ABSTRACT

Title of dissertation: FRAME QUANTIZATION THEORY AND EQUIANGULAR TIGHT FRAMES

Onur Oktay
Doctor of Philosophy, 2007

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

In this thesis, we first consider the finite frame quantization. We make a signal-wise comparison of PCM and first order Sigma-Delta quantization. We show that Sigma-Delta quantization achieves smaller signal-wise quantization error bounds for a class of low amplitude signals. Then, we propose two new quantization methods for finite frames. First method is a variable bit-rate quantization algorithm. Given a finite signal and a predetermined error margin, this method calculates the number of bits necessary to quantize this signal within the pre-specified error margin. Second method is a 1-bit quantization technique that uses functional minimization methods. We first translate the combinatorial quantization problem into an analytic one. Then, we show that the solutions of this this analytic problem correspond to 1-bit quantized estimates of a given finite signal.

Second, we focus on finite equiangular tight frames. We show that equiangular tight frames are minimizers of certain functionals. We also give a characterization of equiangular tight frames with maximum possible redundancy.

FRAME QUANTIZATION THEORY AND EQUIANGULAR TIGHT FRAMES

by

Onur Oktay

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
2007

Advisory Committee:
Professor John J. Benedetto, Chair/Advisor
Professor Raymond L. Johnson
Professor Kasso Okoudjou
Professor Wojciech Czaja
Professor Sennur Ulukus

Table of Contents

List of Figures

Chapter 1

Introduction

1.1 Background

The concept of frames was first introduced by Duffin and Schaeffer [33] in the context of nonharmonic Fourier series. They defined a sequence $\{e_n(t) = e^{i\gamma_n t}\}$ of exponentials to be a frame for $L^2(-\Omega, \Omega)$ if there are global positive constants A and B such that

$$\forall f \in L^2(-\Omega,\Omega), \quad A\|f\|_{L^2}^2 \leq \sum_{n\in\mathbb{Z}} |\int_{-\Omega}^{\Omega} f(t)e^{-i\gamma_n t}dt|^2 \leq B\|f\|_{L^2}^2.$$

Moreover, if $\{e_n\}$ is a frame, then every f has a representation of the form

$$f = \sum_{n\in\mathbb{Z}} c_n(f)e^{i\gamma_n t}$$

for some sequence $\{c_n(f)\} \in \ell^2(\mathbb{Z})$ of coefficients.

Since Duffin and Schaeffer, frames have been studied extensively. A general theory of frames for Hilbert spaces has been developed. According to this general theory, frames are overcomplete systems that have many properties enjoyed by bases, such as the linear reconstruction property. Furthermore, frames have additional properties that bases do not possess. For instance, there is a wide variety of choices of coefficients in a frame expansion due to overcompleteness, whereas the coefficients in a basis expansion are uniquely determined.

Overcompleteness is a distinguishing property of frames that has an important

role in many modern applications. A standard example is sampling theory for bandlimited signals, where oversampling is used for stable reconstruction of signals. Another example is digital signal processing, where redundancy is used to reduce additive noise and overcome the effect of package loss.

A frame expansion of a signal x perfectly represents x. The frame coefficients in a frame expansion generally come from a continuous range of numbers. However, many modern applications require digital data, so any frame representation of x must be in *quantized* form in a digital environment.

Pulse Code Modulation (PCM) and Sigma-Delta quantization are two industry-standards for quantization in digital signal processing. PCM is a memoryless, fine quantization method, which simply rounds off each frame coefficient to the nearest element in a pre-specified *alphabet* of numbers. Sigma-Delta quantization is a coarse quantization method, which is associated with redundant *dictionaries*, such as frames. While PCM relies on fine quantization to minimize quantization error, Sigma-Delta shapes the quantization noise in a way that a major component of the noise stays in a space, which can later be eliminated during reconstruction. For instance, in the setting of bandlimited signals, Sigma-Delta quantization error usually has small in band frequency components and larger out-of-band frequency components [41, 12]. This phenomenon is known as the *noise shaping property* of Sigma-Delta quantization [41]. Daubechies and DeVore gave a more detailed mathematical analysis of Sigma-Delta quantization for bandlimited signals in [28]. They showed that, given a signal x with *stable* rth order Sigma-Delta estimate $\widetilde{x}$, the

quantization noise $x - \widetilde{x}$ satisfies the estimate,

$$\forall t \in \mathbb{R}, \quad |x(t) - \widetilde{x}(t)| \leq K\lambda^{-r},$$

where K is a constant depending on the reconstruction filter (and, thus, also depends on the bandwidth) and λ is the oversampling rate.

Unlike the samples of a bandlimited function, in many applications data does not always naturally come from an infinite dimensional structure. Finite frames are designated to analyze finite dimensional, but potentially large amounts of data.

Finite frames are also potentially useful for data coming from an infinite dimensional structure. One has to be careful with truncation errors for an infinite frame expansion. Depending on the convergence property of an infinite frame expansion, the size of the truncation error might be substantially large. There is no truncation error problem for finite frame expansions.

Finite frames are also useful in other applications, for example, in wireless communications for *codebook* design for code division multiple access (CDMA) systems [74, 72].

Finite frames have been studied extensively, and many properties of finite frames are very well understood, e.g. [5, 78, 16, 69]. Benedetto, Powell, and Yilmaz gave a mathematical analysis of Sigma-Delta quantization for finite frames in [10, 9]. Cvetković [22], Goyal, Vetterli, and Thao [40] showed that PCM quantization error can be improved using *consistent estimates*. There are many other contributions, e.g., [75, 8]. However, there are still many open problems in finite frame quantization theory.

1.2 Organization of the Thesis and New Results

Section 1.3 contains a basic overview of frames for Hilbert spaces.

In Chapter 2, we make a signal-wise comparison of PCM and first order Sigma-Delta quantization for finite frames. Section 2.1 contains a brief overview of the problem, and states established comparison results for the worst case quantization error, as well as for the mean-squared quantization error. Section 2.2 and Section 2.3 present the new results in this chapter.

In Chapter 3, we propose two new quantization techniques for finite frames. Section 3.1 contains a general description and properties of a *perfect quantizer*. In Section 3.2, we discuss Sigma-Delta quantization in the context of sparse matrices and periodic solutions of discrete dynamical systems. In Section 3.3, we propose a new adaptive bit-rate quantization method, and in Section 3.4, we propose another new 1-bit quantization method.

Chapter 4 is devoted to finite equiangular tight frames. Section 4.1 contains known results about equiangular tight frames, and their relations to other problems. Section 4.2 presents the new results of this chapter. Section 4.2.1 shows that equiangular tight frames are the minimizers of a class of scalar-valued functions. Section 4.2.2 gives a characterization of equiangular tight frames with maximum possible redundancy.

1.3 Frames

Definition 1. Let H be a separable Hilbert space. A set $F = \{e_j\}_{j\in J} \subseteq H$ is a *frame* for H if

$$\exists A, B > 0 \quad \text{such that} \quad \forall x \in H, \quad A\|x\|^2 \leq \sum_{j\in J} |\langle x, e_j\rangle|^2 \leq B\|x\|^2.$$

A frame F is a *tight frame* if we can choose $A = B$. If, in addition, each e_j is unit-norm, then F is a *unit-norm tight frame*.

