a way that they satisfy the additional requirements $g_1(0) = g_1(1)$ and $g_2(0) = g_2(1)$, so that (4.1.6) can be applied to the periodic extensions of g_1 and g_2 to $\mathbb{R}$. □

Example 4.1.9. Let $\{x_n\}_{n=1}^{\infty}$ be u.d. mod 1. Then the relation (4.1.5) is not valid for every Lebesgue-integrable function f on $\overline{I}$.

Proof. Let $f = \mathbb{1}_{\mathbb{Q}\cap[0,1]}$. Then we see that f is Lebesgue-integrable with $\int_0^1 f(x)\,dx = \mu(\mathbb{Q}\cap[0,1]) = 0$. We note also that f is not Riemann-integrable. Choose a sequence $\{x_n\}$ that is u.d. mod 1 and $x_n \in \mathbb{Q}$ for all $n \in \mathbb{N}$, e.g., the sequence in Example 4.1.2. With this function f we have $f(\{x_n\}) = 1$ for all $n \in \mathbb{N}$. So

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} f(\{x_n\}) = 1 \neq 0 = \int_0^1 f(x)\,dx.$$

□

We record one more result about the property of u.d. mod 1 sequence before we move on to the second section on the Weyl criterion.

Theorem 4.1.10. *If the sequence $\{x_n\}_{n=1}^{\infty}$ is u.d. mod 1 and if $\{y_n\}_{n=1}^{\infty}$ is a sequence with the property that $\lim_{n\to\infty}(x_n - y_n) = \alpha$, a real constant, then $\{y_n\}_{n=1}^{\infty}$ is u.d. mod 1.*

Proof. See [17] for proof. □

4.2 The Weyl criterion

Considered perhaps one of the most important facts in the theory of uniform distribution modulo 1, the Weyl criterion is used to determine whether a real sequence is u.d. mod 1. This criterion features one of the most versatile functions in analysis, that is the exponential functions $f(x) = e^{2\pi i h x}$, where h is a nonzero integer. We note that functions f satisfy the necessary condition of Corollary 4.1.8.

The Weyl criterion states that these functions suffice to determine the u.d. mod 1 of a sequence.

Theorem 4.2.1 (Weyl Criterion). *The sequence $\{x_n\}_{n=1}^{\infty}$ is u.d. mod 1 if and only if*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} e^{2\pi i h x_n} = 0 \quad \text{for all integers } h \neq 0. \tag{4.2.1}$$

Proof. The necessity follows from Corollary 4.1.8. Now suppose that $\{x_n\}_{n=1}^{\infty}$ possesses property (4.2.1). Then we shall show that (4.1.6) is valid for every complex-valued continuous function f on $\mathbb{R}$ with period 1. Let $\epsilon > 0$. Then by the Weierstrass approximation theorem, there exists a trigonometric polynomial $\Psi(x)$, that is, $\Psi(x) = a_0 + \sum_{k=1}^{K} a_k e^{2\pi i h_k x}$ for some $a_0, \ldots, a_K \in \mathbb{C}$ and $h_1, \ldots, h_K \in \mathbb{Z} \setminus \{0\}$, such that

$$\sup_{0\le x\le 1} |f(x) - \Psi(x)| \le \frac{\epsilon}{2}. \tag{4.2.2}$$

A straightforward calculation yields that $\int_0^1 \Psi(x)\,dx = a_0$. We have the following inequalities:

$$\begin{aligned}
\left| \int_0^1 f(x)\,dx - \frac{1}{N}\sum_{n=1}^{N} f(x_n) \right| &\le \left| \int_0^1 f(x)\,dx - \int_0^1 \Psi(x)\,dx \right| \\
&\quad + \left| \int_0^1 \Psi(x)\,dx - \frac{1}{N}\sum_{n=1}^{N} \Psi(x_n) \right| \\
&\quad + \left| \frac{1}{N}\sum_{n=1}^{N} \Psi(x_n) - \frac{1}{N}\sum_{n=1}^{N} f(x_n) \right| \\
&\le \frac{\epsilon}{2} + \frac{\epsilon}{2} + \left| a_0 - a_0 + \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K} a_k e^{2\pi i h_k x_n} \right| \\
&= \epsilon + \left| \sum_{k=1}^{K} a_k \frac{1}{N}\sum_{n=1}^{N} e^{2\pi i h_k x_n} \right|.
\end{aligned}$$

Passing on the limit supremum, we obtain

$$\begin{aligned}\limsup_{N\to\infty}\left|\int_0^1 f(x)\,dx-\frac{1}{N}\sum_{n=1}^{N}f(x_n)\right| &\leq \epsilon+\limsup_{N\to\infty}\left|\sum_{k=1}^{K}a_k\frac{1}{N}\sum_{n=1}^{N}e^{2\pi ih_kx_n}\right| \\ &\leq \epsilon+\sum_{k=1}^{K}a_k\lim_{N\to\infty}\left|\frac{1}{N}\sum_{n=1}^{N}e^{2\pi ih_kx_n}\right| \\ &= \epsilon.\end{aligned}$$

Since ϵ is arbitrarily small, we have that

$$\limsup_{N\to\infty}\left|\int_0^1 f(x)\,dx-\frac{1}{N}\sum_{n=1}^{N}f(x_n)\right|=0,$$

and so does the limit infimum. Therefore

$$\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}f(x_n)=\int_0^1 f(x)\,dx.$$

It follows from Corollary 4.1.8 that $\{x_n\}_{n=1}^{\infty}$ is u.d. mod 1. Hence the sufficiency follows. □

Example 4.2.2. Let θ be an irrational number. We show, using the Weyl criterion, that the sequence $\{n\theta\}_{n=1}^{\infty}$ is u.d. mod 1. Let h be a nonzero integer. We have

$$\begin{aligned}\left|\frac{1}{N}\sum_{n=1}^{N}e^{2\pi ihn\theta}\right| &= \frac{1}{N}\left|\frac{e^{2\pi ih\theta}(1-e^{2\pi ihN\theta})}{1-e^{2\pi ih\theta}}\right| \\ &\leq \frac{1}{N\left|\sin\pi h\theta\right|}. \qquad (4.2.3)\end{aligned}$$

We note that since θ is irrational, $\sin\pi h\theta\neq 0$. So the right side of the inequality approaches zero as N approaches infinity. Hence by Theorem 4.2.1 the sequence $\{n\theta\}_{n=1}^{\infty}$ is u.d. mod 1.

Example 4.2.3. We show that the sequence $\{\log n\}_{n=1}^{\infty}$ is not u.d. mod 1. This is an interesting fact because we will see later that there is a subsequence $\{x_n\}_{n=1}^{\infty}$ of

natural numbers such that $\{\log x_n\}_{n=1}^{\infty}$ is u.d. mod 1. An example of such sequence $\{x_n\}_{n=1}^{\infty}$ is the Fibonacci sequence $\{F_n\}_{n=1}^{\infty}$, i.e., $F_1 = F_2 = 1$ and $F_n = F_{n-1} + F_{n-2}$ for $n \geq 3$. To show that $\{\log n\}_{n=1}^{\infty}$ is not u.d. mod 1, we need the Euler summation formula which states the following. If $F(t)$ is a complex-valued function with a continuous derivative on $1 \leq t \leq N$, where $N \in \mathbb{N}$, then

$$\sum_{n=1}^{N} F(n) = \int_1^N F(t)\,dt + \frac{1}{2}\big(F(1) + F(N)\big) + \int_1^N \big(\{t\} - \frac{1}{2}\big)F'(t)\,dt. \qquad (4.2.4)$$

In this case we let $F(t) = e^{2\pi i \log t}$. Then we see that the first term on the right of (4.2.4) divided by N is equal to

$$\frac{Ne^{2\pi i \log N} - 1}{N(2\pi i + 1)} = \frac{N\big(\cos(2\pi \log N) + i \sin(2\pi \log N)\big) - 1}{N(2\pi i + 1)},$$

which does not converge as $N \to \infty$. The second term on the right of (4.2.4) divided by N converges to zero as $N \to \infty$. Finally the last term on the right of (4.2.4) divided by N also converges to zero as $N \to \infty$. This can be seen by noticing that $F'(t) = e^{2\pi i \log t} 2\pi i/t$ and $0 \leq \{t\} < 1$, so that

$$\begin{aligned} \left| \int_1^N \big(\{t\} - \frac{1}{2}\big)F'(t)\,dt \right| &\leq \int_1^N \left| \{t\} - \frac{1}{2} \right| |F'(t)|\,dt \\ &\leq \int_1^N \frac{1}{2} \cdot \frac{2\pi}{t}\,dt = \pi \log N. \end{aligned}$$

This shows that (4.2.1) with $x_n = \log n$ and $h = 1$ is not satisfied and therefore $\{\log n\}_{n=1}^{\infty}$ is not u.d. mod 1.

Remark 4.2.4. We note that Euler summation formula (4.2.4) has various applications in analytic number theory. It is used for example in the standard proof of The Prime Number Theorem. Another application of the formula includes the proof of

the Stirling's formula which states that

$$\lim_{n\to\infty} \frac{n!}{\sqrt{2\pi n}}\big(\frac{e}{n}\big)^n = 1.$$

(See, e.g., pp. 288 in [21] for proof.) This formula has an application in combinatorics as it is used to approximate $n!$. We have found another application of (4.2.4) by using it to create the following curious looking integral:

$$\int_0^\infty \frac{x\lfloor x\rfloor}{4x^4+1}\,dx = \frac{\pi}{16}.$$

Theorem 4.2.5. *If a sequence $\{x_n\}_{n=1}^\infty$ has the property that*

$$\Delta x_n = x_{n+1} - x_n \to \theta \text{ (irrational) as } n \to \infty\ , \tag{4.2.5}$$

then the sequence $\{x_n\}_{n=1}^\infty$ is u.d. mod 1.

Proof. Let q be a positive integer, then by (4.2.5) there exists an integer $g_0 = g_0(q)$ such that for any integers $n > g \geq g_0$ and integer $k \geq 0$,

$$|\Delta x_j - \theta| \leq \frac{1}{q^2} \quad (j = g+kq, g+1+kq, \ldots, n-1+kq).$$

Since $\sum_{j=a}^{b-1} \Delta x_j = x_b - x_a$ for $1 \leq a < b$, we have for each integer $k \geq 0$,

$$\begin{aligned} |x_{n+kq} - x_{g+kq} - (n-g)\theta| &= \left|\sum_{j=g+kq}^{n-1+kq} (\Delta x_j - \theta)\right| \\ &\leq \sum_{j=g+kq}^{n-1+kq} |\Delta x_j - \theta| \\ &\leq \sum_{j=g+kq}^{n-1+kq} \frac{1}{q^2} = \frac{n-g}{q^2}. \end{aligned} \tag{4.2.6}$$

For arbitrary real numbers u and v, we have

$$\begin{aligned} \left|e^{2\pi iu} - e^{2\pi iv}\right| = \left|e^{2\pi i(u-v)} - 1\right| &= \left|e^{\pi i(u-v)} - e^{-\pi i(u-v)}\right| \\ &= |2i\sin\pi(u-v)| \leq 2\pi\,|u-v|\,. \end{aligned}$$

Let h be a nonzero integer. Then combining the fact above with (4.2.6) we obtain for each integer $k \geq 0$,

$$\begin{aligned}\left|e^{2\pi i h x_{n+kq}} - e^{2\pi i h(x_{g+kq}+(n-g)\theta)}\right| &\leq 2\pi\,|h|\,|x_{n+kq} - x_{g+kq} - (n-g)\theta| \\ &\leq \frac{2\pi\,|h|\,(n-g)}{q^2}.\end{aligned}$$

From (4.2.3) and the triangle inequality, we have for each integer $k \geq 0$,

$$\begin{aligned}\left|\sum_{n=g}^{g+q-1} e^{2\pi i h x_{n+kq}}\right| &\leq \left|\sum_{n=g}^{g+q-1} e^{2\pi i h(x_{g+kq}-(n-g)\theta)}\right| \\ &\quad + \left|\sum_{n=g}^{g+q-1} e^{2\pi i h x_{n+kq}} - \sum_{n=g}^{g+q-1} e^{2\pi i h(x_{g+kq}-(n-g)\theta)}\right| \\ &\leq \left|\sum_{n=g}^{g+q-1} e^{2\pi i h(x_{g+kq}-(n-g)\theta)}\right| \\ &\quad + \sum_{n=g}^{g+q-1} \left|e^{2\pi i h x_{n+kq}} - e^{2\pi i h(x_{g+kq}-(n-g)\theta)}\right| \\ &\leq \left|\sum_{n=g}^{g+q-1} e^{2\pi i h(x_{g+kq}-(n-g)\theta)}\right| + \frac{2\pi\,|h|}{q^2}\sum_{n=g}^{q+g-1}(n-g) \\ &= \left|\sum_{n=g}^{g+q-1} e^{2\pi i h\theta}\right| + \frac{2\pi\,|h|}{q^2}\cdot\frac{q(q-1)}{2} \leq K,\end{aligned}$$

where $K = \frac{1}{|\sin \pi h\theta|} + \pi\,|h|$. For every positive integer H, we have

$$\begin{aligned}\left|\sum_{n=g}^{g-1+Hq} e^{2\pi i h x_n}\right| &= \left|\sum_{n=g}^{g-1+q} e^{2\pi i h x_n} + \sum_{n=g+q}^{g-1+2q} e^{2\pi i h x_n} + \cdots + \sum_{n=g+(H-1)q}^{g-1+Hq} e^{2\pi i h x_n}\right| \\ &= \left|\sum_{n=g}^{g-1+q} e^{2\pi i h x_n} + \sum_{n=g}^{g-1+q} e^{2\pi i h x_{n+q}} + \cdots + \sum_{n=g}^{g-1+q} e^{2\pi i h x_{n+(H-1)q}}\right| \\ &\leq \sum_{k=0}^{H-1}\left|\sum_{n=g}^{g-1+q} e^{2\pi i h x_{n+kq}}\right| \\ &\leq \sum_{k=0}^{H-1} K = HK.\end{aligned}$$

Let $N \geq g$ be an integer. Choose the largest integer H_N such that $g-1+H_N q < N$.

It follows that

$$g-1+H_Nq \leq N-1 \Rightarrow H_N \leq \frac{N-g}{q}$$

and

$$g-1+(H_N+1)q \leq N \Rightarrow N-g-H_Nq+1 \leq q.$$

With these facts we have

$$\begin{aligned}\left|\sum_{n=1}^{N} e^{2\pi i h x_n}\right| &= \left|\sum_{n=1}^{g-1} e^{2\pi i h x_n} + \sum_{n=g}^{g-1+H_Nq} e^{2\pi i h x_n} + \sum_{n=g+H_Nq}^{N} e^{2\pi i h x_n}\right| \\ &\leq (g-1)+H_NK+(N-g-H_Nq+1) \\ &\leq g-1+\frac{N-g}{q}K+q.\end{aligned}$$

Keeping q fixed, we obtain

$$\limsup_{N\to\infty}\left|\frac{1}{N}\sum_{n=1}^{N} e^{2\pi i h x_n}\right| \leq \frac{K}{q}.$$

Since q can be made arbitrarily large, we have

$$\limsup_{N\to\infty}\left|\frac{1}{N}\sum_{n=1}^{N} e^{2\pi i h x_n}\right| = 0,$$

and hence so does the limit infimum. Therefore

$$\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N} e^{2\pi i h x_n} = 0.$$

Hence by Theorem 4.2.1, $\{x_n\}_{n=1}^{\infty}$ is u.d. mod 1. □

Example 4.2.6. We show that for each irrational number θ, $\{x_n = n\theta + \sqrt{n}\}_{n=1}^{\infty}$ is u.d. mod 1. We have

$$x_{n+1}-x_n = \theta+\sqrt{n+1}-\sqrt{n} \to \theta \quad \text{as} \quad n\to\infty,$$

so that the hypothesis of Theorem 4.2.5 is satisfied and therefore the sequence $\{x_n\}_{n=1}^{\infty}$ is u.d. mod 1.

Example 4.2.7. We show that $\{\log F_n\}_{n=1}^{\infty}$ is u.d. mod 1. It is a well-known fact that F_{n+1}/F_n converges to the golden ratio $\tau = (1+\sqrt{5})/2$ as $n \to \infty$. We then have

$$\log F_{n+1} - \log F_n = \log\big(\frac{F_{n+1}}{F_n}\big) \to \log \tau \quad \text{as} \quad n \to \infty.$$

Since $\log \tau$ is an irrational number, it follows from Theorem 4.2.5 that the sequence $\{\log F_n\}_{n=1}^{\infty}$ is u.d. mod 1.

Example 4.2.8. Let θ be an irrational number, and let a and d be integers with $a \geq 0$ and $d > 0$. For $n \geq 1$, one sets $\epsilon_n = 1$ if the integer closest to $n\theta$ is to the left of $n\theta$; otherwise, $\epsilon_n = 0$. Then we have

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \epsilon_{a+nd} = \frac{1}{2}.$$

Proof. We prove the case when $a = 0$ and $d = 1$. From Example 4.2.2 we know that the sequence $\omega = \{n\theta\}_{n=1}^{\infty}$ is u.d. mod 1. We also note that the definition of ϵ_n can be rewritten as $\epsilon_n = 1$ if $n\theta - \lfloor n\theta \rfloor = \{n\theta\} < 1/2$ and $\epsilon_n = 0$ otherwise.We then have

$$\begin{aligned} \frac{1}{N}\sum_{n=1}^{N} \epsilon_n &= \frac{1}{N}\Big(\sum_{\{n\theta\}<1/2} \epsilon_n + \sum_{\{n\theta\}\geq 1/2} \epsilon_n\Big) \\ &= \frac{1}{N}\sum_{\{n\theta\}<1/2} 1 = \frac{1}{N} A\big([0,\frac{1}{2}); N; \omega\big). \end{aligned}$$

Since ω is u.d. mod 1, the last quantity, by definition, tends to $1/2 - 0 = 1/2$ as $N \to \infty$. For general case $a \geq 0$ and $d > 0$, we check that the sequence $\{(a+nd)\theta\}_{n=1}^{\infty}$ is u.d. mod 1 and then proceed the proof as above. In fact we have for each nonzero integer h,

$$\left|\frac{1}{N}\sum_{n=1}^{N} e^{2\pi i h(a+nd)\theta}\right| = \left|\frac{1}{N}\sum_{n=1}^{N} e^{2\pi i (hd) n\theta}\right|$$

which, by Theorem 4.2.1, tends to 0 as $N \to \infty$, since hd is a nonzero integer and $\{n\theta\}_{n=1}^{\infty}$ is u.d. mod 1. □

4.3 Approximation of exponential sums

Many well-known mathematicians such as Gauss, Weyl, Vinogradov, and van der Corput, to name a few, have studied exponential sums and contributed significant results in the subject. However, perhaps one of the most famous exponential sums is the Gaussian sum. This is an exponential sum of the form

$$S(p,q) = \sum_{n=0}^{q-1} e^{-\pi i n^2 \frac{p}{q}},$$

where p and q are relatively prime integers and $q > 0$. Gauss showed that

$$S(-2,q) = \frac{1-i^q}{1-i}\sqrt{q},$$

where q is odd. (See pp. 235-237 in [13] for an interesting proof by Dirichlet using Fourier Analysis.) Most of the exponential sums cannot be derived into a simple form. Number theorists therefore try to come up with a good way to approximate an exponential sum. One efficient theorem is given by van der Corput.