Example 1. Let $PW_\Omega(\mathbb{R})$ be the set of square integrable functions with compactly supported Fourier transforms, which are supported in the interval $[-\Omega, \Omega]$. Let $T > 0$ such that $2T\Omega \leq 1$, and let $s \in L^2(\mathbb{R})$ with the Fourier transform $\hat{s}$, which satisfies

$$\begin{aligned} \hat{s}(\gamma) &= 1 \quad \text{if } |\gamma| \leq \Omega, \\ \hat{s}(\gamma) &= 0 \quad \text{if } |\gamma| \geq 1/(2T), \\ 0 \leq \hat{s}(\gamma) &\leq 1 \quad \text{if } \Omega < |\gamma| < 1/(2T). \end{aligned}$$

In particular, we can choose

$$s(t) = \frac{\sin 2\pi\Omega t}{\pi t}.$$

Let $s_n(.) = s(. - nT)$. Then, $\{s_n\}_{n\in\mathbb{Z}}$ is a tight frame for $PW_\Omega(\mathbb{R})$ with the frame constant $A = T^{-1}$. In fact, by the Classical Sampling Theorem, we have

$$\forall x \in PW_\Omega(\mathbb{R}), \quad x(t) = T\sum_{n\in\mathbb{Z}} x(nT)s(t - nT), \tag{1.1}$$

and also $\langle x, s_n\rangle = x(nT)$. In particular,

$$\forall x \in PW_\Omega(\mathbb{R}), \quad \|x\|^2_{L^2(\mathbb{R})} = T\sum_{n\in\mathbb{Z}} |\langle x, s_n\rangle|^2 = T\sum_{n\in\mathbb{Z}} |x(nT)|^2.$$

There are four operators associated with every frame. These are given in Definition 2

Definition 2. Let H be a separable Hilbert space, and let $F = \{e_j\}_{j\in J}$ be a frame for H.

(i) The linear function $L : H \to \ell^2(J)$ defined by $Lx = \{\langle x, e_j\rangle\}_{j\in J}$ is the *Bessel map* or the *analysis operator* for F.

(ii) The Hilbert space adjoint of L, L^* is the *synthesis operator*, and it satisfies the property

$$\forall c = (c_j)_{j\in J} \in \ell^2(J), \quad L^*c = \sum_{j\in J} c_j e_j. \tag{1.2}$$

(iii) $S = L^*L : H \to H$ is the *frame operator*, and it satisfies

$$\forall x \in H, \quad Sx = \sum_{j\in J} \langle x, e_j\rangle e_j. \tag{1.3}$$

(iv) $G = LL^* : \ell^2(J) \to \ell^2(J)$ is the *Grammian operator*.

Theorem 1. L^* can, in fact, be defined by (1.2).

Proof. By definition of the Hilbert space adjoints,

$$\forall x \in H, \quad \forall c = (c_j)_{j\in J} \in \ell^2(J), \quad \langle Lx, c\rangle = \langle x, L^*c\rangle.$$

Then,

$$\begin{aligned} \langle x, L^*c\rangle &= \langle Lx, c\rangle \\ &= \sum_{j\in J} \langle x, e_j\rangle \overline{c_j} \\ &= \langle x, \sum_{j\in J} c_j e_j\rangle. \end{aligned}$$

Since this is true for every $x \in H$, the result follows. □

Theorem 2. S is positive definite, and it satisfies $AI \leq S \leq BI$, where I is the identity operator on H.

Proof. By definition of S,

$$\forall x \in H, \quad \langle Sx, x\rangle = \sum_{j\in J} |\langle x, e_j\rangle|^2,$$

so $A\|x\|^2 \leq \langle Sx, x\rangle \leq B\|x\|^2$. Hence, the result follows. □

Definition 3. Let $\widetilde{e_j} = S^{-1}e_j$. Then, $\widetilde{F} = \{\widetilde{e_j}\}_{j\in J}$ is called the *canonical dual frame* of F. In this case, the Bessel map of the canonical dual frame is denoted by $\widetilde{L}$.

Theorem 3. Let $F = \{e_j\}_{j\in J}$ be a frame for H, and let $\widetilde{F} = \{\widetilde{e_j}\}_{j\in J}$ be the its canonical dual. Then, $\widetilde{F} = \{\widetilde{e_j}\}_{j\in J}$ is a frame with frame constants B^{-1} and A^{-1}, and for every $x \in H$, the following reconstruction formulas hold

$$\begin{aligned} x &= \sum_{j\in J} \langle x, \widetilde{e_j}\rangle e_j, \\ x &= \sum_{j\in J} \langle x, e_j\rangle \widetilde{e_j}. \end{aligned}$$

In particular, $L^*\widetilde{L} = I$ and $\widetilde{L}^*L = I$, where I is the identity operator on H. Furthermore, the frame operator of the canonical dual frame is S^{-1}, it is positive definite, and it satisfies $B^{-1}I \leq S^{-1} \leq A^{-1}I$.

Proof. Since S is positive definite, by the spectral theorem [67], there is an orthonormal set $\{v_k\}$ of eigenvectors of S, which is a basis for H, and

$$\forall x \in H, \quad Sx = \sum_k \lambda_k \langle x, v_k\rangle v_k,$$

where λ_k is the eigenvalue of S corresponding to v_k. Since $A\|x\|^2 \leq \langle Sx, x\rangle \leq B\|x\|^2$, any eigenvalue of S satisfies $A \leq \lambda_k \leq B$. In fact,

$$A = A\|v_k\|^2 \leq \langle Sv_k, v_k\rangle = \lambda_k \leq B\|v_k\|^2 = B.$$

S^{-1} clearly satisfies $S^{-1}v_k = \lambda_k^{-1}v_k$, and since $\{v_k\}$ is an orthonormal basis for H, we have

$$\forall x \in H, \quad S^{-1}x = \sum_k \lambda_k^{-1}\langle x, v_k\rangle v_k.$$

Therefore

$$B^{-1} \leq \inf_k \lambda_k^{-1} \leq \langle S^{-1}x, x\rangle \leq \sup_k \lambda_k^{-1} \leq A^{-1}.$$

Therefore, S^{-1} is positive definite, and it satisfies $B^{-1}I \leq S^{-1} \leq A^{-1}I$.

Next, since S^{-1} is positive definite, so self adjoint, we have

$$\widetilde{S}x = \sum_{j\in J}\langle x, \widetilde{e_j}\rangle\widetilde{e_j} = \sum_{j\in J}\langle x, S^{-1}e_j\rangle S^{-1}e_j = \sum_{j\in J} S^{-1}\langle S^{-1}x, e_j\rangle e_j = S^{-1}SS^{-1}x = S^{-1}x,$$

so S^{-1} is the frame operator of the dual frame $\widetilde{F}$.

Finally, for every $x \in H$,

$$\begin{aligned}\sum_{j\in J}\langle x, \widetilde{e_j}\rangle e_j &= \sum_{j\in J}\langle S^{-1}x, e_j\rangle e_j = SS^{-1}x = x,\\ \sum_{j\in J}\langle x, e_j\rangle \widetilde{e_j} &= \sum_{j\in J} S^{-1}\langle x, e_j\rangle e_j = S^{-1}Sx = x.\end{aligned}$$

Also,

$$\begin{aligned}L^*\widetilde{L}x &= L^*(\langle x, \widetilde{e_j}\rangle)_{j\in J} = \sum_{j\in J}\langle x, \widetilde{e_j}\rangle e_j,\\ \widetilde{L}^*Lx &= \widetilde{L}^*(\langle x, e_n\rangle)_{j\in J} = \sum_{j\in J}\langle x, e_j\rangle \widetilde{e_j}.\end{aligned}$$

Hence, $L^*\widetilde{L} = I$ and $\widetilde{L}^*L = I$, by the reconstruction formulas. □

Definition 4. A frame $F = \{e_j\}_{j=1}^N$ for $\mathbb{F}^d$ with finite number of elements is called a *finite frame*. If F is unit-norm and tight, then it is called a finite unit-norm tight frame (FUNTF).

Theorem 4. a. Any spanning set $\{e_j\}_{j=1}^N$ in $\mathbb{F}^d$ is a frame for $\mathbb{F}^d$.

b. If $F = \{e_j\}_{j=1}^N$ is a FUNTF for $\mathbb{F}^d$ with frame constant A, then $A = N/d$.

Proof. *a.* Let $\{e_j\}_{j=1}^N$ be a spanning set for $\mathbb{F}^d$. Since $\{x \in \mathbb{F}^d : \|x\| = 1\}$ is compact and there is an $x_0, \|x_0\| = 1$ at which the continuous function $\sum_{j=1}^N |\langle x, e_j\rangle|^2$ attains its minimum value. Let $A = \sum_{j=1}^N |\langle x_0, e_j\rangle|^2$.

$$A = 0 \Rightarrow \forall j = 1, \ldots, N, \quad \langle x_0, e_j\rangle \Rightarrow x_0 \notin span\{e_j\}_{j=1}^N.$$

Therefore, $A > 0$. Moreover,

$$\forall x \in \mathbb{F}^d, \quad A \leq \sum_{j=1}^N |\langle \frac{x}{\|x\|}, e_j\rangle|^2 \Rightarrow A\|x\|^2 \leq \sum_{j=1}^N |\langle x, e_j\rangle|^2.$$

On the other hand,

$$\forall x \in \mathbb{F}^d, \quad \sum_{j=1}^N |\langle x, e_j\rangle|^2 \leq \|x\|^2 \sum_{j=1}^N \|e_j\|^2.$$

We can choose $B = \sum_{j=1}^N \|e_j\|^2$.

b. If F is a finite frame, L, S and G can be represented as matrices. In particular, since F is a FUNTF, $S = AI$, and $G = (\langle e_i, e_j\rangle)$. Using the property of traces,

$$Ad = trace(S) = trace(G) = \sum_{j=1}^N \|e_j\|^2 = N.$$

□

Chapter 2

Pointwise Comparison of PCM and First Order Sigma-Delta for Finite Frames

2.1 Background

Let $x \in \mathbb{F}^d$ ($\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$) represent a data vector, and let $F = \{e_n\}_{n=1}^N$ be a frame for $\mathbb{F}^d$ with dual frame $\widetilde{F} = \{\widetilde{e_n}\}_{n=1}^N$. In applications, it is sometimes more useful or more convenient to work with the sequence $\{\langle x, e_n\rangle\}$ of frame coefficients rather than the data vector x itself. Frame coefficients represent x perfectly, since we can reconstruct x from these coefficients by the reconstruction formula

$$x = \sum_{n=1}^{N} \langle x, e_n\rangle \widetilde{e}_n.$$

Generally, $\{\langle x, e_n\rangle\}$ consists of arbitrary real or complex numbers. However, many digital signal processing applications require digital data. In such digital applications, a finite set of numbers $\mathcal{A}$ is pre-specified, and all of the components of a datum in a digital system is represented with a number in this *alphabet* $\mathcal{A}$. The larger the size of the alphabet, the more bits are needed to decode the elements in this alphabet.

The *frame quantization problem* is the problem of finding q_n in this alphabet $\mathcal{A}$ such that the quantity

$$\sum_{n=1}^{N} q_n \widetilde{e}_n$$

is equal or close to x in some prescribed way. PCM and Sigma-Delta quantization are two industry standards for quantization.

A quantization method is called *fine quantization* if the method uses a *high resolution* alphabet, i.e., any two elements in the alphabet are very close to each other. Consequently, the size of the alphabet associated with this method is large. A quantization method is *coarse quantization* if the size of the alphabet is small. 16-bit PCM is an example of a fine quantization method, whereas 1-bit Sigma-Delta is a coarse quantization method.

Fine quantization methods rely on the high resolution of the alphabet. As a result, these methods are less robust to noise compared to coarse quantization. By robust, we mean the following: if we have a sequence of numbers q with entries coming from a high resolution alphabet, then even a small perturbation of the entries of q irreversibly changes q. On the other hand, an error caused by a noise of up to a certain magnitude, let us say 1, can be corrected if the entries of q are coming from $\{-1, 1\}$.

Coarse quantization methods can result in small quantization error when used with highly redundant expressions. Recently, Benedetto, Powell, and Yilmaz [10] showed that Sigma-Delta outperforms PCM in the worst-case error, and in the mean-squared error for signals $x \in \mathbb{R}^d$ normalized so that $\|x\| \leq 1$. Building on these results, we make a signal-wise comparison of PCM and Sigma-Delta quantization in this chapter.

We assume that $\mathbb{F} = \mathbb{C}$. Any frame for $\mathbb{R}^d$ is automatically a frame for $\mathbb{C}^d$, and the quantization schemes that we consider in this chapter, when restricted to

$\mathbb{R}^d$, coincide with the quantization schemes for real sequences. Therefore, all of the results in this chapter for $\mathbb{C}$ automatically hold for $\mathbb{R}$.

2.1.1 Overview of PCM and Sigma-Delta

Definition 5. For $K > 0$ and an integer $b \geq 2$, let $\delta = 2^{1-b}$. The *midrise quantization alphabet* is

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

and the associated *scalar uniform quantizer with step size* δ is given by

$$Q(u + iv) = \delta \left(\frac{1}{2} + \left\lfloor \frac{u}{\delta} \right\rfloor + i \left\lfloor \frac{v}{\delta} \right\rfloor \right).$$

Here, $b \geq 2$ represents the number of bits. We define the alphabet and the quantizer for the 1-bit case as follows.

$$\mathcal{A} = \{\pm 1 \pm i\}, \quad Q(u + iv) = \text{sign}(u) + i\text{sign}(v).$$

Notationally, we set

$$\text{sign}(u) = \begin{cases} 1, & \text{if } u \geq 0, \\ -1, & \text{if } u < 0. \end{cases}$$

PCM rounds off each frame coefficient to the nearest element in the alphabet, i.e.,

$$q_n = Q(\langle x, e_n \rangle), \tag{2.1}$$

whereas first order Sigma-Delta scheme defines (q_n) by means of the iterative scheme

$$\begin{aligned} u_n &= u_{n-1} + \langle x, e_n \rangle - q_n \\ q_n &= Q(\langle x, e_n \rangle + u_{n-1}). \end{aligned} \tag{2.2}$$

with the initial condition u_0.

In either quantization scheme, the quantized estimate $\widetilde{x}$ of x is given by

$$\widetilde{x} = \sum_{n=1}^{N} q_n \widetilde{e}_n,$$

where $\{\widetilde{e}_n\}_{n=1}^{N}$ is the dual frame.

Benedetto, Powell and Yilmaz [10] established a uniform upper bound for the first order Sigma-Delta quantization error. Theorem 5 and a proof can be found in [10].

Theorem 5. Let $F = \{e_n\}_{n=1}^{N}$ be a FUNTF for $\mathbb{R}^d$, let p be a permutation of $\{1, \ldots, N\}$, let $|u_0| \leq \delta/2$, and let $x \in \mathbb{R}^d$ satisfy $\|x\| \leq 1$. Let $\widetilde{x}$ denote the first order Sigma-Delta estimate for x. Then,

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

where

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

Theorem 6 generalizes Theorem 5 to the complex case. A proof of Theorem 6 is in [8].