Theorem 4.3.1 (van der Corput). *If a and b are integers with $a < b$ and if f is a twice differentiable function on $[a,b]$ with $f''(x) \geq \rho > 0$ for all $x \in [a,b]$ or $f''(x) \leq -\rho < 0$ for all $x \in [a,b]$, then*

$$\left| \sum_{n=a}^{b} e^{2\pi i f(n)} \right| \leq (|f'(b) - f'(a)| + 2)\Big(\frac{4}{\sqrt{\rho}} + 3\Big). \tag{4.3.1}$$

We need some lemmata to prove Theorem 4.3.1.

Lemma 4.3.2. *Suppose the real-valued function f has a monotone derivative f' on $[a,b]$ with $|f'(x)| \geq \lambda > 0$ for $x \in [a,b]$. Then, if $J = \int_a^b e^{2\pi i f(x)}\, dx$, we have $|J| < 1/\lambda$.*

Proof. We rewrite J as

$$J = \frac{1}{2\pi i} \int_a^b \frac{de^{2\pi i f(x)}}{f'(x)}.$$

Then since f' is monotone on $[a,b]$, it follows from the Second Mean Value Theorem (see, e.g., [22], pp. 279) that there exists $x_0 \in [a,b]$ such that

$$J = \frac{1}{2\pi}\Big(\frac{1}{f'(a)} \int_a^{x_0} de^{2\pi i f(x)} + \frac{1}{f'(b)} \int_{x_0}^b de^{2\pi i f(x)}\Big),$$

so that

$$\begin{aligned} |J| &\leq \frac{1}{2\pi}\Big(\frac{1}{|f'(a)|}\left|e^{2\pi i f(x_0)} - e^{2\pi i f(a)}\right| + \frac{1}{|f'(b)|}\left|e^{2\pi i f(b)} - e^{2\pi i f(x_0)}\right|\Big) \\ &\leq \frac{1}{2\pi}\big(\frac{2}{|f'(a)|} + \frac{2}{|f'(b)|}\big) \leq \frac{2}{\pi\lambda} < \frac{1}{\lambda}. \end{aligned}$$

□

Lemma 4.3.3. *Let f be twice differentiable on $[a,b]$ with $f''(x) \geq \rho > 0$ for $x \in [a,b]$ or $f''(x) \leq -\rho < 0$ for $x \in [a,b]$. Then the integral J from Lemma 4.3.2 satisfies $|J| < 4/\sqrt{\rho}$.*

Proof. We assume that $f''(x) \geq \rho > 0$ for $x \in [a,b]$; otherwise, we replace f by $-f$. So f' is increasing. Suppose for the moment that f' is of constant sign in $[a,b]$. That is either $f'(x) \geq 0$ for all $x \in [a,b]$ or $f'(x) \leq 0$ for all $x \in [a,b]$. We consider each of these cases. Let c be fixed with $a < c < b$. For the case $f' \geq 0$ on $[a,b]$, applying the Mean Value Theorem on $[a,x]$ for $x \in [c,b]$ yields that there exists ξ_x in (a,x)

such that

$$f'(x) - f'(a) = (x-a)f''(\xi_x) \geq (x-a)\rho,$$

so that

$$f'(x) \geq (x-a)\rho + f'(a) \geq (x-a)\rho > (c-a)\rho > 0.$$

Applying Lemma 4.3.2 with $\int_c^b e^{2\pi i f(x)}\,dx$, we have

$$|J| \leq \left|\int_a^c e^{2\pi i f(x)}\,dx\right| + \left|\int_c^b e^{2\pi i f(x)}\,dx\right| < (c-a) + \frac{1}{(c-a)\rho}. \tag{4.3.2}$$

We note that the last quantity is minimized when $c = a + 1/\sqrt{\rho}$. Hence, with this value of c

$$|J| < \frac{2}{\sqrt{\rho}}.$$

For the case $f' \leq 0$ on $[a,b]$, applying the Mean Value Theorem on $[x,b]$ for $x \in [a,c]$ yields that there exists η_x in (x,b) such that

$$f'(b) - f'(x) = (b-x)f''(\eta_x) \geq (b-x)\rho,$$

so that

$$f'(x) \leq -(b-x)\rho + f'(b) \leq -(b-x)\rho \leq -(b-c)\rho < 0.$$

Applying Lemma 4.3.2 with $\int_a^c e^{2\pi i f(x)}\,dx$, we have

$$|J| \leq \left|\int_a^c e^{2\pi i f(x)}\,dx\right| + \left|\int_c^b e^{2\pi i f(x)}\,dx\right| < \frac{1}{(b-c)\rho} + (b-c). \tag{4.3.3}$$

We note that the last quantity is minimized when $c = b - 1/\sqrt{\rho}$. Hence, with this value of c

$$|J| < \frac{2}{\sqrt{\rho}}.$$

In the general case, $[a, b]$ is the union of two intervals in each of which f' is of constant sign, and the desired inequality follows by adding the inequalities for these two intervals and hence obtaining the bound $4/\sqrt{\rho}$. □

Remark 4.3.4. We note that the value c that minimizes the last quantity in either (4.3.2) or (4.3.3) should be contained in (a, b). If this is not the case then we replace c by either b or a whichever appropriate and the bound of J will be replaced by $C/\sqrt{\rho}$ for some absolute constant $C > 0$.

Lemma 4.3.5. *Let f' be monotone on $[a, b]$ with $|f'(x)| \leq 1/2$ for $a \leq x \leq b$. Then, if $J_1 = \int_a^b (\{x\} - 1/2) de^{2\pi i f(x)}$, we have*

$$|J_1| \leq 2.$$

Proof. Since the function $x \mapsto \{x\} - 1/2$ is odd and periodic with period 1, we obtain the Fourier series of this function by

$$\{x\} - \frac{1}{2} = \sum_{n=1}^{\infty} b_n \sin 2\pi n x,$$

valid for all $x \notin \mathbb{Z}$ and b_n is given by

$$b_n = 2 \int_0^1 \big(\{x\} - \frac{1}{2} \big) \sin 2\pi n x \, dx = -\frac{1}{\pi n}.$$

Let $S_N(x)$ denote the Nth partial sum of the series, i.e., $S_N(x) = -\sum_{n=1}^{N} \frac{\sin 2\pi n x}{\pi n}$ for $x \in \mathbb{R}$. Then by summation by parts one can show that S_N is uniformly bounded. Therefore,

$$J_1 = \lim_{N \to \infty} \int_a^b S_N(x) de^{2\pi i f(x)}. \tag{4.3.4}$$

For each integer $N \geq 1$ we have

$$\begin{aligned}
\int_a^b S_N(x) de^{2\pi i f(x)} &= \sum_{n=1}^{N} \frac{1}{n} \int_a^b (-2i \sin 2\pi n x) e^{2\pi i f(x)} f'(x)\, dx \\
&= \sum_{n=1}^{N} \frac{1}{n} \int_a^b (e^{-2\pi i n x} - e^{2\pi i n x}) e^{2\pi i f(x)} f'(x)\, dx \\
&= \frac{1}{2\pi i} \sum_{n=1}^{N} \frac{1}{n} \Big(\int_a^b \frac{f'(x)}{f'(x) - n} de^{2\pi i (f(x) - nx)} - \int_a^b \frac{f'(x)}{f'(x) + n} de^{2\pi i (f(x) + nx)} \Big).
\end{aligned}$$

Since the functions $f'/(f' \pm n)$ are monotone and $|f'| \leq 1/2$, an application of the Second Mean Value Theorem yields that

$$\left| \int_a^b \frac{f'(x)}{f'(x) \pm n} de^{2\pi i (f(x) \pm nx)} \right| \leq \frac{2}{n - \frac{1}{2}},$$

so that

$$\begin{aligned}
\left| \int_a^b S_N(x) de^{2\pi i f(x)} \right| &\leq \frac{2}{\pi} \sum_{n=1}^{N} \frac{1}{n(n - \frac{1}{2})} \\
&= \frac{2}{\pi} \Big(2 + \sum_{n=2}^{N} \frac{1}{n(n - \frac{1}{2})} \Big) \\
&\leq \frac{2}{\pi} \Big(2 + \sum_{n=2}^{N} \frac{1}{n(n-1)} \Big) \\
&= \frac{2}{\pi} (2 + 1 - \frac{1}{N}) < \frac{6}{\pi} < 2.
\end{aligned}$$

Taking the limit $N \to \infty$ we obtain $|J_1| \leq 2$. □

Proof of Theorem 4.3.1. We write

$$\sum_{n=a}^{b} e^{2\pi i f(n)} = \sum_{p=-\infty}^{\infty} S_p \tag{4.3.5}$$

with

$$S_p = \sum_{\substack{a \leq n \leq b \\ p - 1/2 \leq f'(n) < p + 1/2}} e^{2\pi i f(n)}.$$

The sum over p in (4.3.5) is in fact a finite sum. Let p be an integer for which the sum S_p is nonvoid. By the assumption on f'', we have that f' is monotone, and therefore this sum S_p is over consecutive values of n, say from $n = a_p$ to $n = b_p$. With $F_p(x) = f(x) - px$, we get

$$\begin{aligned} S_p &= \sum_{n=a_p}^{b_p} e^{2\pi i f(n)} = \sum_{n=a_p}^{b_p} e^{2\pi i F_p(n)} \\ &= \int_{a_p}^{b_p} e^{2\pi i F_p(x)}\, dx + \frac{1}{2}\Big(e^{2\pi i F_p(a_p)} + e^{2\pi i F_p(b_p)}\Big) + \int_{a_p}^{b_p} \Big(\{x\} - \frac{1}{2}\Big)\, de^{2\pi i F_p(x)} \end{aligned}$$

by the Euler summation formula. Now the first integral is in absolute value less than $4/\sqrt{\rho}$ by Lemma 4.3.3. The last integral is in absolute value at most 2 because of the fact that $|F_p'(x)| \leq 1/2$ for $x \in [a_p, b_p]$ and of Lemma 4.3.5. Therefore,

$$|S_p| < (4/\sqrt{\rho}) + 3. \tag{4.3.6}$$

By counting the values of p such that

$$\min\{f'(a), f'(b)\} - \frac{1}{2} < p \leq \max\{f'(a), f'(b)\} + \frac{1}{2},$$

we obtain that there are at most $|f'(b) - f'(a)| + 1/2 - (-1/2) + 1 = |f'(b) - f'(a)| + 2$ values of p for which S_p is a nonvoid sum. This, (4.3.6), and (4.3.5) imply (4.3.1) and therefore the proof is complete. □

Example 4.3.6. To see how good the van der Corput Theorem is in approximating exponential sums, we experiment it with the Gaussian sum $S(-2, q)$ (q odd) described in the introduction of this section. A straightforward calculation yields that for each positive odd integer q,

$$|S(-2, q)| = \sqrt{q}.$$

To approximate $S(-2,q)$ using Theorem 4.3.1, we let $f(x) = -x^2/q$ so that $f'(x) = -2x/q$ and $f''(x) = -2/q$. We have $|f''(x)| = 2/q$ for all x so that $\rho = 2/q$. Moreover, $|f'(q-1) - f'(0)| = 2(q-1)/q$. Hence from Theorem 4.3.1,

$$|S(-2,q)| \leq \big(\frac{2(q-1)}{q} + 2\big)\big(\frac{4}{\sqrt{2/q}} + 3\big) < 8\sqrt{2}\sqrt{q} + 12 < 17\sqrt{q},$$

for all $q \geq 169$. We see that Theorem 4.3.1 approximates $|S(-2,q)|$ up to the asymptotic order of $\sqrt{q}$ which is considered acceptably good in applications.

4.4 Discrepancy

We have seen from the previous sections the sequences that are uniformly distributed. We have developed some good criteria to determine whether a given sequence is uniformly distributed. However, there are quite a few sequences that are not uniformly distributed. Among these sequences there might be some that are distributed "better" than the others. That is to say we are interested in measuring how well a given sequence is distributed comparing with a sequence that has uniform distribution which we consider as the ideal distribution. The quantity associated with the quality of the distribution of a sequence is called *discrepancy*. In this section we shall develop the notion of discrepancy and prove an important inequality, namely the Erdös-Turán Inequality, that is mainly used to approximate the discrepancy of a given sequence in terms of exponential sums.

Definition 4.4.1. Let $x_1, \ldots, x_N$ be a finite sequence of real numbers. The discrepancy of the given sequence, denoted $D_N(x_1, \ldots, x_N)$ or simply D_N, is defined

by

$$D_N = D_N(x_1, \ldots, x_N) = \sup_{0 \le \alpha < \beta \le 1} \left| \frac{A([\alpha, \beta); N)}{N} - (\beta - \alpha) \right|. \tag{4.4.1}$$

If $\omega = \{x_n\}_{n=1}^{\infty}$ is an infinite sequence of real numbers then we define $D_N(\omega)$ to be the discrepancy of the first N terms of the sequence.

It should be remarked that to compute the discrepancy of a sequence, we consider the supremum in (4.4.1) over all subintervals of the unit interval $I = [0, 1)$. Therefore, when we prove some assertions about discrepancy we can assume without loss of generality that all elements of a given sequence are contained in I.

The following theorem reflects the association of a uniform distribution to an ideal distribution.

Theorem 4.4.2. *The sequence ω is u.d. mod 1 if and only if*

$$\lim_{N \to \infty} D_N(\omega) = 0.$$

Proof. To prove the sufficiency, we let $[a, b)$ be an arbitrary interval in I and note that for each positive integer N,

$$\left| \frac{A([a, b); N)}{N} - (b - a) \right| \le \sup_{0 \le \alpha < \beta \le 1} \left| \frac{A([\alpha, \beta); N)}{N} - (\beta - \alpha) \right|.$$

Since as $N \to \infty$, the right-hand side quantity approaches 0 by assumption, it follows that

$$\lim_{N \to \infty} \left| \frac{A([a, b); N)}{N} - (b - a) \right| = 0$$

and therefore ω is u.d. mod 1. Now we show the necessity. Let $m \ge 2$ be a positive integer and let $I_k = [k/m, (k + 1)/m)$ for $k = 0, 1, \ldots, m - 1$, so that the the family

$\{I_k\}_{k=0}^{m-1}$ partitions the interval I. From the definition of u.d. mod 1, there exists an integer $N_0(m)$ such that for each integer $N \geq N_0$, and for each $k = 0, 1, \ldots, m-1$,

$$\frac{1}{m}\Big(1 - \frac{1}{m}\Big) \leq \frac{A(I_k; N)}{N} \leq \frac{1}{m}\Big(1 + \frac{1}{m}\Big). \tag{4.4.2}$$

Let $J = [\alpha, \beta)$ be an arbitrary interval in I. Then we see that there exist intervals J_1 and J_2 being finite unions of intervals I_k such that $J_1 \subseteq J \subseteq J_2$ with $\lambda(J) - \lambda(J_1) < 2/m$ and $\lambda(J_2) - \lambda(J) < 2/m$. We note that J_1 is possibly an empty interval. From (4.4.2) we have that for each $N \geq N_0$,

$$\lambda(J_1)\Big(1 - \frac{1}{m}\Big) \leq \frac{A(J_1; N)}{N} \leq \frac{A(J; N)}{N} \leq \frac{A(J_2; N)}{N} \leq \lambda(J_2)\Big(1 - \frac{1}{m}\Big)$$

from which it follows that

$$\Big(\lambda(J) - \frac{2}{m}\Big)\Big(1 - \frac{1}{m}\Big) < \frac{A(J; N)}{N} < \Big(\lambda(J) + \frac{2}{m}\Big)\Big(1 + \frac{1}{m}\Big).$$

Now since $\lambda(J) \leq 1$ we have

$$-\frac{3}{m} - \frac{2}{m^2} < \frac{A(J; N)}{N} - \lambda(J) < \frac{3}{m} + \frac{2}{m^2}.$$

Since the bounds are independent of interval J it follows that for each $N \geq N_0$,

$$D_N(\omega) \leq \frac{3}{m} + \frac{2}{m^2}.$$

By letting m be arbitrarily large, we have

$$\lim_{N \to \infty} D_N(\omega) = 0. \qquad \square$$

We note that the necessity of Theorem 4.4.2 asserts that if a sequence ω is u.d. mod 1, then $\lim_{N\to\infty} A([a, b); N)/N = b - a$ uniformly in all subintervals $[a, b)$ of I. The following theorem gives basic estimates for a discrepancy.

Theorem 4.4.3. *For any sequence of N numbers, we have*

$$\frac{1}{N} \leq D_N \leq 1.$$

Proof. The rightmost inequality follows by noting that for every arbitrary interval $[\alpha, \beta)$ in I,

$$\frac{A([\alpha, \beta); N)}{N} - (\beta - \alpha) \leq 1 - (\beta - \alpha) < 1.$$

To prove the leftmost inequality, we fix an element x in the sequence and let $\epsilon > 0$ be small such that $x + \epsilon \leq 1$. We then have

$$\frac{A([x, x + \epsilon); N)}{N} - \epsilon \geq \frac{1}{N} - \epsilon.$$

This implies $D_N \geq 1/N - \epsilon$. Since ϵ can be made arbitrarily small, it follows that

$$D_N \geq \frac{1}{N}.$$ □

Example 4.4.4. The leftmost inequality in Theorem 4.4.3 is in fact sharp. Consider the following sequence $0, 1/N, 2/N, \ldots, (N-1)/N$ in some order. Let $[\alpha, \beta)$ be an arbitrary interval in I. We try to compute $A([\alpha, \beta); N)$. From the terms of the sequence this amounts to counting integers k for $k = 0, 1, \ldots, N-1$ such that $\alpha < k/N \leq \beta$. Hence

$$A([\alpha, \beta); N) = N(\beta - \alpha) + \theta,$$

such that $0 \leq \theta \leq 1$. This means

$$\left| \frac{A([\alpha, \beta); N)}{N} - (\beta - \alpha) \right| = \frac{\theta}{N} \leq \frac{1}{N}.$$

Therefore $D_N \leq 1/N$. From Theorem 4.4.3 we have $D_N = 1/N$.