Theorem 6. Let $F = \{e_n\}_{n=1}^{N}$ be a FUNTF for $\mathbb{C}^d$, let p be a permutation of $\{1, \ldots, N\}$, let $|u_0| \leq \delta/2$, and let $x \in \mathbb{C}^d$ satisfy $\|x\| \leq 1$. Let $\widetilde{x}$ denote the first order Sigma-Delta estimate for x. Then,

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

Both for the real and the complex case, the *state variable* u is bounded by $\delta/2$ in absolute value if u_0 is bounded by $\delta/2$ [10, 8]. Then, by the definition (2.2) of the first order Sigma-Delta scheme, one can show that

$$\forall n = 1, \ldots, N, \quad \left| \sum_{k=1}^{n} \langle x, e_n \rangle - \sum_{k=1}^{n} q_k \right| \leq \sqrt{2} \frac{\delta}{2}, \tag{2.3}$$

i.e., first order Sigma-Delta minimizes the running sums.

Building on the result of Theorem 5, Wang [75] gave an upper bound for the *frame variation* $\sigma(\{e_n\}_{n=1}^N, p)$ that increases slower than $\mathcal{O}(N)$ as $N \to \infty$. Using this upper bound, one can prove the Theorem 7. A proof of Theorem 7 can be found in [8].

Theorem 7. Let $F = \{e_n\}_{n=1}^N$ be a unit norm frame for $\mathbb{F}^d$, $d \geq 3$. There exists a permutation p of $\{1, \ldots, N\}$ such that

i. if $\mathbb{F} = \mathbb{R}$, then $\sigma(F, p) \leq 4\sqrt{d+3}\ N^{1-\frac{1}{d}} - 4\sqrt{d+3}$,

ii. if $\mathbb{F} = \mathbb{C}$, then $\sigma(F, p) \leq 4\sqrt{2d+3}\ N^{1-\frac{1}{2d}} - 4\sqrt{2d+3}$.

Moreover, if $x \in \mathbb{C}^d$, $\|x\| \leq 1$, then the first order Sigma-Delta quantization (2.2) error $\|x - \widetilde{x}\|$ satisfies

$$\begin{aligned} \|x - \widetilde{x}\| &\leq \sqrt{2}\delta d \left((1 - 2\sqrt{2d+3}) N^{-1} + 2\sqrt{2d+3} N^{-\frac{1}{2d}} \right) \\ &\leq M N^{-\frac{1}{2d}}, \end{aligned} \tag{2.4}$$

where $M = \sqrt{2}\delta d$.

2.1.2 Rate Distortion

Rate distortion theory was created by Claude Shannon in his foundational work on information theory, and now it is a major branch of information theory. Rate distortion theory addresses the problem of determining the minimal amount of information R that should be used, so that the input signal or data can be reconstructed at the receiver without exceeding a given distortion D.

The term *rate* refers to the minimal amount of information R. Therefore, the rate is a function of the distortion and the input signal.

Lossy compression techniques that are used in many of the existing audio, speech, image, and video compression uses the concept of rate distortion. Given a signal or a data stream, a lossy compression technique looks for an estimate, which can be stored using small number of bits, and, at the same time, is close to the original signal or data stream in some sense. This process is irreversible, i.e., we cannot obtain the original signal/data stream back from its estimate, hence the name lossy compression. In this context, the rate is understood as the number of bits per sample to be stored or transmitted, and the distortion is essentially the size of the error, which is the difference of the original signal/data stream and its estimate.

Both PCM and Sigma-Delta quantization can be considered as lossy compression techniques. For our discussion, the distortion is the distance between the original data vector $x \in \mathbb{C}^d$ and the quantized vector with respect to a FUNTF $F = \{e_n\}_{n=1}^N$, in a suitable metric on $\mathbb{C}^d$. The rate is bN, where b is the number of

bits for quantization, and N is the redundancy of the frame. If ρ is a metric on $\mathbb{C}^d$, $x \in \mathbb{C}^d$, and if

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

is the quantized estimate of x that b-bit PCM or Sigma-Delta produces, then, the rate distortion problem in our setting is the problem of finding b and N that results in the smallest value for bN such that

$$\rho(x, \widetilde{x}_b) \leq D,$$

for a given distortion D.

Generally, a b-bit quantization scheme with a FUNTF $F = \{e_n\}_{n=1}^N$ maps an $x \in \mathbb{C}^d$ to an element in the set of all possible quantized expansions

$$S = \left\{ \frac{d}{N} \sum_{n=1}^{N} q_n e_n : q_n \in \mathcal{A}_\delta^K \right\},$$

where $\mathcal{A}_\delta^K$ is given as in Definition 5. It is not hard to show that S has at most 2^{bN} elements (exactly 2^{bN} if all of the elements in S are distinct). Theorem 8 provides an information theoretic lower bound for the worst case error, which is independent of the quantization scheme.

Theorem 8. Let $\|.\|$ be a norm on $\mathbb{C}^d$. For a b-bit finite frame quantization scheme, let $\widetilde{x}_b$ denote the quantized estimate of an $x \in \mathbb{C}^d$. Then, the worst case error

$$\max_{\|x\| \leq 1} \|x - \widetilde{x}_b\|$$

is bounded below by $2^{-bN/d}$ for the unit ball $\{x \in \mathbb{C}^d : \|x\| \leq 1\}$.

Proof. Let r be equal to the worst case error. Let

$$\mathcal{B}_r(x) = \{\xi : \|x - \xi\| < r\}$$

denote the ball centered at x with radius r, and let $\mathcal{L}^d$ denote the Lebesgue measure on $\mathbb{C}^d$. Then, $\mathcal{B}_1(0) \subseteq \bigcup_{x\in S} \mathcal{B}_r(x)$. It is well known that the Lebesgue measure on $\mathbb{C}^d$ is translation invariant, and it satisfies

$$\forall A \subseteq \mathbb{C}^d, \forall t > 0, \quad \mathcal{L}^d(tA) = t^d \mathcal{L}^d(A).$$

Then,

$$\begin{aligned} \mathcal{B}_1(0) \subseteq \bigcup_{x\in S} \mathcal{B}_r(x) \quad &\Rightarrow \quad \mathcal{L}^d(\mathcal{B}_1(0)) \leq \sum_{x\in S} \mathcal{L}^d(\mathcal{B}_r(x)) = |S| r^d \mathcal{L}^d(\mathcal{B}_1(0)) \\ &\Rightarrow \quad r \geq |S|^{-1/d} \geq 2^{-bN/d}. \end{aligned}$$

□

If $\{v_n\}_{n=1}^d$ is an orthonormal basis for $\mathbb{C}^d$, then the b-bit PCM quantization error satisfies

$$2^{-b} \leq \|x - \sum_{k=1}^{d} Q(\langle x, v_n\rangle) v_n\| \leq \sqrt{2}(\sum_{k=1}^{d} \|v_n\|) 2^{-b}.$$

Therefore, PCM with an orthonormal basis is an asymptotically optimal quantization method in the rate distortion sense. However, using redundant expressions has its advantages over bases in some applications. For example, frames are used in noise reduction in communications (Theorem 44). Also, redundancy has a key role in overcoming the erasure problem in communications [49, 52, 53, 54]. We shall talk about these problems more in Section 4.1.4.