Computationally speaking, it might be easier to restrict the family of intervals over which the supremum is formed in the definition of discrepancy. One type of restriction is to consider only intervals $[0, \alpha)$ for all $0 < \alpha \leq 1$. This type of restriction will prove very useful in computing discrepancy. We define discrepancy associated to this restriction as follows.

Definition 4.4.5. Let $x_1, \ldots, x_N$ be a finite sequence of real numbers. We define the discrepancy D_N^* of this sequence by

$$D_N^* = D_N^*(x_1, \ldots, x_N) = \sup_{0<\alpha\leq 1} \left| \frac{A([0,\alpha); N)}{N} - \alpha \right|. \tag{4.4.3}$$

The following theorem relates the size of D_N^* with D_N.

Theorem 4.4.6. *For any sequence of N numbers, we have*

$$D_N^* \leq D_N \leq 2D_N^*.$$

Proof. The leftmost inequality is clear from the fact that $\{[0,\alpha) : 0 < \alpha \leq 1\} \subset \{[\alpha,\beta) : 0 \leq \alpha < \beta \leq 1\}$. To show the rightmost inequality, we let $[\alpha, \beta)$ be an arbitrary interval in I and observe that $A([\alpha,\beta); N) = A([0,\beta); N) - A([0,\alpha); N)$, so that

$$\left| \frac{A([\alpha,\beta); N)}{N} - (\beta - \alpha) \right| \leq \left| \frac{A([0,\beta); N)}{N} - \beta \right| + \left| \frac{A([0,\alpha); N)}{N} - \alpha \right|.$$

Since each term on the right-hand side is no greater than D_N^*, it follows that

$$\left| \frac{A([\alpha,\beta); N)}{N} - (\beta - \alpha) \right| \leq 2D_N^*.$$

This implies $D_N \leq 2D_N^*$ as we wish to show. □

Corollary 4.4.7. *The sequence ω is u.d. mod 1 if and only if*

$$\lim_{N\to\infty} D_N^*(\omega) = 0.$$

Proof. This follows immediately from Theorem 4.4.2 and Theorem 4.4.6 above. □

We note that when we compute D_N or D_N^* of a finite sequence, the order of the terms of the sequence does not matter; so we may order the terms of the sequence in increasing order.

One advantage of D_N^* over D_N is that one can actually compute it explicitly (involving only finitely many steps of computation) as stated by the following theorem.

Theorem 4.4.8. *Let $x_1 \le x_2 \le \cdots \le x_N$ be N numbers in I. Then their discrepancy D_N^* is given by*

$$\begin{aligned} D_N^* &= \max_{1\le i\le N} \max\left\{ \left| x_i - \frac{i}{N} \right|, \left| x_i - \frac{i-1}{N} \right| \right\} \\ &= \frac{1}{2N} + \max_{1\le i\le N} \left| x_i - \frac{2i-1}{2N} \right|. \end{aligned} \tag{4.4.4}$$

Proof. We let $x_0 = 0$ and $x_{N+1} = 1$ and observe that the distinct values of the numbers $x_i, 0 \le i \le N+1$, form a partition of interval $[0,1]$. Therefore,

$$\begin{aligned} D_N^* &= \max_{\substack{0\le i\le N \\ x_i<x_{i+1}}} \sup_{x_i<\alpha\le x_{i+1}} \left| \frac{A([0,\alpha);N)}{N} - \alpha \right| \\ &= \max_{\substack{0\le i\le N \\ x_i<x_{i+1}}} \sup_{x_i<\alpha\le x_{i+1}} \left| \frac{i}{N} - \alpha \right|. \end{aligned}$$

Now we note that whenever $x_i < x_{i-1}$, the function $g_i(\alpha) = |i/N - \alpha|$ attains its maximum in $[x_i, x_{i+1}]$ at one of the end points of the interval. Hence,

$$D_N^* = \max_{\substack{0\le i\le N \\ x_i<x_{i+1}}} \max\left\{ \left| \frac{i}{N} - x_i \right|, \left| \frac{i}{N} - x_{i+1} \right| \right\}. \tag{4.4.5}$$

Now we show that we may drop the restriction $x_i < x_{i+1}$ in the first maximum. Suppose we have $x_i < x_{i+1} = x_{i+2} = \cdots = x_{i+r} < x_{i+r+1}$ with some $r \geq 2$. The indices not admitted in the first maximum in (4.4.5) are the integers $i + j$ with $1 \leq j \leq r - 1$. We shall prove that the numbers

$$\left| \frac{i+j}{N} - x_{i+j} \right| \text{ and } \left| \frac{i+j}{N} - x_{i+j+1} \right|$$

for $1 \leq j \leq r - 1$, which are excluded in (4.4.5), are in fact dominated by numbers already occurring in (4.4.5). For $1 \leq j \leq r - 1$, we get by the same reasoning as above (by considering the function $h_{i+1}(y) = |y - x_{i+1}|$) that

$$\begin{aligned} \left| \frac{i+j}{N} - x_{i+j} \right| &= \left| \frac{i+j}{N} - x_{i+1} \right| < \max \left\{ \left| \frac{i}{N} - x_{i+1} \right|, \left| \frac{i+r}{N} - x_{i+1} \right| \right\} \\ &= \max \left\{ \left| \frac{i}{N} - x_{i+1} \right|, \left| \frac{i+r}{N} - x_{i+r} \right| \right\}, \end{aligned}$$

and both numbers in the last maximum occur in (4.4.5). The same argument can be applied for $|(i + j)/N - x_{i+j+1}|$, $1 \leq j \leq r - 1$. Thus, we may drop the restriction $x_i < x_{i+1}$ and arrive at

$$\begin{aligned} D_N^* &= \max_{0 \leq i \leq N} \max \left\{ \left| \frac{i}{N} - x_i \right|, \left| \frac{i}{N} - x_{i+1} \right| \right\} \\ &= \max_{1 \leq i \leq N} \max \left\{ \left| \frac{i}{N} - x_i \right|, \left| \frac{i-1}{N} - x_i \right| \right\}. \end{aligned}$$

The last equality is valid since the only terms we dropped are $|0/N - x_0|$ and $|N/N - x_{N+1}|$ which are both zero. We now show the second equality in (4.4.4). It suffices to show that

$$\max \left\{ \left| x_i - \frac{i}{N} \right|, \left| x_i - \frac{i-1}{N} \right| \right\} = \frac{1}{2N} + \left| x_i - \frac{2i-1}{2N} \right|.$$

We consider two cases. The first case is that $\max \left\{ \left| x_i - \frac{i}{N} \right|, \left| x_i - \frac{i-1}{N} \right| \right\} = \left| x_i - \frac{i}{N} \right|$. This means $\left| x_i - \frac{i-1}{N} \right| < \left| x_i - \frac{i}{N} \right|$ which is equivalent to $\left(x_i - \frac{i-1}{N} \right)^2 < \left(x_i - \frac{i}{N} \right)^2$

which, after expanding, is equivalent to

$$x_i < \frac{2i-1}{2N}.$$

Therefore,

$$0 < \frac{1}{2N} + \left| x_i - \frac{2i-1}{2N} \right| = \frac{1}{2N} + \frac{2i-1}{2N} - x_i = \frac{i}{N} - x_i = \left| \frac{i}{N} - x_i \right|.$$

The second case is that $\max\left\{ \left| x_i - \frac{i}{N} \right|, \left| x_i - \frac{i-1}{N} \right| \right\} = \left| x_i - \frac{i-1}{N} \right|$. By the same argument above we have

$$x_i > \frac{2i-1}{2N}.$$

Therefore,

$$0 < \frac{1}{2N} + \left| x_i - \frac{2i-1}{2N} \right| = \frac{1}{2N} + x_i - \frac{2i-1}{2N} = x_i - \frac{i-1}{N} = \left| x_i - \frac{i-1}{N} \right|. \quad \square$$

Example 4.4.9. For $n \geq 1$, let ω_n be the finite sequence

$$\frac{0}{n^2}, \frac{1}{n^2}, \frac{4}{n^2}, \ldots, \frac{(n-1)^2}{n^2}.$$

We show that

$$\lim_{n\to\infty} D_n^*(\omega_n) = \frac{1}{4}.$$

First let's compute $D_n^*(\omega_n)$ for n large. From the second equation of (4.4.4) we have

$$D_n^*(\omega_n) = \frac{1}{2n} + \max_{1\leq k\leq n} \left| \frac{(k-1)^2}{n^2} - \frac{2k-1}{2n} \right|.$$

Let $h_n(x) = (x-1)^2/n^2 - (2x-1)/2n$ for $x \in [1, n]$. We see that the graph of h_n is a parabola with vertex at $x = n/2 + 1$. If n is even then

$$\begin{aligned} D_n^*(\omega_n) &= \frac{1}{2n} + \max\left\{ |h(n/2+1)|, |h(1)|, |h(n)| \right\} \\ &= \frac{1}{2n} + \max\left\{ \frac{1}{4} + \frac{1}{2n}, \frac{1}{2n}, \frac{3}{2n} - \frac{1}{n^2} \right\} \\ &= \frac{1}{2n} + \frac{1}{4} + \frac{1}{2n} = \frac{1}{4} + \frac{1}{n} \to \frac{1}{4} \text{ as } n \to \infty. \end{aligned}$$

Similarly, if n is odd then

$$\begin{aligned} D_n^*(\omega_n) &= \frac{1}{2n} + \max\left\{\left|h((n-1)/2+1)\right|, \left|h(1)\right|, \left|h(n)\right|\right\} \\ &= \frac{1}{2n} + \max\left\{\frac{1}{4} + \frac{1}{2n} - \frac{1}{4n^2}, \frac{1}{2n}, \frac{3}{2n} - \frac{1}{n^2}\right\} \\ &= \frac{1}{2n} + \frac{1}{4} + \frac{1}{2n} - \frac{1}{4n^2} = \frac{1}{4} + \frac{1}{n} - \frac{1}{4n^2} \to \frac{1}{4} \text{ as } n \to \infty. \end{aligned}$$

Since in either case $D_n^*(\omega_n) \to 1/4$ as $n \to \infty$, we have the desired result.

Corollary 4.4.10. *For any sequence of N numbers in I we have*

$$D_N^* \geq \frac{1}{2N},$$

with equality only for the sequence $1/(2N), 3/(2N), \ldots, (2N-1)/(2N)$ or its rearrangements.

Proof. This follows immediately by looking at the second equality in (4.4.4). □

The discrepancy D_N^* of infinite sequence has different lower bounds than the one given by Corollary 4.4.10 as stated by the following theorem.

Theorem 4.4.11. *For any infinite sequence ω of real numbers, we have*

$$N D_N^*(\omega) > c \log N$$

for infinitely many positive integers N, where $c > 0$ is an absolute constant.

Proof. See [17], pp. 109 for proof. □

We present two types of inequalities that give the upper bounds of discrepancies in terms of exponential sums.

Theorem 4.4.12 (LeVeque's Inequality). *The discrepancy D_N of the finite sequence $x_1, \ldots, x_N$ in I satisfies*

$$D_N \leq \left(\frac{6}{\pi^2} \sum_{k=1}^{\infty} \frac{1}{k^2} \left| \frac{1}{N} \sum_{n=1}^{N} e^{2\pi i k x_n} \right|^2 \right)^{1/3}. \tag{4.4.6}$$

Proof. See [17], pp. 111 for proof. □

Remark 4.4.13. We remark that the constant $6/\pi^2$ in LeVeque's Inequality is best possible. In fact, let $x_1 = x_2 = \cdots = x_N = 0$. Then $D_N = 1$, and the right-hand side is equal to

$$\left(\frac{6}{\pi^2} \sum_{n=1}^{\infty} \frac{1}{n^2} \right)^{1/3} = 1.$$

Theorem 4.4.14 (Erdös-Turán Inequality). *For any finite sequence $x_1, \ldots, x_N$ of real numbers and any positive integer K, we have*

$$D_N \leq \frac{6}{K+1} + \frac{4}{\pi} \sum_{k=1}^{K} \left(\frac{1}{k} - \frac{1}{K+1} \right) \left| \frac{1}{N} \sum_{n=1}^{N} e^{2\pi i k x_n} \right|. \tag{4.4.7}$$

Proof. Without loss of generality, we assume that $x_1 \leq x_2 \leq \cdots \leq x_N$. Let

$$\Delta_N(x) = \frac{A([0,x); N)}{N} - x \quad \text{for } 0 \leq x \leq 1,$$

and extend this function with period 1 to $\mathbb{R}$. The function Δ_N can be written

explicitly as follows:

$$\Delta_N(x) = \begin{cases} -x & 0 \le x \le x_1, \\ 1/N - x & x_1 < x \le x_2, \\ \vdots & \\ k/N - x & x_k < x \le x_{k+1}, \\ \vdots & \\ N/N - x & x_N < x \le 1. \end{cases} \tag{4.4.8}$$

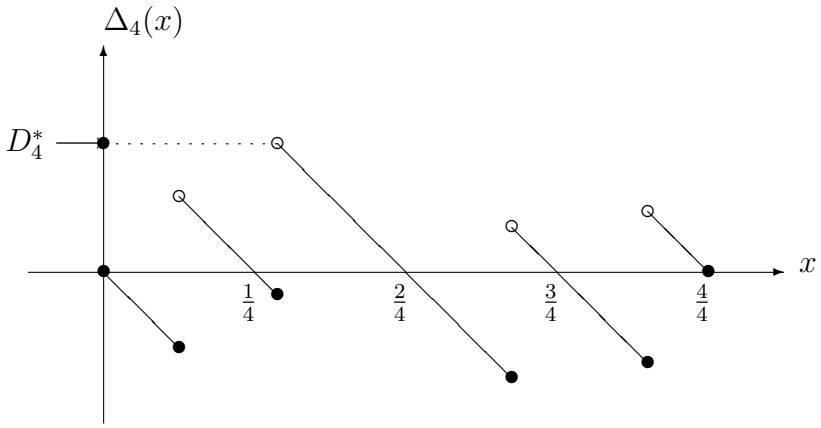

Figure 4.1: A typical graph of the function Δ_N. This particular graph draws Δ_4 of the sequence $x_1 = 1/8$, $x_2 = 2.3/8$, $x_3 = 5.4/8$, $x_4 = 7.2/8$.

We consider first a sequence $x_1, \ldots, x_N$ in I for which

$$\int_0^1 \Delta_N(x)\, dx = 0. \tag{4.4.9}$$

For convenience we also let

$$S_h = \frac{1}{N} \sum_{n=1}^{N} e^{2\pi i h x_n} \quad \text{for } h \in \mathbb{Z}.$$

We note that by computing the Fourier coefficients of Δ_N, we have

$$\frac{S_h}{-2\pi ih} = \int_0^1 \Delta_N(x) e^{2\pi ihx}\, dx \quad \text{for } h \in \mathbb{Z} \smallsetminus \{0\}. \tag{4.4.10}$$

In fact we note that $A([0,x);N) = \sum_{n=1}^N \mathbb{1}_{(x_n,1]}(x)$, so that for every nonzero integer h, we have

$$\begin{aligned}
\int_0^1 \Delta_N(x) e^{2\pi ihx}\, dx &= \frac{1}{N}\sum_{n=1}^N \int_0^1 \mathbb{1}_{(x_n,1]}(x) e^{2\pi ihx}\, dx - \int_0^1 x e^{2\pi ihx}\, dx \\
&= \frac{1}{N}\sum_{n=1}^N \int_{x_n}^1 e^{2\pi ihx}\, dx - \frac{1}{2\pi ih} \\
&= \frac{1}{2\pi ihN}\sum_{n=1}^N \left(1 - e^{2\pi ihx_n}\right) - \frac{1}{2\pi ih} \\
&= -\frac{1}{2\pi ihN}\sum_{n=1}^N e^{2\pi ihx_n}.
\end{aligned}$$

Choose a positive integer m, and let a be a real number to be determined later. From (4.4.9) and (4.4.10) it follows that

$$\begin{aligned}
&\sum_{h=-m}^{m}{}^{*} (m+1-|h|) e^{-2\pi iha} \frac{S_h}{-2\pi ih} \\
&\qquad = \int_0^1 \Delta_N(x) \left(\sum_{h=-m}^{m} (m+1-|h|) e^{2\pi ih(x-a)} \right) dx \\
&\qquad = \int_{-a}^{1-a} \Delta_N(x+a) \left(\sum_{h=-m}^{m} (m+1-|h|) e^{2\pi ihx} \right) dx,
\end{aligned} \tag{4.4.11}$$

where the asterisk indicates that $h = 0$ is deleted from the range of summation. Because of the periodicity of the integrand, the last integral may also be taken over $[-1/2, 1/2]$. We note that

$$\sum_{h=-m}^{m} (m+1-|h|) e^{2\pi ihx} = \frac{\sin^2(m+1)\pi x}{\sin^2 \pi x}, \tag{4.4.12}$$

where the right-hand side is interpreted as $(m+1)^2$ in case x is an integer. This is

the well-known discrete version of Fejér kernel. We infer from (4.4.11) that

$$\begin{aligned}\left| \int_{-1/2}^{1/2} \Delta_N(x+a) \frac{\sin^2(m+1)\pi x}{\sin^2 \pi x} dx \right| &\leq \frac{1}{2\pi} \sum_{h=-m}^{m}{}^{*} (m+1-|h|) \frac{|S_h|}{|h|} \\ &= \frac{1}{\pi} \sum_{h=1}^{m} (m+1-h) \frac{|S_h|}{h}.\end{aligned} \tag{4.4.13}$$

The last equality follows from the fact that $|S_h| = |S_{-h}|$. We note from the nature of the graph of Δ_N (Figure 4.1) that we either have $\Delta_N(b) = -D_N^*$ or $\Delta_N(b+0) = \lim_{x \to b^+} \Delta_N(x) = D_N^*$ for some $b \in [0,1]$. We shall deal with the second case, the first case being completely similar.

For $b < t \leq b + D_N^*$, we have

$$\Delta_N(t) = D_N^* + \Delta_N(t) - \Delta_N(b+0) \geq D_N^* + b - t. \tag{4.4.14}$$

In fact, since the slope of each linear function in the function Δ_N equals -1, it follows from the the nature of the graph of Δ_N (Figure 4.2) that

$$\frac{\Delta_N(t) - \Delta_N(b+0)}{t-b} \geq -1.$$

Now choose $a = b + \frac{1}{2} D_N^*$. Then for $|x| < \frac{1}{2} D_N^*$, we have $b < x + a < b + D_N^*$. So $x + a$ plays the role of t in (4.4.14) and therefore,

$$\Delta_N(x+a) \geq D_N^* + b - x - a = \frac{1}{2} D_N^* - x \quad \text{for } |x| < \frac{1}{2} D_N^*. \tag{4.4.15}$$

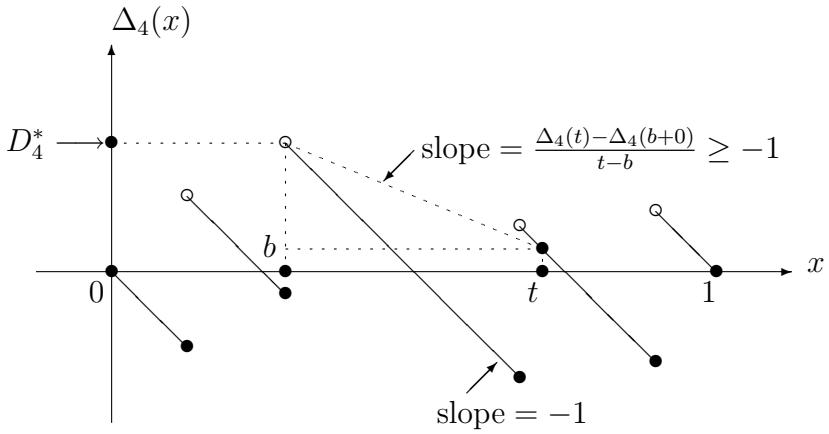

Figure 4.2: Pictorial proof of (4.4.14)

Consequently, we have

$$\begin{aligned}
&\int_{-1/2}^{1/2} \Delta_N(x+a)\frac{\sin^2(m+1)\pi x}{\sin^2 \pi x}\,dx \\
&\qquad = \Big(\int_{-1/2}^{-D_N^*/2} + \int_{-D_N^*/2}^{D_N^*/2} + \int_{D_N^*/2}^{1/2} \Big) \Delta_N(x+a)\frac{\sin^2(m+1)\pi x}{\sin^2 \pi x} dx \\
&\qquad \geq \int_{-D_N^*/2}^{D_N^*/2} (\frac{D_N^*}{2} - x)\frac{\sin^2(m+1)\pi x}{\sin^2 \pi x} dx - D_N^* \int_{-1/2}^{-D_N^*/2} \frac{\sin^2(m+1)\pi x}{\sin^2 \pi x} dx \\
&\qquad\quad - D_N^* \int_{D_N^*}^{1/2} \frac{\sin^2(m+1)\pi x}{\sin^2 \pi x} dx \\
&\qquad = D_N^* \int_0^{D_N^*/2} \frac{\sin^2(m+1)\pi x}{\sin^2 \pi x} dx - 2D_N^* \int_{D_N^*/2}^{1/2} \frac{\sin^2(m+1)\pi x}{\sin^2 \pi x} dx \\
&\qquad = D_N^* \int_0^{1/2} \frac{\sin^2(m+1)\pi x}{\sin^2 \pi x} dx - 3D_N^* \int_{D_N^*/2}^{1/2} \frac{\sin^2(m+1)\pi x}{\sin^2 \pi x} dx \\
&\qquad \geq \frac{m+1}{2} D_N^* - 3D_N^* \int_{D_N^*/2}^{1/2} \frac{dx}{4x^2} > \frac{m+1}{2} D_N^* - \frac{3}{2}.
\end{aligned} \tag{4.4.16}$$

The second inequality follows from (4.4.15) and the fact that $|\Delta_N(x)| \leq D_N^*$ for all x. The third equality follows from the definition of even function. We note that

(4.4.12) yields

$$\int_0^{1/2} \frac{\sin^2(m+1)\pi x}{\sin^2 \pi x} dx = \frac{m+1}{2}. \tag{4.4.17}$$

By convexity of the graph of sine function, we obtain the following inequality:

$$\sin x \geq \frac{2}{\pi} x \quad \text{for } x \in [0, \pi/2]. \tag{4.4.18}$$

We see that (4.4.17) and (4.4.18) together yield the penultimate inequality in (4.4.16). Combining (4.4.16) and (4.4.13), we arrive at

$$D_N^* \leq \frac{3}{m+1} + \frac{2}{\pi} \sum_{h=1}^{m} (\frac{1}{h} - \frac{1}{m+1}) |S_h|.$$

Since $D_N \leq 2D_N^*$ (Theorem 4.4.6), it follows that

$$D_N \leq \frac{6}{m+1} + \frac{4}{\pi} \sum_{h=1}^{m} (\frac{1}{h} - \frac{1}{m+1}) |S_h|. \tag{4.4.19}$$

We shall show that for any finite sequence $x_1, \ldots, x_N$ in $[0,1)$, there exists $c \in [0,1)$ such that the shifted sequence $\{x_1 + c\}, \ldots, \{x_N + c\}$ satisfies (4.4.9). This will prove the theorem, since both the left-hand and the right-hand side of (4.4.19) are invariant under the transition from $x_1, \ldots, x_N$ to the shifted sequence. Now since

$$\begin{aligned}
\int_0^1 \Delta_N(x)\, dx &= \frac{1}{N} \int_0^1 \Big(\sum_{n=1}^{N} \mathbb{1}_{(x_n,1]}(x) - Nx \Big)\, dx \\
&= \frac{1}{N} \sum_{n=1}^{N} \int_{x_n}^1 1\, dx - \int_0^1 x\, dx \\
&= \frac{1}{N} \sum_{n=1}^{N} (1 - x_n) - \frac{1}{2} \\
&= \frac{1}{2} - \frac{1}{N} \sum_{n=1}^{N} x_n,
\end{aligned}$$

we have to prove the existence of a number $c \in I$ for which

$$\frac{1}{N}\sum_{n=1}^{N}\{x_n + c\} = \frac{1}{2}.$$

But for every $c \in I$, we have

$$\begin{aligned}
\frac{1}{N}\sum_{n=1}^{N}(\{x_n + c\} - x_n) &= \frac{1}{N}\sum_{x_n<1-c} c + \frac{1}{N}\sum_{x_n\geq 1-c}(c-1) \\
&= \frac{c}{N}\Big(\sum_{x_n<1-c} 1 + \sum_{x_n\geq 1-c} 1\Big) - \frac{1}{N}\sum_{x_n\geq 1-c} 1 \\
&= c - \frac{1}{N}\Big(N - \sum_{x_n<1-c} 1\Big) \\
&= c - 1 + \frac{1}{N}\sum_{x_n<1-c} 1 = \Delta_N(1-c).
\end{aligned}$$

Therefore it remains to show that there exists $c \in I$ such that

$$\Delta_N(1-c) = \frac{1}{2} - \frac{1}{N}\sum_{n=1}^{N} x_n =: s.$$

We consider only the case $s > 0$, the case $s < 0$ being completely analogous. Since $\int_0^1 \Delta_N(t)\,dt = s$, we have $\Delta_N(x) \geq s$ for some $x \in (0,1)$. But since $\Delta_N(1) = 0$ and since Δ_N is piecewise linear with positive jumps only, the function Δ_N must attain the value s in the interval $[x, 1)$. □

We end this chapter by stating Koksma's Inequality which gives the upper bound in terms of discrepancy for the error of the integral from the average value of function over finitely many points in I.

Theorem 4.4.15 (Koksma's Inequality). *Let f be a function on $\overline{I}$ of bounded variation $Var(f)$, and suppose we are given N points $x_1, \ldots, x_N$ in I with discrepancy D_N^*. Then*

$$\left|\frac{1}{N}\sum_{n=1}^{N} f(x_n) - \int_0^1 f(t)\,dt\right| \leq \mathrm{Var}(f) D_N^*.$$

Chapter 5

Number Theoretic Approximation Theorem

5.1 Statement of the main theorem

In this chapter we prove the main theorem on improved quantization error estimate. The construction of proof has been modified from the original proof given in [16]. We begin with the setup of the problem. Let $\{F_N\}_{N=d+1}^{\infty}$ be a family of unit norm tight frames for $\mathbb{R}^d$, with $F_N = \{e_n^N\}_{n=1}^N$, so that F_N has frame bound N/d. If $x \in \mathbb{R}^d$, then $\{x_n^N\}_{n=1}^N$ will denote the corresponding sequence of frame coefficients with respect to F_N, i.e., $x_n^N = \langle x, e_n^N \rangle$. Let $\{q_n^N\}_{n=1}^N$ be the quantized sequence which is obtained by running the $\Sigma\Delta$ scheme of Definition 3.3.2 on the input sequence $\{x_n^N\}_{n=1}^N$ with respect to the identity permutation of $\{1, 2, \ldots, N\}$, and let $\{u_n^N\}_{n=0}^N$ be the associated auxiliary sequence of state variables. Thus, if $x \in \mathbb{R}^d$ is expressed as a frame expansion with respect to F_N, and if this expansion is quantized by the first order $\Sigma\Delta$ scheme, then the resulting quantized expansion is

$$\widetilde{x}_N = \frac{d}{N} \sum_{n=1}^{N} q_n^N e_n^N.$$

With $u_0^N = 0$, Abel Summation by Parts yields

$$\begin{aligned} x - \widetilde{x}_N &= \frac{d}{N}\Big(\sum_{n=1}^{N-1} u_n^N (e_n^N - e_{n+1}^N) + u_N^N e_N^N \Big) \\ &= \frac{d}{N}\Big(\sum_{n=1}^{N-2} v_n^N (f_n^N - f_{n+1}^N) + v_{N-1}^N f_{N-1}^N + u_N^N e_N^N \Big), \end{aligned} \tag{5.1.1}$$

where we have defined

$$f_n^N = e_n^N - e_{n+1}^N,\ v_n^N = \sum_{j=1}^{n} u_j^N,\ \text{ and } v_0^N = 0.$$

Our goal is to achieve a good estimate for $|v_n|$ for all $n = 1, \ldots, N-1$ and N is sufficiently large.

We let $\mathcal{B}_\Omega$ be the class of Ω-bandlimited functions consisting of all functions in $L^\infty(\mathbb{R})$ whose Fourier transforms (as distributions) are supported in $[-\Omega, \Omega]$. By Paley-Wiener theorem, elements of $\mathcal{B}_\Omega$ are restrictions of entire functions to the real line. A function f is said to be in class $\mathcal{M}_\Omega$ if $f \in \mathcal{B}_\Omega$, $f' \in L^\infty(\mathbb{R})$ and all the zeros $z_1, \ldots, z_{n^*}$ of f' contained in $[0,1]$ are simple, that is, $f''(z_j) \neq 0$ for all $j = 1, ..., n^*$. We use the notation $A \lesssim B$ to mean that there exists an absolute constant $C > 0$ such that $A \leq CB$. The following is the main theorem leading to the improvement of quantization error estimate of $\Sigma\Delta$ quantization.

Theorem 5.1.1 (Number Theoretic Theorem on Sigma-Delta Quantization). *Let $\{F_N\}_{N=d+1}^\infty$ be a family of unit norm tight frames for $\mathbb{R}^d$, with $F_N = \{e_n^N\}_{n=1}^N$. Suppose $x \in \mathbb{R}^d$ satisfies $\|x\| \leq (K - 1/2)\delta$, and let $\{x_n\}_{n=1}^N$ be the sequence of frame coefficients of x with respect to F_N. If, for some $\Omega > 0$, there exists $h \in \mathcal{M}_\Omega$ such that*

$$\forall N \text{ and } 1 \leq n \leq N,\ x_n^N = h(n/N),$$

then for all sufficiently large integers N and for all $n = 1, \ldots, N-1$,

$$|v_n^N| \lesssim \delta\Big(\frac{n}{N^{1/4}} + N^{3/4}\log N\Big) \lesssim \delta N^{3/4}\log N.$$

Moreover, if h' has no zero inside $[0,1]$, then for all sufficiently large integers N and

for all $n = 1, \ldots, N-1$,

$$|v_n^N| \lesssim \delta N^{2/3} + \sqrt{\delta} N^{2/3} + N^{1/3} \lesssim (\delta + \sqrt{\delta}) N^{2/3}.$$

The implicit constants are independent of N *and* δ*, but they do depend on* x *and hence* h*. The value of what constitutes a sufficiently large* N *depends on* δ*.*

5.2 Güntürk's theorem

In what follows we list two additional results which are not discussed in the previous chapters and are necessary for the proof of Theorem 5.1.1. We shall prove a special case of Güntürk's theorem in detail and shall derive an explicit bound associated to the inequality. The proof of the general case of Güntürk's theorem can be found in [8].

Theorem 5.2.1 (Bernstein's Inequality). *Let* $f \in \mathcal{B}_\Omega$*. Then* $\|f^{(r)}\|_{L^\infty} \leq \Omega^r \|f\|_{L^\infty}$.

Theorem 5.2.2 (Güntürk). *Let* $\Omega > 0$*,* $\lambda_0 > 1$ *and let* $h \in \mathcal{B}_\Omega$*, then for each* $0 < T \leq \frac{1}{2\lambda_0 \Omega}$ *there exists an analytic function* X_T *satisfying the following conditions.*

- $$\forall t \in \mathbb{R}, \quad X_T(t) - X_T(t-1) = h(tT) \text{ and } X_T(0) = 0, \tag{5.2.1}$$

- $$\|X_T' - h(\cdot T)\|_{L^\infty} \leq K_{\lambda_0,\Omega} T \|h\|_{L^\infty}, \tag{5.2.2}$$

  where $K_{\lambda_0,\Omega}$ *is a constant depending on* Ω *and* λ_0*.*

If it is given that $\widehat{h}$ is real-valued and locally integrable, then the constant $K_{\lambda_0,\Omega}$ can be taken as

$$K_{\lambda_0,\Omega} = 3\sqrt{\pi^2+4}\,\lambda_0\Omega\left(\frac{1}{2}+\frac{\lambda_0}{\pi(\lambda_0-1)}\right)^{2/3}.$$

Proof. As mentioned earlier, we shall prove the special case of the function h, that is, we shall assume that $h \in \mathcal{B}_\Omega$ for some $\Omega < \pi$, and $\widehat{h}$ is real-valued and locally integrable. Let X_T be a function in $L^\infty(\mathbb{R})$ such that its Fourier transform is given by

$$\widehat{X}_T(\gamma) = \frac{\widehat{h}(\gamma/T)}{T(1-e^{-2\pi i\gamma})} + c\delta_0(\gamma),$$

where $c = c(T)$ is chosen such that $X_T(0) = 0$. It follows that $\widehat{X}_T$ is a compactly supported distribution on $\mathbb{R}$. To see this, it suffices to show that for every function $g \in C_c^\infty(\mathbb{R})$ the following condition holds

$$\lim_{\epsilon\to 0^+}\int_{\epsilon\le|\gamma|\le\Omega}\frac{\widehat{h}(\gamma/T)}{1-e^{-2\pi i\gamma}}g(\gamma)\,d\gamma < \infty \quad \text{for } T>0.$$

We may assume without loss of generality that $T = 1$. We have, for each $0 < \epsilon < \Omega$,

$$\int_{\epsilon\le|\gamma|\le\Omega}\frac{\widehat{h}(\gamma)}{1-e^{-2\pi i\gamma}}g(\gamma)\,d\gamma = \int_{\epsilon\le|\gamma|\le\Omega}\gamma\frac{\widehat{h}(\gamma)}{1-e^{-2\pi i\gamma}}\cdot\frac{g(\gamma)-g(0)}{\gamma}+\frac{g(0)\widehat{h}(\gamma)}{1-e^{-2\pi i\gamma}}\,d\gamma.$$

The first term in the integral is integrable since

(i) $\dfrac{g(\gamma)-g(0)}{\gamma} \to g'(0)$ as $|\gamma|\to 0$,

(ii) $\dfrac{\gamma}{1-e^{-2\pi i\gamma}}$ is bounded away from zero as $|\gamma|\to 0$, and

(iii) $\widehat{h}$ is locally integrable.