Even though PCM is optimal with an orthonormal basis, it is far from being optimal with redundant expressions for the worst case error [10]. First order $\Sigma\Delta$ is not (asymptotically) optimal in the rate distortion sense, either, but it outperforms PCM in the worst case error and in the expected mean-square error [10]. However, this leaves open the question of whether the signal-wise PCM error can be less than the signal-wise error for Sigma-Delta at specific signals.

In the remainder of this chapter, we investigate the class of signals where the signal-wise first order Sigma-Delta quantization error is less than the signal-wise PCM error. We use the same redundant frame $\{e_n\}_{n=1}^N$ for both quantization methods, which we choose to be a finite unit-norm tight frame.

In Section 2.2, we show that 1-bit Sigma-Delta totally outperforms 1-bit PCM for each x, $\|x\| \leq 1$. In Section 2.3, we show that 1-bit Sigma-Delta outperforms multibit PCM for a class of low amplitude signals. We also give certain properties of the quantization error function $\mathtt{err}_{PCM}(.)$ of multibit PCM for a family of structured FUNTFs.

2.2 Comparison of 1-bit PCM and 1-bit Sigma-Delta

Definition 6. Let $x \in \mathbb{C}^d$, let $F = \{e_n\}_{n=1}^N$ be a FUNTF for $\mathbb{C}^d$ with the analysis matrix L, and let $q_{PCM}(x,b)$ and $q_{\Sigma\Delta}(x,b)$ denote the quantized sequences given by b-bit PCM and b-bit Sigma-Delta, respectively. We define the *quantization error functions*

$$\mathtt{err}_{PCM}(x,F,b) = \|x - \frac{d}{N}L^* q_{PCM}(x)\|, \quad \mathtt{err}_{\Sigma\Delta}(x,F,b) = \|x - \frac{d}{N}L^* q_{\Sigma\Delta}(x)\|.$$

Notationally, we omit writing F when we compare two schemes with the same FUNTF, and we omit writing b when we compare at the same bit rate.

Theorem 9. Let $x \in \mathbb{C}^d$ satisfy $0 < \|x\| \leq 1$, and let $F = \{e_n\}_{n=1}^N$ be a FUNTF for $\mathbb{C}^d$. Then, the 1-bit PCM error satisfies

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

where

$$\alpha_F := \inf_{\|x\|=1} \frac{d}{N} \sum_{n=1}^{N} |Re(\langle x, e_n \rangle)| + |Im(\langle x, e_n \rangle)| - 1 \geq 0. \tag{2.5}$$

Proof. First, $Re(Q(a+ib)(a-ib)) = |a| + |b|$. Then,

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

□

In Lemma 2, we prove that α_F is always nonzero for a FUNTF, but first we need the following.

Definition 7. A frame F is *robust to 1-erasure* if for any $x \in F$, $F \backslash \{x\}$ still constitute a frame.

Theorem 10. Let $N > d$. Every FUNTF $F = \{e_n\}_{n=1}^N$ for $\mathbb{C}^d$ is robust to 1-erasure.

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

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

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

In general, if $N > rd$, then any FUNTF $\{e_n\}_{n=1}^N$ for $\mathbb{C}^d$ is robust to r-erasures, i.e., if we remove any r elements of the frame, the remaining vectors constitute a frame for $\mathbb{C}^d$ (Theorem 45).

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

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

Proof.

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

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

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

□

Lemma 2. Let $F = \{e_n\}_{n=1}^N$ be a FUNTF for $\mathbb{C}^d$ with the property

$$\forall k = 1,\dots,N, \quad e_k \in F \quad \text{and} \quad |\lambda| = 1 \Rightarrow \lambda e_k \notin F.$$

Then $\alpha_F > 0$.

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

$$|Re(\langle x, e_n\rangle)| \geq |Re(\langle x, e_n\rangle)|^2, \quad |Im(\langle x, e_n\rangle)| \geq |Im(\langle x, e_n\rangle)|^2. \tag{2.6}$$

Then,

$$\begin{aligned} \alpha_F &= \inf_{\|x\|=1} \frac{d}{N} \sum_{n=1}^{N} |Re(\langle x, e_n\rangle)| + |Im(\langle x, e_n\rangle)| - 1 \qquad (2.7) \\ &= \inf_{\|x\|=1} \frac{d}{N} \sum_{n=1}^{N} |Re(\langle x, e_n\rangle)| - |Re(\langle x, e_n\rangle)|^2 + |Im(\langle x, e_n\rangle)| - |Im(\langle x, e_n\rangle)|^2 \geq 0. \end{aligned}$$

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

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

In the latter case, we must have

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

by (2.6). Then, since

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

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

$$|\langle x_0, e_n\rangle| = |Re(\langle x_0, e_n\rangle)| + |Im(\langle x_0, e_n\rangle)|. \tag{2.8}$$

Then, (2.7) and (2.8) imply

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

Then, by Lemma 1, there is a w such that $\langle x_0, e_n\rangle e_n = w$ if $\langle x_0, e_n\rangle \neq 0$. Hence, there is a $v \in \mathbb{C}^d$, such that for every e_n, for which $\langle x_0, e_n\rangle \neq 0$, there is a $\lambda_n \in \mathbb{C}, |\lambda| = 1$ and $e_n = \lambda_n v$. But, by the hypothesis, there can only be one such frame element. Thus, there is only one frame element nonorthogonal to x_0. Erasing this element, remaining vectors would not span $\mathbb{C}^d$, i.e., F would not be robust. But, this is a contradiction to Theorem 10.

Therefore, $\alpha_F > 0$. □

Theorem 11. Let $\{F_N = \{e_n^N\}_{n=1}^N\}$ be a family of FUNTFs for $\mathbb{C}^d$. Then,

$$\forall \varepsilon > 0 \quad \exists N_0 > 0 \quad \forall N \geq N_0 \quad \mathtt{err}_{\Sigma\Delta}(x, F_N, 1) \leq \mathtt{err}_{PCM}(x, F_N, 1)$$

for every $0 < \|x\| \leq 1 - \varepsilon$.

Proof. By Theorem 7, for any $\|x\| \leq 1$ and for any N,

$$\mathtt{err}_{\Sigma\Delta}(x) \leq MN^{-1/2d}.$$

Then, by Theorem 9

$$\forall \varepsilon > 0 \quad \exists N_0 > 0 \quad \forall N \geq N_0 \quad MN^{-1/2d} \leq \varepsilon \leq 1 - \|x\| + \alpha_F \leq \mathtt{err}_{PCM}(x)$$

for every x, $0 < \|x\| \leq 1 - \varepsilon$. □

We want to note that the bound $N \geq (M/\varepsilon)^{2d}$ is a crude lower bound for N. In practice, we can choose a significantly small N that satisfies the condition of Theorem 11.

Let $\{F_N = \{e_n^N\}_{n=1}^N\}$ be a family of FUNTFs. If there is a positive uniform lower bound for (α_{F_N}), then we can improve the result of Theorem 11. Namely, we

can replace Theorem 11 by the assertion

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

The families $\{F_N\}$ of FUNTFs for which $\alpha_{F_N} \to 0$ are extreme cases, which we describe in Theorem 12.

Theorem 12. Let $\{F_N = \{e_n^N\}_{n=1}^N\}$ be a family of FUNTFs for $\mathbb{C}^d$ such that $\lim_{N\to\infty} \alpha_{F_N} = 0$. Then, there is an $x_0 \in \mathbb{C}^d$, $\|x_0\| = 1$ such that

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

Proof.