We investigate the second integral carefully. We have, for each $0 < \epsilon < \Omega$,

$$\begin{aligned}
\int_{\epsilon \le |\gamma| \le \Omega} \frac{\widehat{h}(\gamma)}{1 - e^{-2\pi i \gamma}}\, d\gamma &= \frac{1}{2i} \int_{-\Omega}^{-\epsilon} \frac{\widehat{h}(\gamma) e^{\pi i \gamma}}{\sin \pi\gamma}\, d\gamma + \frac{1}{2i} \int_{\epsilon}^{\Omega} \frac{\widehat{h}(\gamma) e^{\pi i \gamma}}{\sin \pi\gamma}\, d\gamma \\
&= -\frac{1}{2i} \int_{\epsilon}^{\Omega} \frac{\widehat{h}(-\gamma) e^{-\pi i \gamma}}{\sin \pi\gamma}\, d\gamma + \frac{1}{2i} \int_{\epsilon}^{\Omega} \frac{\widehat{h}(\gamma) e^{\pi i \gamma}}{\sin \pi\gamma}\, d\gamma \\
&= \frac{1}{2i} \int_{\epsilon}^{\Omega} \frac{\widehat{h}(\gamma)\big(e^{\pi i \gamma} - e^{-\pi i \gamma}\big)}{\sin \pi\gamma}\, d\gamma \\
&= \int_{\epsilon}^{\Omega} \widehat{h}(\gamma)\, d\gamma < \infty \text{ as } \epsilon \to 0^{+} \quad \big(\because \widehat{h} \in L^1_{\mathrm{loc}}(\mathbb{R})\big).
\end{aligned}$$

The third equality follows from the assumption that $\widehat{h}$ is real-valued, so that $\widehat{h}(-\gamma) = \overline{\widehat{h}(\gamma)} = \widehat{h}(\gamma)$. We therefore verify that $\widehat{X}_T$ is a compactly supported distribution on $\mathbb{R}$. Hence, by the Paley-Wiener Theorem, X_T is a real analytic function on $\mathbb{R}$. Now since

$$\begin{aligned}
(X_T(t) - X_T(t-1))\widehat{\ }(\gamma) &= \widehat{X}_T(\gamma) - e^{-2\pi i \gamma} \widehat{X}_T(\gamma) \\
&= (1 - e^{-2\pi i \gamma}) \widehat{X}_T(\gamma) = \frac{1}{T} \widehat{h}(\gamma/T) + c\delta_0(\gamma)(1 - e^{-2\pi i \gamma}) \\
&= \big(h(tT)\big)\widehat{\ }(\gamma) + c(\mathbb{1}(t) - \mathbb{1}(t-1))\widehat{\ }(\gamma) = \big(h(tT)\big)\widehat{\ }(\gamma),
\end{aligned}$$

it follows that (5.2.1) holds. Let φ be a fixed smoothing kernel defined by

$$\widehat{\varphi}(\gamma) = \begin{cases} 1 & \text{if} \quad |\gamma| \le \Omega, \\ 0 & \text{if} \quad |\gamma| \ge \lambda_0 \Omega. \end{cases}$$

Let ϕ_T be such that

$$\widehat{\phi}_T(\gamma) = \Big(\frac{2\pi i \gamma}{1 - e^{-2\pi i \gamma}} - 1\Big) \widehat{\varphi}(\gamma/T).$$

Since, by computation of Fourier transform,

$$\begin{aligned}
\big(X_T' - h(\cdot\, T)\big)\widehat{\ }(\gamma) &= 2\pi i\gamma \widehat{X}_T(\gamma) - \frac{1}{T}\widehat{h}(\gamma/T) \\
&= 2\pi i\gamma\Big(\frac{\widehat{h}(\gamma/T)}{T(1-e^{-2\pi i\gamma})} + c\delta_0(\gamma)\Big) - \frac{1}{T}\widehat{h}(\gamma/T) \\
&= \frac{2\pi i\gamma \widehat{h}(\gamma/T)}{T(1-e^{-2\pi i\gamma})} + 2\pi i c \underbrace{\gamma\delta_0(\gamma)}_{=0} - \frac{1}{T}\widehat{h}(\gamma/T) \\
&= \Big(\frac{2\pi i\gamma}{1-e^{-2\pi i\gamma}} - 1\Big)\frac{1}{T}\widehat{h}(\gamma/T) \\
&= \big(\phi_T * h(\cdot\, T)\big)\widehat{\ }(\gamma),
\end{aligned}$$

it follows that

$$X_T' - h(\cdot\, T) = \phi_T * h(\cdot\, T). \tag{5.2.3}$$

Hence we obtain the following inequality

$$\|X_T' - h(\cdot\, T)\|_{L^\infty} \le \|\phi_T\|_{L^1}\|h\|_{L^\infty}. \tag{5.2.4}$$

Next we establish that $\|\phi\|_{L^1} \le K_{\lambda_0,\Omega} T$ for some constant $K_{\lambda_0,\Omega} > 0$. We first note that

$$\frac{2\pi i\gamma}{1-e^{-2\pi i\gamma}} = 1 + i\pi\gamma + \mathcal{O}(\gamma^2).$$

Hence, by convexity, for $|\gamma| \le 1/2$,

$$\Big|\frac{2\pi i\gamma}{1-e^{-2\pi i\gamma}} - 1\Big| \le \sqrt{\pi^2+4}\,|\gamma| \quad \text{and} \quad \Big|\frac{d}{d\gamma}\Big(\frac{2\pi i\gamma}{1-e^{-2\pi i\gamma}}\Big)\Big| \le \frac{\pi}{2}\sqrt{\pi^2+4}.$$

Now since $|\gamma| \le \lambda_0\Omega T \le 1/2$ in the support of $\widehat{\phi}_T$, we obtain

$$|\widehat{\phi}_T(\gamma)| \le \Big|\frac{2\pi i\gamma}{1-e^{-2\pi i\gamma}} - 1\Big||\widehat{\varphi}(\gamma/T)| \le \sqrt{\pi^2+4}\,|\gamma| \le \sqrt{\pi^2+4}\,\lambda_0\Omega T$$

and

$$\begin{aligned}\Big|\frac{d\widehat{\phi}_T}{d\gamma}\Big| &\le \Big|\frac{d}{d\gamma}\Big(\frac{2\pi i\gamma}{1-e^{-2\pi i\gamma}}\Big)\Big||\widehat{\varphi}(T\gamma)| + \Big|\frac{2\pi i\gamma}{1-e^{-2\pi i\gamma}} - 1\Big|\Big|\frac{1}{T}(\widehat{\varphi})'(\gamma/T)\Big| \\ &\le \frac{\pi}{2}\sqrt{\pi^2+4} + \sqrt{\pi^2+4}\,|\gamma|\frac{1}{T}\frac{1}{\Omega(\lambda_0-1)} \\ &\le \frac{\pi}{2}\sqrt{\pi^2+4} + \sqrt{\pi^2+4}\,\lambda_0\Omega T\frac{1}{T}\frac{1}{\Omega(\lambda_0-1)} \\ &= \sqrt{\pi^2+4}\Big(\frac{\pi}{2} + \frac{\lambda_0}{\lambda_0-1}\Big).\end{aligned}$$

This implies that

$$\|\phi_T\|_{L^\infty} \le \int |\widehat{\phi}_T(\gamma)|\,d\gamma \le \int_{|\gamma|\le\lambda_0\Omega T} \sqrt{\pi^2+4}\,\lambda_0\Omega T\,d\gamma \le 2\sqrt{\pi^2+4}\,\lambda_0^2\Omega^2T^2$$

and

$$\begin{aligned}\Big\|\frac{d\widehat{\phi}_T}{d\gamma}\Big\|_{L^2} &\le \Big(\int_{|\gamma|\le\lambda_0\Omega T}(\pi^2+4)\Big(\frac{\pi}{2}+\frac{\lambda_0}{\lambda_0-1}\Big)^2 d\gamma\Big)^{1/2} \\ &\le \sqrt{2}\sqrt{\pi^2+4}\Big(\frac{\pi}{2}+\frac{\lambda_0}{\lambda_0-1}\Big)\sqrt{\lambda_0}\sqrt{\Omega}\sqrt{T}.\end{aligned}$$

Combining these two estimates, we obtain, for any $A > 0$,

$$\begin{aligned}\|\phi_T\|_{L^1} &\le \int_{|t|\le A}\|\phi_T\|_{L^\infty}\,dt + \int_{|t|>A}\frac{1}{|t|}|t\phi_T(t)|\,dt \\ &\le 4A\sqrt{\pi^2+4}\,\lambda_0^2\Omega^2T^2 + \frac{1}{2\pi}\Big(\int_{|t|>A}\frac{1}{t^2}\,dt\Big)^{1/2}\Big(\int|-2\pi it\phi_T(t)|^2\,dt\Big)^{1/2} \\ &= 4A\sqrt{\pi^2+4}\,\lambda_0^2\Omega^2T^2 + \frac{1}{2\pi}\frac{\sqrt{2}}{\sqrt{A}}\Big\|\frac{d\widehat{\phi}_T}{d\gamma}\Big\|_{L^2} \\ &\le 4A\sqrt{\pi^2+4}\,\lambda_0^2\Omega^2T^2 + \frac{1}{\pi}\frac{1}{\sqrt{A}}\sqrt{\pi^2+4}\Big(\frac{\pi}{2}+\frac{\lambda_0}{\lambda_0-1}\Big)\sqrt{\lambda_0}\sqrt{\Omega}\sqrt{T}.\end{aligned}$$

The second inequality follows from Hölder's inequality and the third equality from Parseval's formula and the fact that $(-2\pi itf(t))\widehat{\ }(\gamma) = (\widehat{f})'(\gamma)$. To minimize the right-hand side of the last inequality, we choose $A = (w/2v)^{2/3}$ where $v = 4\lambda_0^2\Omega^2T^2$ and $w = \frac{1}{\pi}\big(\frac{\pi}{2}+\frac{\lambda_0}{\lambda_0-1}\big)\sqrt{\lambda_0}\sqrt{\Omega}\sqrt{T}$. We obtain

$$\|\phi_T\|_{L^1} \le \frac{3}{\sqrt[3]{4}}\sqrt{\pi^2+4}\,w^{2/3}v^{1/3} = 3\sqrt{\pi^2+4}\,\lambda_0\Omega\Big(\frac{1}{2}+\frac{\lambda_0}{\pi(\lambda_0-1)}\Big)^{2/3}T$$

and hence the proof is complete. □

Remark 5.2.3. With the setting of hypothesis of Theorem 5.1.1, Theorem 5.2.2 states the following: For each $N > d$ there exists an analytic function X_N such that

$$\forall\, t \in \mathbb{R}, \quad X_N(t) - X_N(t-1) = h\Big(\frac{t}{N}\Big) \text{ and } X_N(0) = u_0^N = 0. \tag{5.2.5}$$

Moreover, we have

$$\Big|X_N'(t) - h\Big(\frac{t}{N}\Big)\Big| \lesssim \frac{1}{N}. \tag{5.2.6}$$

We note that condition (5.2.5) implies

$$\forall\, 0 \le n \le N, \exists k_n \in \mathbb{Z}, k_n \equiv n \pmod 2, \quad X_N(n) = u_n^N + k_n \frac{\delta}{2}. \tag{5.2.7}$$

To see this, we proceed by induction: For $n = 0$ we have $X_N(0) = u_0^N = 0 = 0 \cdot \frac{\delta}{2}$. So the statement is true for $n = 0$.

For $n = 1$,

$$X_N(1) = X_N(1) - X_N(0) = h\Big(\frac{1}{N}\Big) = x_1^N = u_1^N - u_0^N + q_1^N = u_1^N + q_1^N.$$

Since q_1^N is one of the quantization alphabet, it follows that q_1^N is an odd multiple of $\delta/2$.

Assume the result for $n-1$, then from (5.2.5)

$$\begin{aligned} X_N(n) &= X_N(n-1) + h\Big(\frac{n}{N}\Big) \\ &= X_N(n-1) + x_n^N \\ &= X_N(n-1) + u_n^N - u_{n-1}^N + q_n^N. \end{aligned}$$

So

$$X_N(n) - u_n^N = \Big(X_N(n-1) - u_{n-1}^N\Big) + q_n^N.$$

By induction hypothesis, $X_N(n-1) - u_{n-1}^N$ is either an odd or even multiple of $\delta/2$. If it is an odd (even) multiple of $\delta/2$, then since q_n^N is an odd multiple of $\delta/2$, we have that $X_N(n) - u_n^N$ is an even (odd) multiple of $\delta/2$.

5.3 Proof of the main theorem

We are now ready to present the proof of Theorem 5.1.1.

Proof of Theorem 5.1.1. Let u_n^N be the state variable of the $\Sigma\Delta$ scheme and define $\widetilde{u}_n^N = u_n^N/\delta$. From the definition of v_n^N, and Koksma's Inequality, we obtain

$$\begin{aligned} |v_j^N| &= \delta\Big|\sum_{n=1}^{j}\widetilde{u}_n^N\Big| = j\delta\Big|\frac{1}{j}\sum_{n=1}^{j}\widetilde{u}_n^N - \int_{-1/2}^{1/2} y\,dy\Big| \\ &\leq j\delta \mathrm{Var}(x) D_j(\widetilde{u}_1^N, \ldots, \widetilde{u}_j^N), \end{aligned}$$

where $D_j(\cdot)$ denotes the discrepancy of a sequence as defined by (4.4.1). Next we estimate the discrepancy using Erdös-Turán's inequality: There exists a constant $C > 0$ such that for all integers $K \geq 1$,

$$D_j(\widetilde{u}_1^N, \ldots, \widetilde{u}_j^N) \leq C\Big(\frac{1}{K} + \frac{1}{j}\sum_{k=1}^{K}\frac{1}{k}\Big|\sum_{n=1}^{j} e^{2\pi i k \widetilde{u}_n^N}\Big|\Big).$$

Finally it suffices to estimate $|\sum_{n=1}^{j} e^{2\pi i k \widetilde{u}_n^N}|$.

Applying Bernstein's inequality to (5.2.6) yields

$$\Big|X_N''(t) - \frac{1}{N}h'\Big(\frac{t}{N}\Big)\Big| \lesssim \frac{1}{N^2}. \tag{5.3.1}$$

Let $z_1, \ldots, z_{n^*}$ be the zeros of h' contained in [0,1] in increasing order. Let $0 < \alpha < 1$ be fixed. With an integer N sufficiently large, we define sequence of intervals as

follows:

$$\forall\, j = 1, \ldots, n^*, \quad I_j = [Nz_j - N^\alpha,\, Nz_j + N^\alpha],$$

$$\forall\, j = 1, \ldots, n^* - 1, \quad J_j^{(1)} = [Nz_j + N^\alpha,\, \frac{N}{2}(z_j + z_{j+1})]$$

and

$$J_j^{(2)} = [\frac{N}{2}(z_j + z_{j+1}),\, Nz_{j+1} - N^\alpha],$$

$$J_0 = [1,\, Nz_1 - N^\alpha] \text{ and } J_{n^*} = [Nz_{n^*} + N^\alpha,\, N].$$

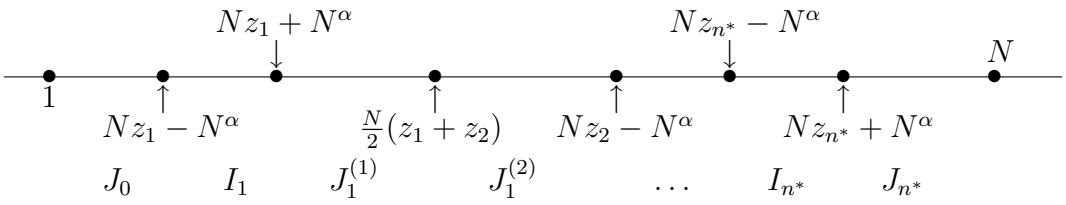

Figure 5.1: The partition of interval $[1, N]$

If 0 or 1 is a zero of h', then we adjust the intervals as follows: if $z_1 = 0$, then discard J_0 and adjust I_1 by $I_1 = [1,\, N^\alpha]$, if $z_{n^*} = 1$, then discard J_{n^*} and adjust I_{n^*} by $I_{n^*} = [N - N^\alpha,\, N]$. If h' has no zero in $[0,\, 1]$ then we just consider the whole interval $[1,\, N]$. In the following proof, we treat only the case when $z_1 \neq 0$ and $z_{n^*} \neq 1$. The proof can be slightly modified when handling with these two cases. We see that

$$J_0 \cup I_1 \cup J_1^{(1)} \cup J_1^{(2)} \cup I_2 \cup \cdots \cup I_{n^*} \cup J_{n^*} = [1,\, N].$$

Claim. *There exists a constant $C > 0$ such that for all sufficiently large integer N and for all $j = 1, \ldots, n^* - 1, k = 1, 2$ and for all $t \in J_0$ or $t \in J_{n^*}$ or $t \in J_j^{(k)}$,*

$$\left|h'\Big(\frac{t}{N}\Big)\right| \geq C\frac{1}{N^{1-\alpha}}.$$

Proof of Claim. Fix $1 \leq j \leq n^* - 1$, and let $t \in J_j^{(1)}$. Then

$$z_j + \frac{1}{N^{1-\alpha}} \leq \frac{t}{N} \leq w_j,$$

where $w_j = \frac{1}{2}(z_j + z_{j+1})$.

Since z_j and z_{j+1} are two consecutive zeros of h' in $[0, 1]$ and $z_j < t/N < z_{j+1}$, we have two cases.

case 1 $h'(t/N) < 0$. Applying Mean Value Theorem on $[z_j, t/N]$, we have that there exists $c_{t,j} \in (z_j, t/N)$ such that

$$h'\Big(\frac{t}{N}\Big) - h'(z_j) = h'\Big(\frac{t}{N}\Big) = h''(c_{t,j})\Big(\frac{t}{N} - z_j\Big).$$

By the first inequality above, we have

$$\begin{aligned} h'\Big(\frac{t}{N}\Big) &\leq h''(c_{t,j})\frac{1}{N^{1-\alpha}} \\ &\leq \sup_{\xi\in(z_j,w_j]} \frac{h'(\xi) - h'(z_j)}{\xi - z_j} \cdot \frac{1}{N^{1-\alpha}} \\ &= \sup_{\xi\in(z_j,w_j]} \frac{h'(\xi)}{\xi - z_j} \cdot \frac{1}{N^{1-\alpha}}. \end{aligned}$$

Since t was arbitrary in $J_j^{(1)}$ and since z_j and z_{j+1} are consecutive zeros of h', we have $h'(t/N) < 0$ for all $t \in J_j^{(1)}$. Hence for all $t \in J_j^{(1)}$

$$-h'\Big(\frac{t}{N}\Big) \geq \inf_{\xi\in(z_j,w_j]} \frac{-h'(\xi)}{\xi - z_j} \cdot \frac{1}{N^{1-\alpha}}.$$

case 2 $h'(t/N) > 0$. By the same argument as above, we have that for all $t \in J_j^{(1)}$

$$h'\Big(\frac{t}{N}\Big) \geq \inf_{\xi \in (z_j, w_j]} \frac{h'(\xi)}{\xi - z_j} \cdot \frac{1}{N^{1-\alpha}}.$$

Combining case 1 and 2 together, we have that for all $t \in J_j^{(1)}$,

$$\Big|h'\Big(\frac{t}{N}\Big)\Big| \geq \inf_{\xi \in (z_j, w_j]} \frac{|h'(\xi)|}{\xi - z_j} \cdot \frac{1}{N^{1-\alpha}}.$$

Now since for each $\xi \in (z_j, w_j]$, the quantity $\frac{h'(\xi)}{\xi - z_j}$ is the slope of line joining $(\xi, h'(\xi))$ and $(z_j, h'(z_j))$ and since $h''(z_j) \neq 0$, we see that

$$C_j^{(1)} = \inf_{\xi \in (z_j, w_j]} \frac{|h'(\xi)|}{\xi - z_j} \cdot \frac{1}{N^{1-\alpha}} > 0.$$

Thus for each $j = 1, \ldots, n^* - 1$ and $t \in J_j^{(1)}$,

$$\Big|h'\Big(\frac{t}{N}\Big)\Big| \geq C_j^{(1)} \frac{1}{N^{1-\alpha}}.$$

Fix $1 \leq j \leq n^* - 1$ and let $t \in J_j^{(2)}$. Then

$$w_j \leq \frac{t}{N} \leq z_{j+1} - \frac{1}{N^{1-\alpha}},$$

where $w_j = \frac{1}{2}(z_j + z_{j+1})$. Using the same argument as in the first part of the proof above, we have that for each $j = 1, \ldots, n^* - 1$, there exists $C_j^{(2)} > 0$ such that for all $t \in J_j^{(2)}$,

$$\Big|h'\Big(\frac{t}{N}\Big)\Big| \geq C_j^{(2)} \frac{1}{N^{1-\alpha}}.$$

The proofs for the intervals J_0 and J_{n^*} are analogous and we assume the constants obtained from these two cases are C_0 and C_{n^*}, respectively. By letting

$$C = \min\{C_0, C_{n^*}, C_1^{(1)}, C_1^{(2)}, \ldots, C_{n^*-1}^{(1)}, C_{n^*-1}^{(2)}\},$$

we have the statement of the claim.

Let J denote one of the intervals $J_0, J_{n^*}, J_j^{(k)}, j = 1, \ldots, n^*-1, k = 1, 2$. From (5.3.1), we then have, for all $t \in J$,

$$\frac{1}{N}\Big|h'\Big(\frac{t}{N}\Big)\Big| - |X_N''(t)| \lesssim \frac{1}{N^2},$$

so that

$$\frac{1}{N^{2-\alpha}} \lesssim \frac{1}{N}\Big|h'\Big(\frac{t}{N}\Big)\Big| \lesssim \frac{1}{N^2} + |X_N''(t)|.$$

Hence, by continuity of X_N'',

$$\frac{1}{N^{2-\alpha}} \lesssim X_N''(t)\ \forall t \in J \quad \text{or} \quad X_N''(t) \lesssim -\frac{1}{N^{2-\alpha}}\ \forall t \in J. \tag{5.3.2}$$

Also, since $h \in \mathcal{B}_\Omega \subseteq L^\infty(\mathbb{R})$, and from (5.2.6), we obtain

$$\forall t \in J, \qquad |X_N'(t)| \lesssim 1. \tag{5.3.3}$$

Now we consider the bound of exponential sum. For each integer $k \geq 1$, and for each interval J, we have

$$\begin{aligned}
\Big|\sum_{n\in\mathbb{N}\cap J} e^{2\pi i k\tilde{u}_n^N}\Big| &= \Big|\sum_{n\in\mathbb{N}\cap J} e^{2\pi i\frac{k}{\delta}u_n^N}\Big| \\
&= \Big|\sum_{\substack{n\in\mathbb{N}\cap J\\ n \text{ even}}} e^{2\pi i\frac{k}{\delta}(X_N(n)+k_n\frac{\delta}{2})} + \sum_{\substack{n\in\mathbb{N}\cap J\\ n \text{ odd}}} e^{2\pi i\frac{k}{\delta}(X_N(n)+\ell_n\frac{\delta}{2})}\Big|,\ k_n \text{ even},\ \ell_n \text{ odd} \\
&= \Big|\sum_{\substack{n\in\mathbb{N}\cap J\\ n \text{ even}}} e^{2\pi i\frac{k}{\delta}X_N(n)} + (-1)^k\sum_{\substack{n\in\mathbb{N}\cap J\\ n \text{ odd}}} e^{2\pi i\frac{k}{\delta}X_N(n)}\Big| \\
&\leq \Big|\sum_{\substack{n\in\mathbb{N}\cap J\\ n \text{ even}}} e^{2\pi i\frac{k}{\delta}X_N(n)}\Big| + \Big|\sum_{\substack{n\in\mathbb{N}\cap J\\ n \text{ odd}}} e^{2\pi i\frac{k}{\delta}X_N(n)}\Big|.
\end{aligned}$$

We consider the bound for n even, and the case for n odd is analogous. We have

$$\left| \sum_{\substack{n\in\mathbb{N}\cap J \\ n \text{ even}}} e^{2\pi i \frac{k}{\delta} X_N(n)} \right| = \left| \sum_{\ell=a}^{b} e^{2\pi i \frac{k}{\delta} \widetilde{X}_N(\ell)} \right|,$$

where $\widetilde{X}_N(\ell) = X_N(2\ell)$ for $\ell = a, a+1, \ldots, b$. From inequalities (5.3.2) and (5.3.3), we have

$$\left|\widetilde{X}'_N(t)\right| = 2\left|X'_N(2t)\right| \lesssim 1,$$

and

$$\left|\widetilde{X}''_N(t)\right| = 4\left|X''_N(2t)\right| \gtrsim \frac{1}{N^{2-\alpha}}.$$

So we can apply Van der Corput Theorem (Theorem 4.3.1) to get

$$\begin{aligned}
\left| \sum_{\substack{n\in\mathbb{N}\cap J \\ n \text{ even}}} e^{2\pi i \frac{k}{\delta} X_N(n)} \right| &\lesssim \left| \frac{k}{\delta}\widetilde{X}'_N(b) - \frac{k}{\delta}\widetilde{X}'_N(a) + 2 \right| \left(4\sqrt{\frac{\delta}{k}} N^{1-\frac{\alpha}{2}} + 3 \right) \\
&\lesssim \sqrt{\frac{k}{\delta}} N^{1-\frac{\alpha}{2}} + \frac{k}{\delta} + \sqrt{\frac{\delta}{k}} N^{1-\frac{\alpha}{2}} + 1 \\
&\lesssim \sqrt{\frac{k}{\delta}} N^{1-\frac{\alpha}{2}} + \frac{k}{\delta}.
\end{aligned}$$

We get the same bound for n odd. Hence, for sufficiently large N, for each integer $k \geq 1$, and for each interval J.

$$\left| \sum_{n\in\mathbb{N}\cap J} e^{2\pi i k \tilde{u}_n^N} \right| \lesssim \sqrt{\frac{k}{\delta}} N^{1-\frac{\alpha}{2}} + \frac{k}{\delta}.$$

We use the trivial estimate on interval I where I is one of the intervals $I_j, j = 1, \ldots, n^*$, i.e.,

$$\left| \sum_{n\in\mathbb{N}\cap I} e^{2\pi i k \tilde{u}_n^N} \right| \lesssim 2N^{\alpha}.$$

We observe that if we take a subset of consecutive integers inside intervals I or J, the same bound of the exponential sum holds. Thus, for all $j = 1, \ldots, N-1$,

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

Let $K = N^{\beta}$, for some $0 < \beta < 1$. By the bound of discrepancy we have earlier, it follows that for all sufficiently large integers N,

$$\begin{aligned} D_j(\widetilde{u}_1^N, \ldots, \widetilde{u}_j^N) &\lesssim \frac{1}{K} + \frac{N^{\alpha} \log K}{j} + \frac{K^{1/2} N^{1-\alpha/2}}{\delta^{1/2} j} + \frac{K}{\delta j} \\ &\lesssim \frac{1}{N^{\beta}} + \frac{N^{\alpha} \log N}{j} + \frac{N^{\beta/2-\alpha/2+1}}{\delta^{1/2} j} + \frac{N^{\beta}}{\delta j}. \end{aligned}$$

Thus we obtain the bound of v_n^N:

$$\begin{aligned} |v_n^N| &\lesssim \frac{\delta n}{N^{\beta}} + \delta N^{\alpha} \log N + \delta^{1/2} N^{\beta/2-\alpha/2+1} + N^{\beta} \\ &\lesssim \delta N^{1-\beta} + \delta N^{\alpha} \log N + \delta^{1/2} N^{\beta/2-\alpha/2+1} + N^{\beta} \end{aligned}$$

To minimize the right-hand side quantity, we choose $\alpha = 3/4$ and $\beta = 1/4$. We then have, for each $\epsilon > 0$, there exists N_ϵ such that for all $N \geq N_\epsilon$,

$$|v_n^N| \lesssim \delta N^{3/4} + \delta N^{3/4} \log N + \delta^{1/2} N^{3/4} + N^{1/4} \lesssim \delta N^{3/4} \log N \lesssim \delta N^{3/4+\epsilon}.$$

If h' has no zero in $[0,1]$ then there exists a constant $C > 0$ such that $|h'(t)| \geq C$ for all $t \in [0,1]$. This allows us to let $\alpha = 1$ in the construction of the proof above. So we can approximate the exponential sum over interval $[1, N]$ directly and obtain that for each integer $k \geq 1$, for all sufficiently large integers N and for all $j = 1, \ldots, N-1$,

$$\Big| \sum_{n=1}^{j} e^{2\pi i k \tilde{u}_n^N} \Big| \lesssim \sqrt{\frac{k}{\delta}} N^{1/2} + \frac{k}{\delta}.$$

Let $K = N^\beta$. Then we have, for all sufficiently large integers N,

$$\begin{aligned} D_j(\widetilde{u}_1^N, \ldots, \widetilde{u}_j^N) &\lesssim \frac{1}{K} + \frac{K^{1/2}N^{1/2}}{\delta^{1/2}j} + \frac{K}{\delta j} \\ &\lesssim \frac{1}{N^\beta} + \frac{N^{\beta/2+1/2}}{\delta^{1/2}j} + \frac{N^\beta}{\delta j}. \end{aligned}$$

Thus we obtain the bound of v_n^N :

$$|v_n^N| \lesssim \frac{\delta n}{N^\beta} + \sqrt{\delta}N^{\beta/2+1/2} + N^\beta \lesssim \delta N^{1-\beta} + \sqrt{\delta}N^{\beta/2+1/2} + N^\beta$$

To minimize the right-hand side quantity, we choose $\beta = 1/3$. This proves the second part of the theorem. □

Corollary 5.3.1. *Let $\{F_N\}_{N=d}^\infty$ be a family of unit norm tight frames for $\mathbb{R}^d$, for which each $F_N = \{e_n^N\}_{n=1}^N$ satisfies the zero sum condition. Suppose $x \in \mathbb{R}^d$ satisfies $\|x\| \leq (K - 1/2)\delta$ for some positive integer K and $\delta > 0$ in the $\Sigma\Delta$ scheme. Let $\{x_n^N\}_{n=1}^N$ be the sequence of frame coefficients of x with respect to F_N, and suppose there exists $h \in \mathcal{M}_\Omega$, $\Omega > 0$, such that*

$$\forall N \text{ and } 1 \leq n \leq N, \quad x_n^N = h(n/N).$$

Additionally, suppose that $f_n^N = e_n^N - e_{n+1}^N$ satisfies

$$\forall N \text{ and } 1 \leq n \leq N, \quad \|f_n^N\| \lesssim \frac{1}{N} \quad \text{and} \quad \|f_n^N - f_{n+1}^N\| \lesssim \frac{1}{N^2},$$

and set $u_0^N = 0$ in the $\Sigma\Delta$ scheme.

If N is even and sufficiently large, then

$$\|x - \widetilde{x}_N\| \lesssim \frac{\delta \log N}{N^{5/4}}.$$

If N is odd and sufficiently large, then

$$\frac{\delta}{N} \lesssim \|x - \widetilde{x}_N\| \lesssim \frac{\delta d}{2N}(\sigma(F_N, p_N) + 1).$$

If h' has no zero in $[0,1]$, then for N even and sufficiently large,

$$\|x - \widetilde{x}_N\| \lesssim \frac{\delta}{N^{4/3}},$$

and for N odd and sufficiently large,

$$\frac{\delta}{N} \lesssim \|x - \widetilde{x}_N\| \lesssim \frac{\delta d}{2N}(\sigma(F_N, p_N) + 1).$$

The implicit constants are independent of δ and N, but do depend on x, and hence h.

Proof. From the first part of Theorem 5.1.1 we have

$$\begin{aligned}
&\left\| \frac{d}{N}\Big(\sum_{n=1}^{N-2} v_n^N (f_n^N - f_{n+1}^N) + v_{N-1}^N f_{N-1}^N \Big) \right\| \\
&\qquad \le \frac{d}{N}\Big(\sum_{n=1}^{N-2} \left|v_n^N\right| \left\|f_n^N - f_{n+1}^N\right\| + \left|v_{N-1}^N\right| \left\|f_{N-1}^N\right\| \Big) \\
&\qquad \lesssim \frac{d}{N}\big(\delta N^{3/4} \log N\big)\Big(\frac{N-2}{N^2} + \frac{1}{N}\Big) \\
&\qquad \lesssim \frac{\delta \log N}{N^{5/4}}.
\end{aligned}$$

Combining this with Theorem 3.3.6 and (5.1.1), we have that for N even and sufficiently large

$$\|x - \widetilde{x}_N\| \lesssim \frac{\delta \log N}{N^{5/4}},$$

and for N odd and sufficiently large

$$\begin{aligned}
\frac{\delta}{N} \lesssim \frac{d\delta}{2N} = \frac{d\left|u_N^N\right| \left\|e_N^N\right\|}{N} &\lesssim \|x - \widetilde{x}_N\| + \frac{\delta \log N}{N^{5/4}} \lesssim \|x - \widetilde{x}_N\| \\
&\lesssim \frac{\delta d}{2N}(\sigma(F_N, p_N) + 1).
\end{aligned}$$

The rightmost inequality follows from Theorem 3.3.7.

If h' has no zero in $[0,1]$, then by Theorem 5.1.1 we have $\left|v_n^N\right| \lesssim \delta N^{2/3}$. We proceed as above by replacing the bound of $\left|v_n^N\right|$ with $\delta N^{2/3}$ and obtain that for N even and sufficiently large

$$\|x - \widetilde{x}_N\| \lesssim \frac{\delta}{N^{4/3}},$$

and for N odd and sufficiently large

$$\frac{\delta}{N} \lesssim \|x - \widetilde{x}_N\| \lesssim \frac{\delta d}{2N}(\sigma(F_N, p_N) + 1). \qquad \square$$

5.4 Examples

In this section we give examples to verify the theorems we have proved in Chapter 3 and in this chapter.

Example 5.4.1 (Error estimates for H_N^d with d even). We shall assume throughout that d is even. We show first that a harmonic frame H_N^d satisfies the zero sum condition and has uniformly bounded frame variation with respect to identity permutation. We recall the definition (2.4.3) of harmonic frame $H_N^d = \{e_n\}_{n=0}^{N-1}$, $N > d$ for the case when d is even that

$$e_n = \sqrt{\frac{2}{d}}\left[\cos\frac{2\pi n}{N}, \sin\frac{2\pi n}{N}, \cos\frac{2\pi 2n}{N}, \sin\frac{2\pi 2n}{N}, \ldots, \cos\frac{2\pi\frac{d}{2}n}{N}, \sin\frac{2\pi\frac{d}{2}n}{N}\right]$$

for $n = 0, 1, \ldots, N-1$. The verification that H_N^d satisfies the zero sum condition follows by noting that, for each integer k not divisible by N,

$$\sum_{j=0}^{N-1} \cos\frac{2\pi kj}{N} = \Re\left[\sum_{j=0}^{N-1}(e^{2\pi ik/N})^j\right] = 0$$

and

$$\sum_{j=0}^{N-1} \sin \frac{2\pi kj}{N} = \Im \left[\sum_{j=0}^{N-1} (e^{2\pi ik/N})^j \right] = 0.$$

Now we show that the kth order frame variation $\sigma_k(H_N^d, p)$ (see Definition 2.4.3) of H_N^d with respect to identity permutation p is uniformly bounded. From the proof of Theorem 2.4.5, we have

$$\begin{aligned}
\sigma_k(H_N^d, p) &= \sum_{n=0}^{N-k-1} \left\| \Delta^k e_n \right\| \\
&= \sum_{n=0}^{N-k-1} \sqrt{\frac{2}{d}} \left(\sum_{j=1}^{d/2} \left(\Delta^k \cos n\theta_j \right)^2 + \left(\Delta^k \sin n\theta_j \right)^2 \right)^{1/2} \qquad (\text{ where } \theta_j = 2\pi j/N) \\
&= \sqrt{\frac{2}{d}} \sum_{n=0}^{N-k-1} \left(\sum_{j=1}^{d/2} \left(2 \sin \frac{\theta_j}{2} \right)^{2k} \right)^{1/2} \\
&\leq (N-k) \sqrt{\frac{2}{d}} \left(\sum_{j=1}^{d/2} \left(\frac{2\pi j}{N} \right)^{2k} \right)^{1/2} \\
&= (N-k) \sqrt{\frac{2}{d}} \left(\frac{2\pi}{N} \right)^k \left(\sum_{j=1}^{d/2} j^{2k} \right)^{1/2} \\
&\leq \sqrt{\frac{2}{d}} (2\pi)^k \left(\sum_{j=1}^{d/2} j^{2k} \right)^{1/2}.
\end{aligned}$$

Thus, for $k = 1$,

$$\begin{aligned}
\sigma_1(H_N^d, p) = \sigma(H_N^d, p) &\leq \sqrt{\frac{2}{d}} 2\pi \left(\frac{d}{2} \left(\frac{d}{2} + 1 \right) (d+1) \frac{1}{6} \right)^{1/2} \\
&= \frac{\pi}{\sqrt{3}} \sqrt{(d+2)(d+1)} \qquad (5.4.1) \\
&< \frac{\pi}{\sqrt{3}} (d+2).
\end{aligned}$$

We now derive error estimates for $\Sigma\Delta$ quantization of harmonic frames in their natural order. If we set $u_0 = 0$ and assume that $x \in \mathbb{R}^d$ satisfies $\|x\| \leq (K - 1/2)\delta$,

then combining (5.4.1) and Corollary 3.3.7 gives

$$\|x-\widetilde{x}\| \leq \begin{cases} \frac{\delta d}{2N}\frac{\pi}{\sqrt{3}}(d+2) & \text{if } N \text{ is even,} \\ \frac{\delta d}{2N}\Big(\frac{\pi}{\sqrt{3}}(d+2)+1\Big) & \text{if } N \text{ is odd.} \end{cases}$$

Example 5.4.2 (Refined estimates for H_N^d with d even). As before, let the dimension d be even. Suppose that $x \in \mathbb{R}^d$ satisfies $\|x\| \leq (K-1/2)\delta$, and that N is sufficiently large with respect to δ. The frame coefficients of $x = (a_1, b_1, \ldots, a_{d/2}, b_{d/2}) \in \mathbb{R}^d$ with respect to H_N^d are given by $\{x_n^N\}_{n=1}^N = \{h(n/N)\}_{n=1}^N$, where

$$h(t) = \sqrt{\frac{2}{d}}\Big(\sum_{j=1}^{d/2} a_j \cos(2\pi jt) + \sum_{j=1}^{d/2} b_j \sin(2\pi jt)\Big).$$

We shall verify that for $f_n^N = e_n^N - e_{n+1}^N$ we have

$$\big\|f_n^N\big\| \lesssim \frac{1}{N} \quad \text{and} \quad \big\|f_n^N - f_{n+1}^N\big\| \lesssim 1N^2.$$

From the proof of Theorem 2.4.5, we have

$$\begin{aligned} \big\|\Delta^k e_n\big\| &= \sqrt{\frac{2}{d}}\left(\sum_{j=1}^{d/2}\Big(2\sin\frac{\pi j}{N}\Big)^{2k}\right)^{1/2} \\ &\leq \sqrt{\frac{2}{d}}2^k\left(\sum_{j=1}^{d/2}\Big(\frac{\pi j}{N}\Big)^{2k}\right)^{1/2} \\ &= \sqrt{\frac{2}{d}}2^k\Big(\frac{\pi}{N}\Big)^k\Big(\sum_{j=1}^{d/2} j^{2k}\Big)^{1/2} \\ &\lesssim \frac{1}{N^k}. \end{aligned}$$

So for $k=1$, $\big\|f_n^N\big\| = \|\Delta e_n\| \lesssim 1/N$ and for $k=2$, $\big\|f_n^N - f_{n+1}^N\big\| = \|\Delta^2 e_n\| \lesssim 1/N^2$.