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

Let $x_N \in \mathbb{C}^d$, $\|x_N\| = 1$ be a point where $\sum_{n=1}^N |\langle x, e_n^N\rangle|$ attains its minimum. Since $\alpha_{F_N} \to 0$, we have

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

On the other hand,

$$\left| \frac{d}{N} \sum_{n=1}^{N} |\langle x_N, e_n^N\rangle| - \frac{d}{N} \sum_{n=1}^{N} |\langle x_0, e_n^N\rangle| \right| \leq d\|x_N - x_0\|. \tag{2.9}$$

Since the unit ball of $\mathbb{C}^d$ is compact, (x_N) has a convergent subsequence. Without loss of generality, assume that $\lim_{N\to\infty} x_N = x_0$. Letting $N \to \infty$ in (2.9), we obtain

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

Next, define the sets A_N^ε and B_N^ε,

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

Then,

$$\frac{d\varepsilon \; card B_N^\varepsilon}{N} \leq \frac{d}{N} \sum_{n=1}^{N} |\langle x_0, e_n^N\rangle| - 1.$$

Therefore,

$$\lim_{N\to\infty} \frac{card B_N^\varepsilon}{N} = 0, \quad \text{and so} \quad \lim_{N\to\infty} \frac{card A_N^\varepsilon}{N} = 1.$$

□

Theorem 13 gives an example of a family $\{F_N\}$ of frames for which the sequence (α_{F_N}) is bounded from below. The families given by Theorem 13 comes from a continuous curve in $\mathbb{R}^d$, which is of bounded variation. Such curves were named *frame paths* in [13].

Definition 8. A function $e : [a, b] \to \mathbb{C}^d$ is of *bounded variation (BV)* if there is a $K > 0$ such that for every $a \leq t_1 < t_2 < \cdots < t_N \leq b$,

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

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

Theorem 13. Let $e : [0, 1] \to \{x \in \mathbb{C}^d : \|x\| = 1\}$ be continuous function of bounded variation such that $F_N = (e(n/N))_{n=1}^N$ is a FUNTF for $\mathbb{C}^d$ for every N.

Then,

$$\exists N_0 > 0 \quad \forall N \geq N_0 \quad \texttt{err}_{\Sigma\Delta}(x, F_N, 1) \leq \texttt{err}_{PCM}(x, F_N, 1)$$

for every $0 < \|x\| \leq 1$.

Proof. Let $e_n^N = e(n/N)$. By Lemma 2, for any x, $\|x\| = 1$, we have

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

Also,

$$\begin{aligned}
& \lim_{N\to\infty} \frac{d}{N} \sum_{n=1}^{N} |Re(\langle x, e_n^N \rangle)| + |Im(\langle x, e_n^N \rangle)| - 1 \\
= & \lim_{N\to\infty} \left(\frac{d}{N} \sum_{n=1}^{N} |Re(\langle x, e(n/N) \rangle)| + |Im(\langle x, e(n/N) \rangle)| \right) \\
& - \lim_{N\to\infty} \left(\frac{d}{N} \sum_{n=1}^{N} |Re(\langle x, e(n/N) \rangle)|^2 + |Im(\langle x, e(n/N) \rangle)|^2 \right) \\
= & \; d \int_0^1 |Re(\langle x, e(t) \rangle)| + |Im(\langle x, e(t) \rangle)| - |Re(\langle x, e(t) \rangle)|^2 - |Im(\langle x, e(t) \rangle)|^2 \; dt.
\end{aligned} \tag{2.10}$$

The integrand in (2.10) cannot be equal to zero for every t. For a contradiction, assume the integrand is zero for every t. Then,

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

But, since

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

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

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

Since $x \neq 0$, there should exist a t^* such that $|\langle x, e(t^*) \rangle| = 1$ which implies that there is a $|\lambda_0| = 1$ such that $x = \lambda_0 e(t^*)$, and that $\langle x, e(t) \rangle = 0$ for every t for which

there is a $\lambda \in \mathbb{C}$, $|\lambda| = 1$ and $e(t) \neq \lambda e(t^*)$. But this contradicts the continuity of e. By contradiction, the integrand in (2.10) is not zero at every point.

Next, since the integrand is continuous, for each x, $\|x\| = 1$,

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

Then, since the unit ball of $\mathbb{C}^d$ is compact,

$$\alpha := d \inf_{\|x\|=1} \int_0^1 |Re(\langle x, e(t)\rangle)| + |Im(\langle x, e(t)\rangle)| - |Re(\langle x, e(t)\rangle)|^2 - |Im(\langle x, e(t)\rangle)|^2 \, dt > 0.$$

Clearly, $\lim_{N\to\infty} \alpha_{F_N} = \alpha$. Then, (α_{F_N}) is bounded below by a $\beta > 0$. For this β

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

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

Third, $\sum_{n=1}^N \|e_n - e_{n+1}\| \leq |e|_{BV} =: M$. Then, by Theorem 5, for every N,

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

Choose $N_0 \geq d(1 + M)/\beta$. Then

$$\forall N \geq N_0, \quad \mathtt{err}_{\Sigma\Delta}(x) \leq \frac{d}{N}(1 + M) \leq \beta \leq \alpha_{F_N} + 1 - \|x\| \leq \mathtt{err}_{PCM}(x)$$

for every $0 < \|x\| \leq 1$. □

Example 2. Real Harmonic Frames $H_N^d = \{e_n^N\}_{n=1}^N$ for $\mathbb{R}^d$ for $d = 2k$ are defined by

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

H_N^d come from the curve

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

by regularly sampling that curve. H_N^d is a FUNTF for each N. It can be shown that the frame variation of each H_N^d can be bounded by the number

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

The family H_N^2 is also known as the family of roots of unity frames for $\mathbb{R}^2$. Our simulations show that the smallest N_0 that satisfy the condition given in Theorem 13 is 17.

Real Harmonic Frames H_N^d for $\mathbb{R}^d$ for $d = 2k+1$ are defined by

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

In this case, H_N^d come from the curve

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

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

2.3 Comparison of Multibit PCM and 1-bit Sigma-Delta

If the amplitude of a signal x is low, then a b-bit PCM does not use all of its dynamic range. For instance, if $\|x\| \leq \delta/2$, then for each frame coefficient, $|\langle x, e_n\rangle| \leq \delta/2$, so $Q_\delta(\langle x, e_n\rangle) = \pm\delta/2$. Therefore, b-bit PCM uses only 1-bit to quantize x. As a result, we have the following result by Theorem 9

Theorem 14. Let $b \geq 2$, $\delta = 2^{1-b}$ and let $x \in \mathbb{C}^d$ satisfy $0 < \|x\| \leq \delta/2$. Let $F = \{e_n\}_{n=1}^N$ be a FUNTF for $\mathbb{C}^d$. Then, the b-bit PCM error satisfies

$$\mathtt{err}_{PCM}(x, F, b) \geq \frac{\delta}{2}(\alpha_F + 1) - \|x\|,$$

where α_F is defined as in (2.5).

Proof. For $0 < \|x\| \leq \delta/2$,

$$\mathtt{err}_{PCM}(x, F, b) = \frac{\delta}{2}\mathtt{err}_{PCM}(\frac{2}{\delta}x, F, 1) \geq \frac{\delta}{2}(\alpha_F + 1 - \|\frac{2}{\delta}x\|) = \frac{\delta}{2}(\alpha_F + 1) - \|x\|.$$

□

As a result of Theorem 14, we have the counterparts of 1-bit comparison theorems, Theorem 11 and Theorem 13, for the multibit case.