To be specific, let $x = (1/\pi, 1/50, \sqrt{3/17}, 1/e) \in \mathbb{R}^4$. Then for this choice of x it is not hard to verify that $h \in \mathcal{M}_{d/2}$. Hence, the first part of Corollary 5.3.1 gives

that if N is even then

$$\|x - \widetilde{x}\| \lesssim \frac{\delta \log N}{N^{5/4}},$$

and if N is odd then

$$\frac{\delta}{N} \lesssim \|x - \widetilde{x}\| \leq \frac{2\delta}{N}\Big(\frac{6\pi}{\sqrt{3}} + 1\Big).$$

In the next several pages are the series of figures showing the plots of the quantization error $\|x - \widetilde{x}_N\|$ as a function of N when the harmonic frames H_N^2 or H_N^4 are used to quantize various inputs. The variables $K = 1$ and $\delta = 2$ are used in the $\Sigma\Delta$ scheme producing the alphabet $\{-1, 1\}$. For comparison, the figures also show the plots of $1/N$ as a dash-dotted line and of $1/N^{1.25}$ as a solid line. The parity of N is also shown on the plots as "crosses" when N is an odd integer and as "dots" when N is an even integer. The y-axis is scaled logarithmically while the x-axis is scaled linearly for the purpose of spreading the plots. In general, we see that the plots seem to confirm the theory well. We see that each plot is globally decreasing and has two portions that are globally "parallel" to the line $1/N$ if the integer N is odd and to the line $1/N^{1.25}$ if the integer N is even.

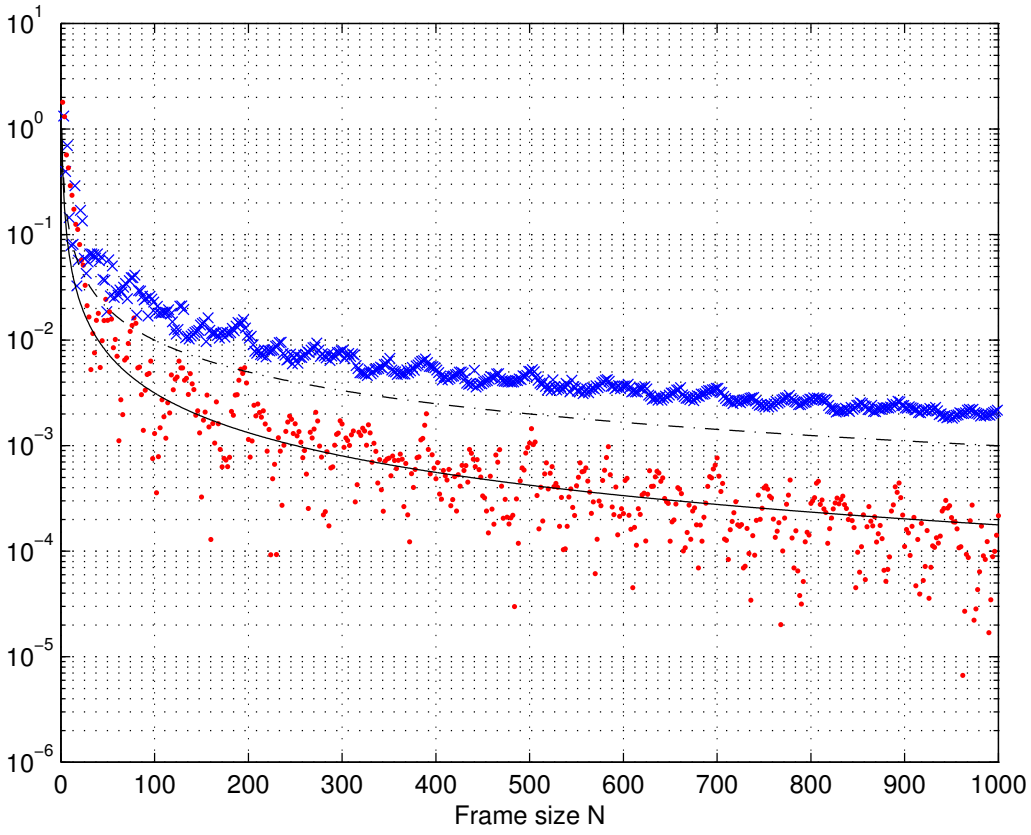

Figure 5.2: This plot shows the quantization error associated to quantizing the signal $x = (1/10, 1/20)$. It exemplifies one of the most typical shape patterns in the plots of quantization error.

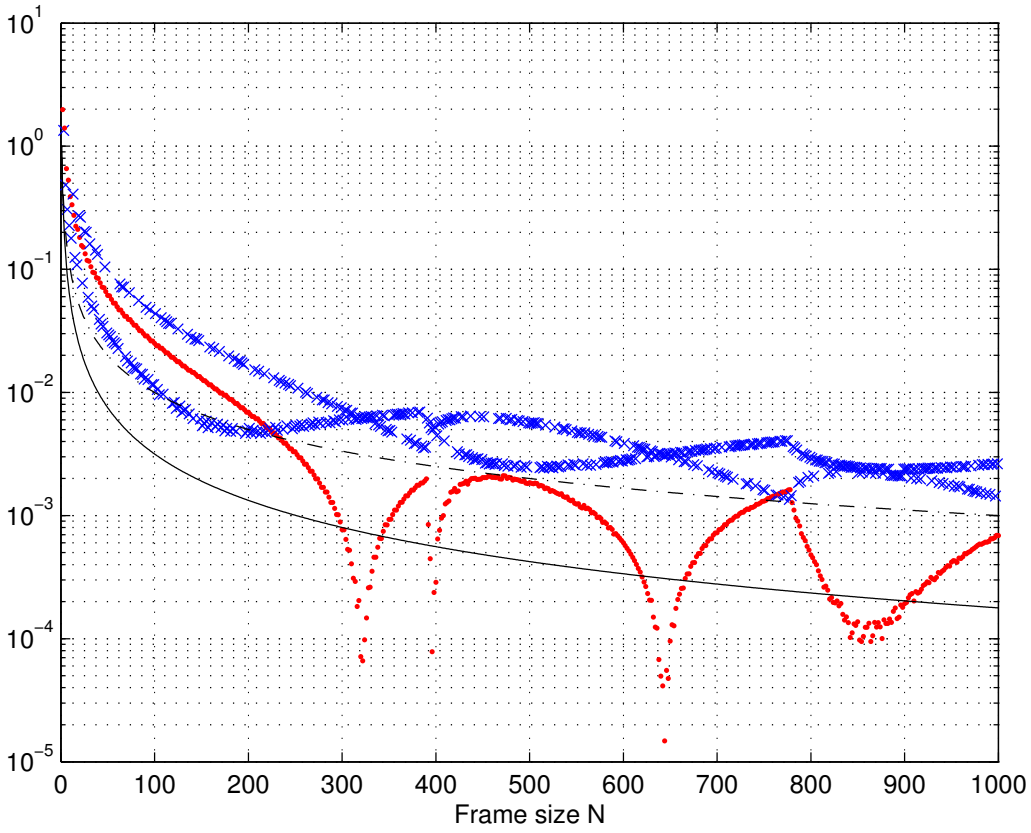

Figure 5.3: This plot shows the quantization error associated to quantizing the signal $x = (1/100, 1/200)$. It exemplifies another interesting shape pattern. The graph corresponding to odd integers splits itself into two parts while the graph corresponding to even integers stays together but jumps up and down.

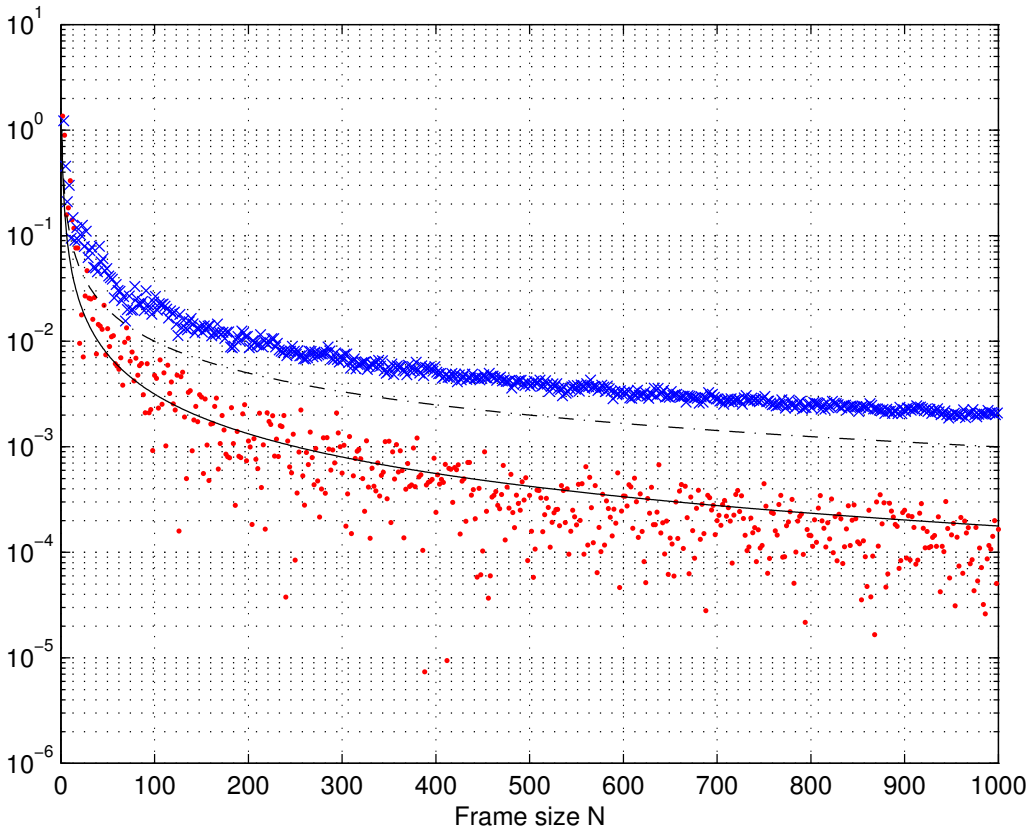

Figure 5.4: This plot shows the quantization error associated to quantizing the signal $x = (1/\pi, \sqrt{3/17})$. This is considered a classic plot as it appears in the original paper [16].

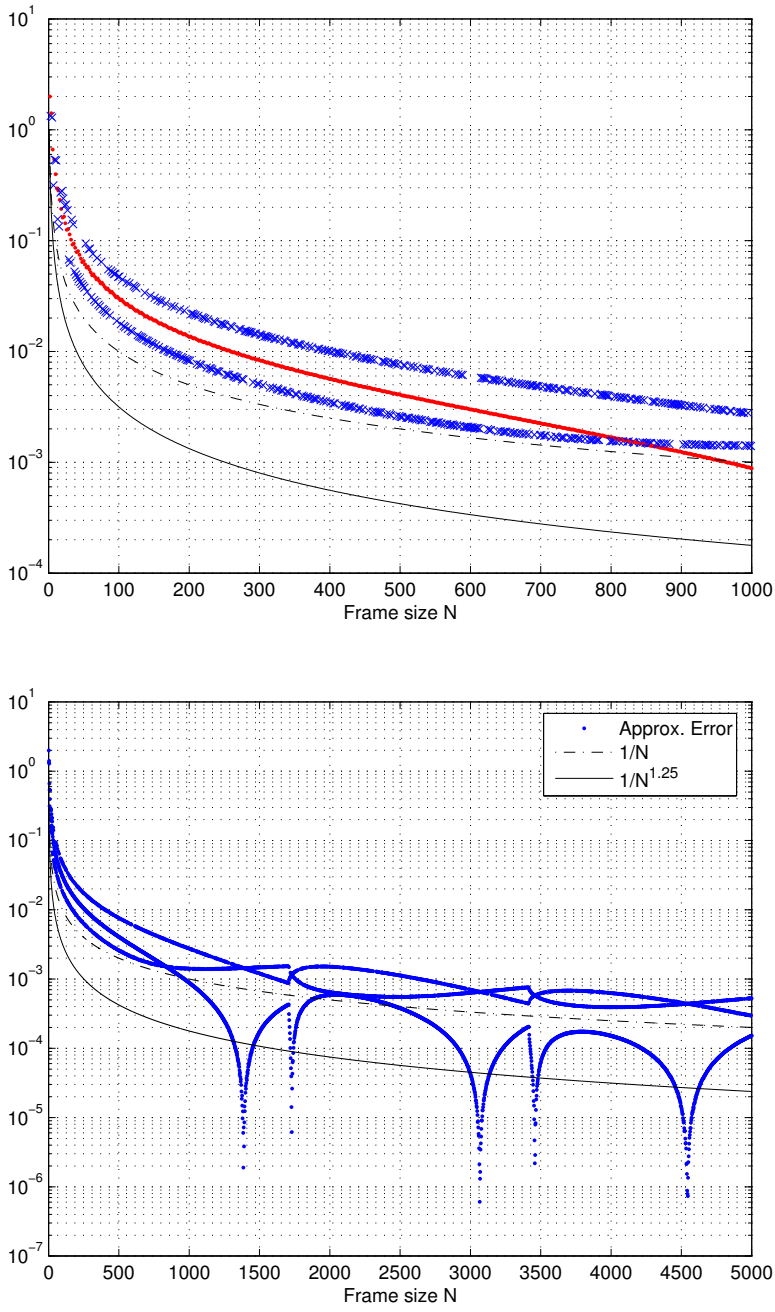

Figure 5.5: These two plots show the quantization error associated to quantizing the signal $x = (0.0018, 0.0014)$. The first plot includes the frame size up to 1000 while the second one includes up to 5000. We see in this case that one would have to increase the frame size in order to see the behavior of graphs more properly.

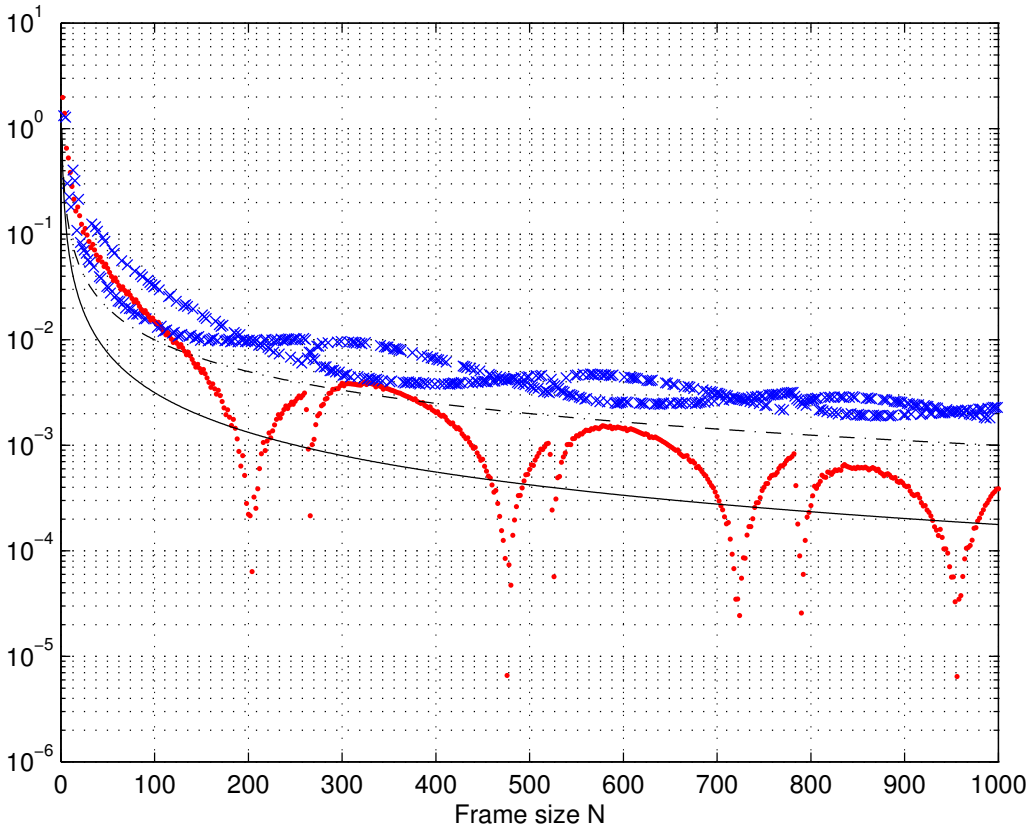

Figure 5.6: This plot shows the quantization error associated to quantizing the signal $x = (1/100, 1/100))$ with $\|x\| = \sqrt{2}/100$.