Theorem 15. Let $b \geq 2$ and let $\delta = 2^{1-b}$. Let $\{F_N = \{e_n^N\}_{n=1}^N\}$ be a family of FUNTFs for $\mathbb{C}^d$. Then,

$$\forall \varepsilon > 0, \quad \exists N_0 > 0, \qquad \forall N \geq N_0, \quad \mathtt{err}_{\Sigma\Delta}(x, F, 1) \leq \mathtt{err}_{PCM}(x, F, b).$$

for every x, $0 < \|x\| \leq (\delta/2) - \varepsilon$.

Proof. By Theorem 7, $\mathtt{err}_{\Sigma\Delta}(x, F, 1) \leq KN^{-1/2d}$ for some constant K. Given $\varepsilon > 0$, choose $N_0 \geq (K/\varepsilon)^{2d}$. Then, for any $N \geq N_0$, and for every x, $0 < \|x\| \leq (\delta/2) - \varepsilon$,

$$\mathtt{err}_{\Sigma\Delta}(x, F, 1) \leq KN^{-1/2d} \leq \varepsilon \leq \mathtt{err}_{PCM}(\frac{2}{\delta}x, F, 1).$$

□

Theorem 16. Let $e : [0, 1] \to \{x : \|x\| = 1\}$ be continuous function of bounded variation for which $F_N = \{e(n/N)\}_{n=1}^N$ is a FUNTF for $\mathbb{C}^d$ for every N. Then,

$$\exists N_0 > 0 \quad \text{such that} \quad \forall N \geq N_0, \quad \mathtt{err}_{\Sigma\Delta}(x, F, 1) \leq \mathtt{err}_{PCM}(x, F, b)$$

for every x, $0 < \|x\| \leq \delta/2$.

Proof. The proof is essentially the same as the proof of Theorem 13. □

The class of frames that we consider in Theorem 13 and Theorem 16 include the family of Harmonic frames for $\mathbb{R}^d$ (Example 2), and Harmonic frames for $\mathbb{C}^d$. If we choose any d columns of the $N \times N$ DFT matrix, and form a new matrix L using these d columns, then, the rows of $(1/\sqrt{d})L$ constitute a finite unit norm tight frame for $\mathbb{C}^d$. We think that it is important to understand how the multibit PCM quantization error function behaves for this family of frames.

In the remainder of this section, we focus on a family $\{F_N = \{e(n/N)\}_{n=1}^N\}$ of FUNTFs for $\mathbb{C}^d$ coming from a continuous curve e of bounded variation. For any $x \in \mathbb{C}^d$, $\|x\| \leq 1$, the b-bit PCM quantized estimate $\widetilde{x}_b^N$ is given by

$$\widetilde{x}_b^N = \frac{d}{N} \sum_{n=1}^{N} Q_\delta(\langle x, e(\frac{n}{N})\rangle) e(\frac{n}{N}).$$

where $\delta = 2^{1-b}$. Then, $\widetilde{x}_b^N$ is nothing but a Riemann sum of the integral

$$\Phi(x) := d \int_0^1 Q_\delta(\langle x, e(t)\rangle) e(t) dt.$$

The integrand is piecewise constant and it has finitely many jumps since e is of bounded variation. Therefore,

$$\left| \frac{d}{N} \sum_{n=1}^{N} Q_\delta(\langle x, e(\frac{n}{N})\rangle) e(\frac{n}{N}) - d \int_0^1 Q_\delta(\langle x, e(t)\rangle) e(t) dt \right| = \mathcal{O}(\frac{1}{N}), \quad \text{as} \quad N \to \infty. \tag{2.11}$$

We would like to note that $\Phi(x)$ might not be equal to x, for every x. Figure 2.1 and Figure 2.2 depict two such examples. Thus, if $\Phi(x) \neq x$, the PCM quantization error $\mathtt{err}_{\Sigma\Delta}(x, F_N, b)$ does not even converge zero, as $N \to \infty$. Moreover, N^{-1} is the

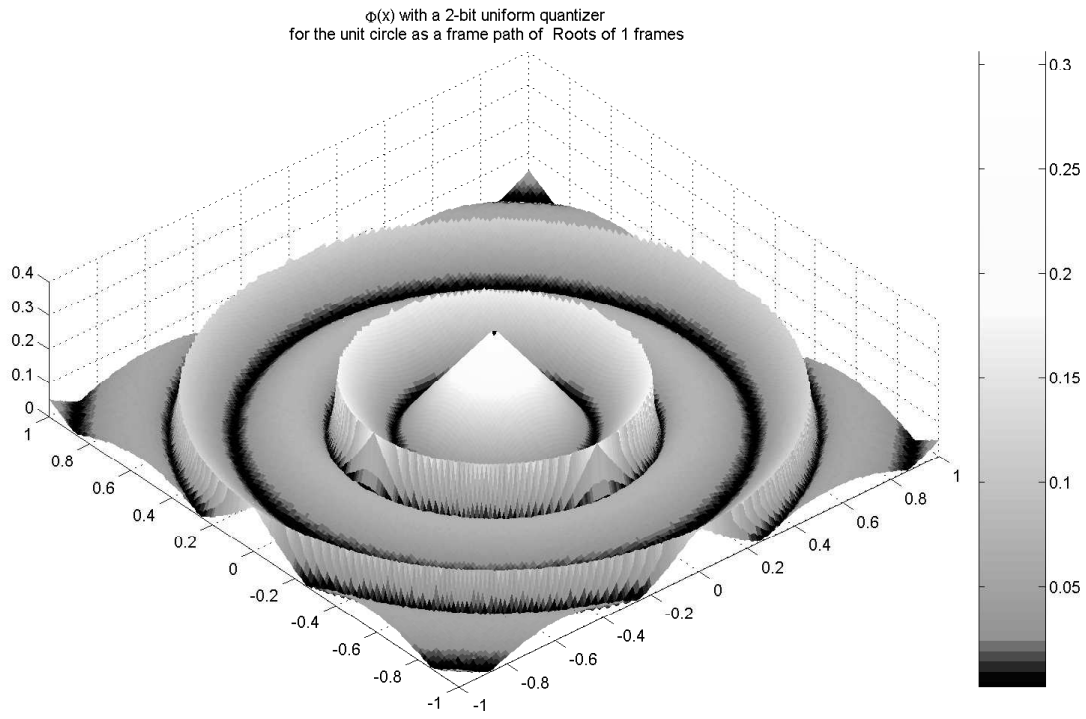

Figure 2.1: The limit Φ of 2-bit PCM quantization error function for the family H_N^2.

best possible error decay rate the quantity in (2.11) when the integrand has jump discontinuities.