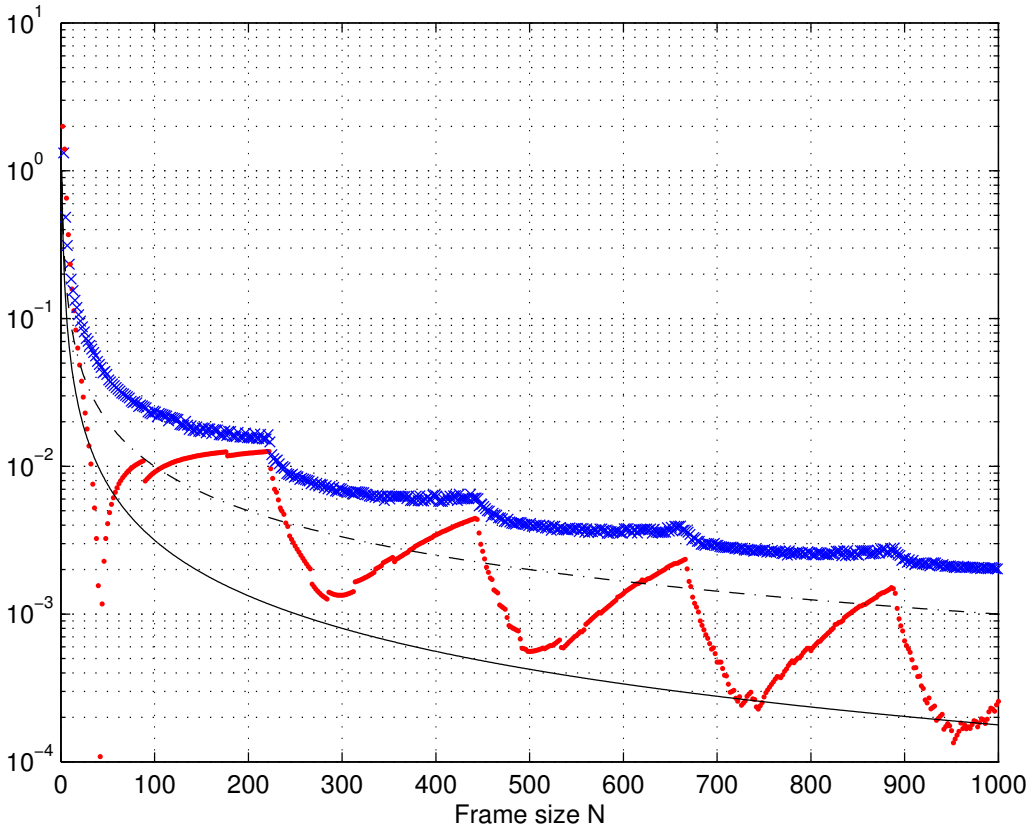

Figure 5.7: This plot shows the quantization error associated to quantizing the signal $x = (1/1000, \sqrt{199}/1000)$ with $\|x\| = \sqrt{2}/100$, which is the same norm as that of the signal in the previous Figure 5.6. This plot demonstrates that the shape of graphs does not depend on the norm of signals.

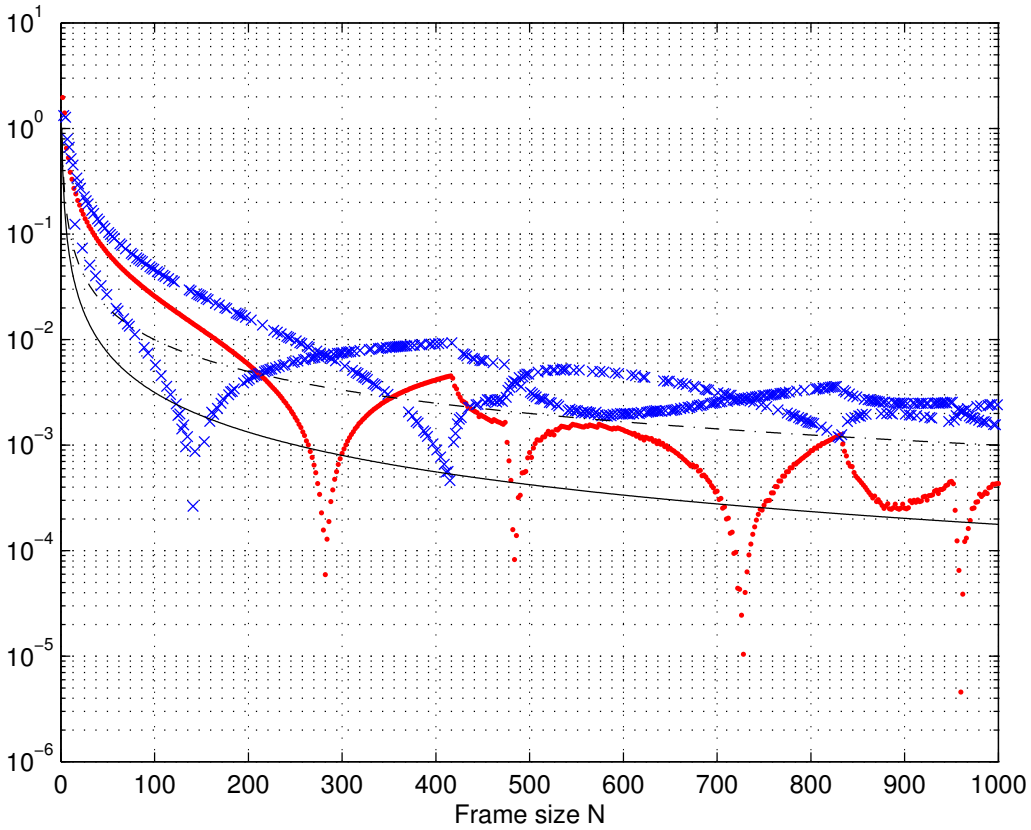

Figure 5.8: This plot shows the quantization error associated to quantizing the signal $x = (\sqrt{199}/1000, 1/1000)$ which is in a different order from the signal in the previous Figure 5.7. This demonstrates that the order of quantization affects the shape of graphs.

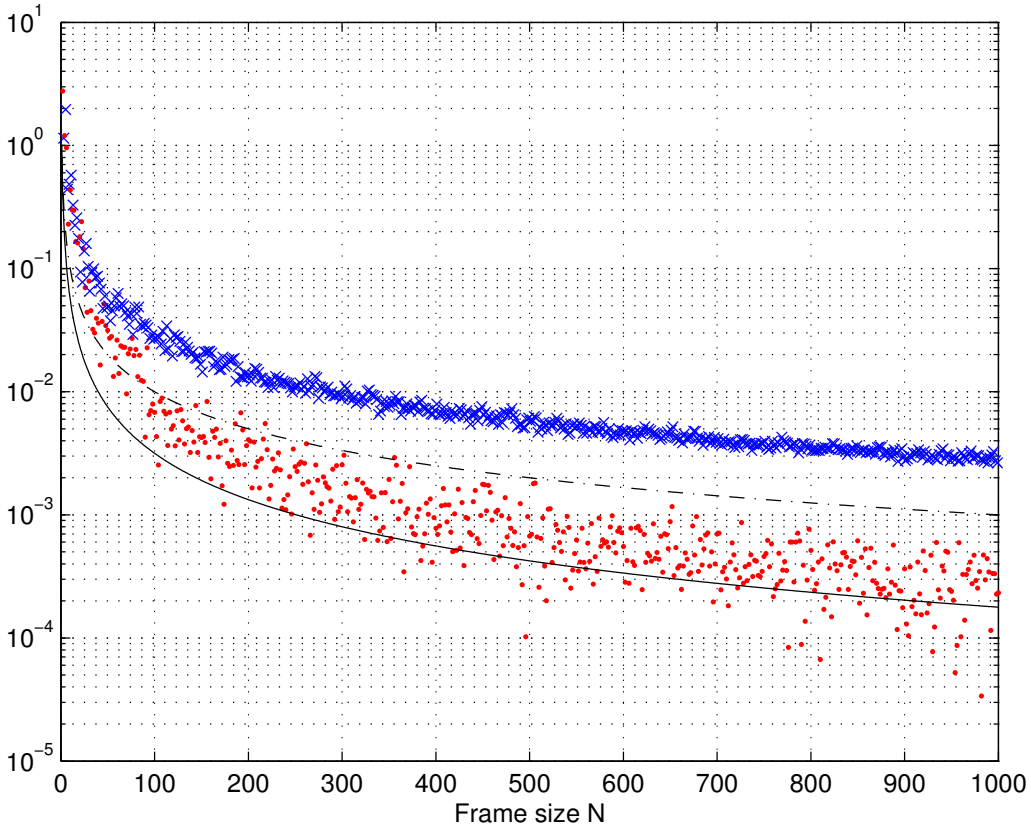

Figure 5.9: This plot shows another classic example in [16] which is the quantization error associated to quantizing the signal $x = (1/\pi, 1/50, \sqrt{3/17}, 1/e) \in \mathbb{R}^4$.

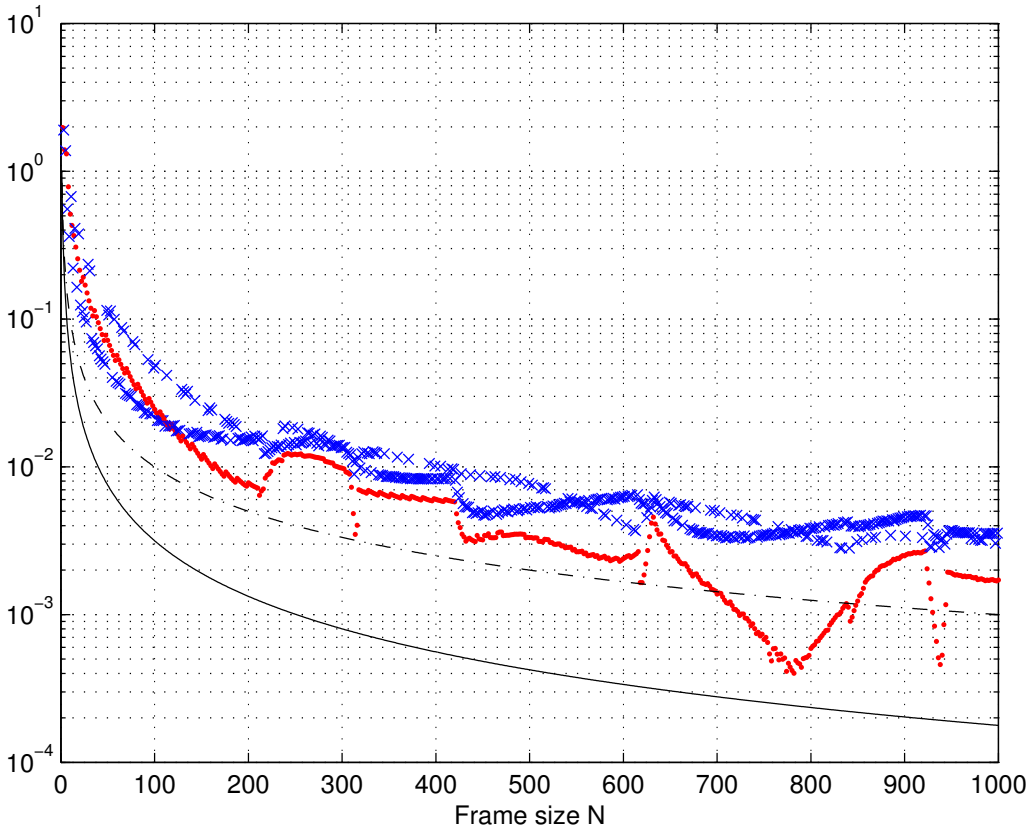

Figure 5.10: This plot shows the quantization error associated to quantizing the signal $x = (1/100, 1/100, 1/100, 1/100)$ with $\|x\| = 2/50$. It demonstrates that the shape pattern gets more complicated as the dimension increases.

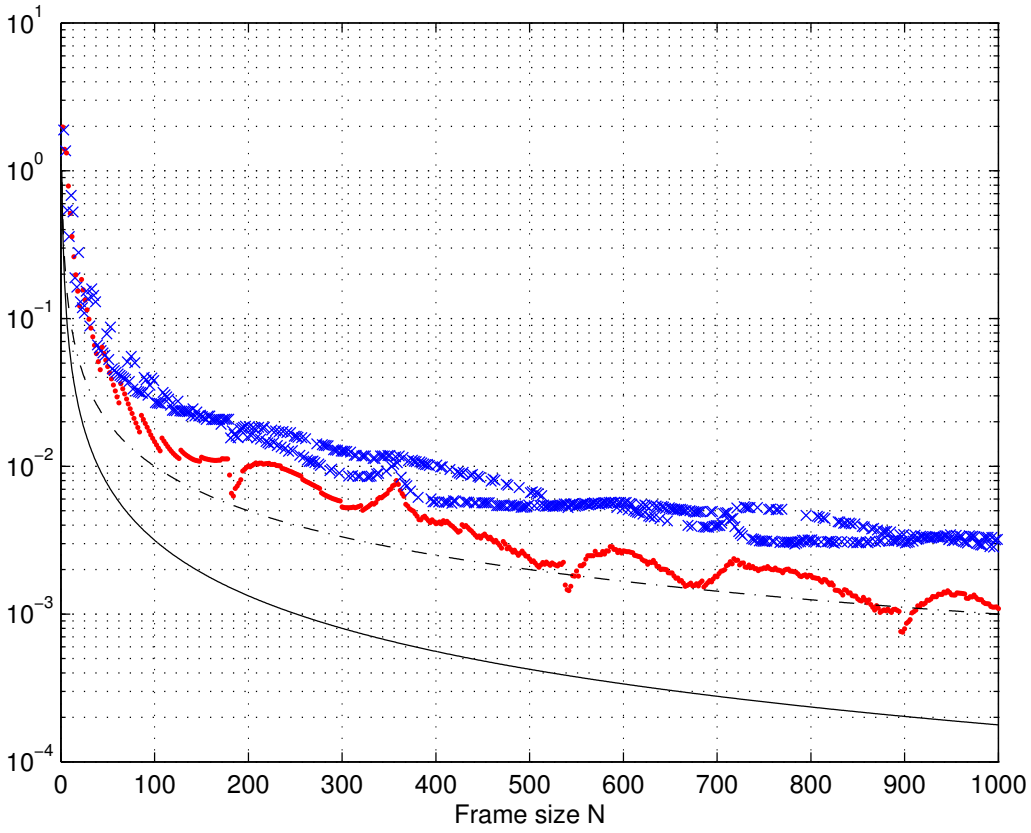

Figure 5.11: This plot shows the quantization error associated to quantizing the signal $x = (1/100, 1/100, 1/1000, \sqrt{1/2500 - 2 \times 10^{-4} - 10^{-6}})$ with $\|x\| = 2/50$, which is the same norm as that of the signal in the previous Figure 5.10. It emphasizes the independence of norms to the shape of graphs in higher dimensions.

BIBLIOGRAPHY

[1] N. Thao and M. Vetterli, "Deterministic analysis of oversampled A/D conversion and decoding improvement based on consistent estimates," *IEEE Transactions on Signal Processing*, vol. 42, no. 3, pp. 519–531, March 1994.

[2] V. Goyal, M. Vetterli, and N. Thao, "Quantized overcomplete expansions in $\mathbb{R}^n$: Analysis, synthesis, and algorithms," *IEEE Transactions on Information Theory*, vol. 44, no. 1, pp. 16–31, January 1998.

[3] 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*, vol. 158, no. 2, pp. 679–710, 2003.

[4] C. Güntürk, J. Lagarias, and V. Vaishampayan, "On the robustness of single loop sigma-delta modulation," *IEEE Trans. Information Theory*, vol. 47, No. 5, pp. 1735–1744, July 2001.

[5] Ö. Yılmaz, "Stability analysis for several sigma-delta methods of coarse quantization of bandlimited functions," *Constructive Approximation*, vol. 18, pp. 599–623, 2002.

[6] Ö. Yılmaz, "Coarse quantization of highly redundant time-frequency representations of square-integrable functions," *Appl. Comput. Harmon. Anal.* , vol. 14, pp. 107–132, 2003.

[7] C. Güntürk, "Approximating a bandlimited function using very coarsely quantized data: Improved error estimates in sigma-delta modulation," *J. Amer. Math. Soc.* , vol. 17, no. 1, pp.229–242, August 2003.

[8] C. Güntürk, "Harmonic analysis of two problems in signal quantization and compression," PhD dissertation of Program in Applied and Computational Mathematics, Princeton University, October 2000.

[9] W. Chen and B. Han, "Improving the accuracy estimate for the first order sigma-delta modulator," *J. Amer. Math. Soc.*, submitted in 2003.

[10] P. Casazza and J. Kovačević, "Equal-norm tight frames with erasures," *Advances in Computational Mathematics*, vol. 18, pp. 387–430, 2003.

[11] O. Christensen, *An Introduction to Frames and Riesz Bases*. Boston, MA: Birkhäuser, 2003.

[12] G. Zimmermann, "Normalized tight frames in finite dimensions," in *Recent Progress in Multivariate Approximation*, K. Jetter, W. Haussmann, and M. Reimer, Eds. Birkhäuser, 2001.

[13] J. Benedetto, *Harmonic Analysis and Applications*. Boca Raton, FL: CRC Press, 1997.

[14] J. Benedetto and M. Fickus, "Finite normalized tight frames," *Advances in Computational Mathematics*, vol. 18, pp. 357–385, 2003.

[15] J. Benedetto, A. Powell, and Ö. Yılmaz, "Second order sigma-delta ($\Sigma\Delta$) quantization of finite frame expansions," *Preprint*, 2004.

[16] J. Benedetto, A. Powell, and Ö. Yılmaz, "Sigma-delta ($\Sigma\Delta$) quantization and finite frames," *Preprint*, 2004.

[17] L. Kuipers and H. Niederreiter, *Uniform Distribution of Sequences*. New York: Wiley-Interscience, 1974.

[18] R. Duffin and A. Schaeffer, "A class of nonharmonic Fourier series," *Trans. Amer. Math. Soc.*, vol. 72, pp. 341–366, 1952.

[19] I. Daubechies, A. Grossmann, Y. Meyer, "Painless nonorthogonal expansions," *J. Math. Phys.*, vol. 27, pp. 1271–1283, 1986.

[20] G. Zimmermann, "Normalized tight frames in finite dimensions," *Recent Progress in Multivariate Approximation*, vol. 18, pp. 249–252, 2001.

[21] W. Kosmala, *A Friendly Introduction to Analysis: Single and Multivariable*. Upper Saddle River, NJ: Pearson Prentice Hall, 2004.

[22] H. Sohrab, *Basic Real Analysis*. Boston, MA: Birkhäuser, 2003.

[23] H. Inose, Y. Yasuda, and J. Murakami, "A telemetering system code modulation-$\Delta-\Sigma$ modulation," *IRE Trans. Space Elect. Telemetry*, vol. SET-8, pp. 204–209, September 1962.

[24] J. Candy, "A use of double integeration in Sigma-Delta modulation," *IEEE Trans. on Communications*, vol. COM-33, pp. 249–258, March 1985.

[25] J. Candy, "A use of limit cycle oscillation to obtain robust analog-to-digital converters," *IEEE Trans. Communications*, vol. 22, pp. 298–305, 1974.

[26] J. Candy and Y. Benjamin, "The structure of quantization noise from sigma-delta modulation," *IEEE Trans. Communications*, vol. 29, pp. 1316–1323, 1981.

[27] Y. Meyer, *Wavelets and Operator*. Cambridge University Press, 1992.