By (2.11), 1-bit Sigma-Delta can potentially outperform b-bit PCM at every point in the unit ball of $\mathbb{C}^d$. Since the families of the type have *bounded frame variation*, i.e.,

$$\exists M > 0, \quad \text{such that} \quad \forall N, \quad \sigma(F_N, p) \leq M,$$

1-bit Sigma-Delta error $\mathtt{err}_{\Sigma\Delta}(x, F_N, 1)$ asymptotically decays at least as fast as N^{-1} as $N \to \infty$. In fact, by Theorem 5 and Theorem 6, and the bounded frame

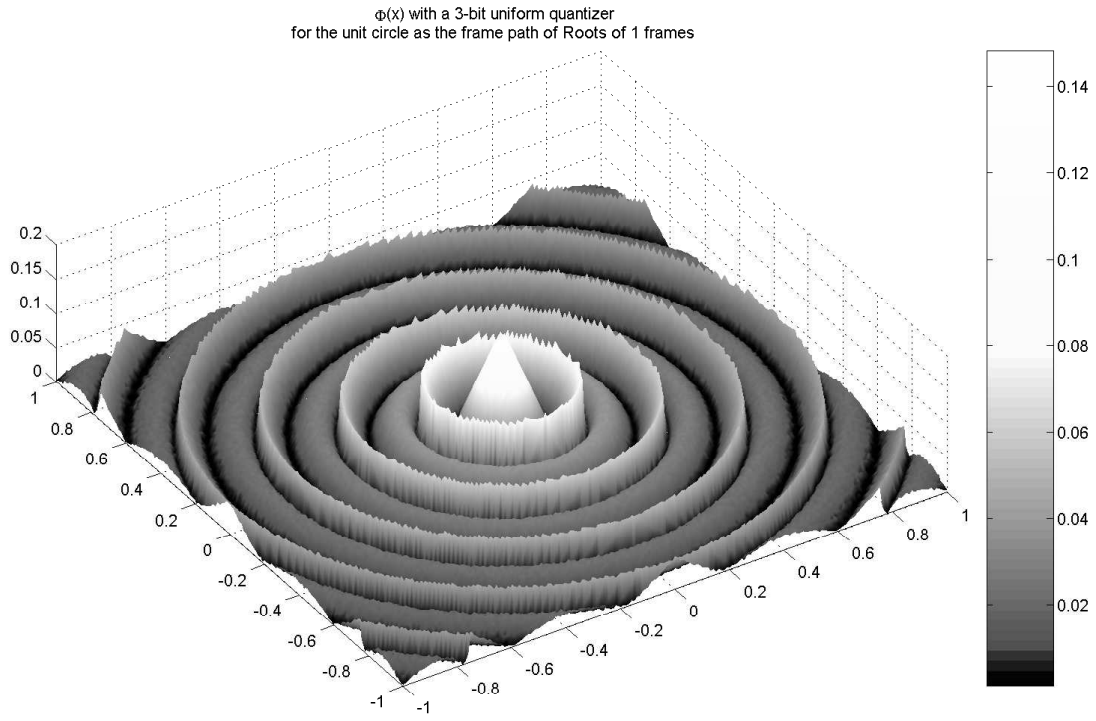

Figure 2.2: The limit Φ of 3-bit PCM quantization error function for the family H_N^2.

variation property, we have

$$\mathtt{err}_{\Sigma\Delta}(x, F_N, 1) \leq \frac{d}{N}(\sigma(F_N, p) + 1) \leq \frac{d}{N}(M + 1).$$

We can always choose $M = |e|_{BV}$.

In [47] Güntürk showed that 1-bit Sigma-Delta error for bandlimited signals can be bounded above by a bound that decays asymptotically in the oversampling rate λ, faster than λ^{-1} by using number theoretical tools. In fact, he proved that for every bandlimited signal x and $\varepsilon > 0$, there is a constant $C_{\varepsilon,x}$ such that the Sigma-Delta error can be uniformly bounded above by $C_{\varepsilon,x}\lambda^{-4/3+\varepsilon}$. Benedetto, Powell and Yilmaz [10] proved that b-bit Sigma-Delta error decays faster than N^{-1} for certain classes of frames. In fact, they proved a more general version of Theorem 17 with additional assumptions. Theorem 17 and a proof can also be found in [10].

Theorem 17. Let d be an even integer and let $\{H_N^d\}$ be the family of real Harmonic frames for $\mathbb{R}^d$. For an $x \in \mathbb{R}^d$, $\|x\| \leq 1$, let $\widetilde{x}_b$ denote the first order b-bit Sigma-Delta estimate. Let $\delta = 2^{1-b}$. Then there is a constant C_x depending on x, such that

$$\|x - \widetilde{x}_b\| \leq C_x \delta \frac{\log N}{N^{5/4}}.$$

These improved error bounds for Sigma-Delta show that, Sigma-Delta error, in fact, is decaying faster than the PCM error for certain families of frames, including $\{H_N^d\}$ for d even. PCM error function for these families of frames is closely related to $\Phi(.)$. Therefore, we investigate the function $\Phi(.)$ more carefully.

Definition 9. $t \in [0, 1]$ is a *quantization crossing of* x if

$$\exists n \in \mathbb{N} \quad \text{such that} \quad \langle x, e(t) \rangle = n\delta.$$

Lemma 3. Let $x \in \mathbb{R}^d$, $\|x\| \leq 1$, and let t^* be a quantization crossing of x. Suppose further that e is differentiable at every point. If $\langle x, e'(t^*)\rangle \neq 0$, then there is a neighborhood W of x and a $\mathcal{C}^1$ function $\tau : W \to [0,1]$ such that

- $\tau(x) = t^*$, and
- $\langle y, e(\tau(y))\rangle = \langle x, e(t^*)\rangle$, $\forall y \in W$.

Proof. Let $G(y,t) = \langle y, e(t)\rangle - \langle x, e(t^*)\rangle$. Then, $G(x,t^*) = 0$, and

$$\frac{\partial G}{\partial t}(x,t^*) = \langle x, e'(t^*)\rangle \neq 0.$$

The result follows by the Implicit Function Theorem. □

Theorem 18. Let $x_0 \in \mathbb{R}^d$, $\|x_0\| \leq 1$, and assume that $\langle x_0, e'(t^*)\rangle \neq 0$ for any quantization crossing t^* of x_0. Moreover, if e is differentiable at every point in a neighborhood of t^*, then $\Phi(.)$ is $\mathcal{C}^1$ around a neighborhood of x_0.

Proof. Let $0 \leq t_1 < t_2 < \cdots < t_r \leq 1$ be distinct quantization crossings of x_0. Then, by Lemma 3, there is a neighborhood W of x_0 and $\mathcal{C}^1$ functions $\tau_j : W \to [0,1]$ such that $\tau_j(x_0) = t_j$ and

$$\langle x, e(\tau_j(x))\rangle = \langle x_0, e(t_j)\rangle, \ \forall j = 1, \ldots r.$$

For notational convenience, we let $\tau_0 \equiv 0$ and $\tau_{r+1} \equiv 1$ on W.

W can be chosen such that

$$Q(\langle x, e(t)\rangle) = \left(n_j\delta + \frac{\delta}{2}\right), \forall t \in [\tau_j(x), \tau_{j+1}(x)),$$

for every $j = 0, \ldots, r$ and some integers n_j. Since e is continuous, n_j and n_{j+1} must be successive integers. Moreover, $n_{j+1} - n_j = \text{sign}(\langle x, e'(t_j)\rangle)$. Then, on W, Φ has the form

$$\Phi(x) = \sum_{j=0}^{r} \left(n_j \delta + \frac{\delta}{2} \right) \int_{\tau_j(x)}^{\tau_{j+1}(x)} e(t) dt.$$

Since τ_j are $\mathcal{C}^1$ on W, so is Φ. In particular,

$$\begin{aligned} \mathcal{D}\Phi(x_0) &= \sum_{j=0}^{r} \left(n_j \delta + \frac{\delta}{2} \right) [e(t_{j+1}) \mathcal{D}\tau_{j+1}(x_0) - e(t_j) \mathcal{D}\tau_j(x_0)] \\ &= \delta \sum_{j=0}^{r} \text{sign}(\langle x, e'(t_j)\rangle)\ e(t_j) \mathcal{D}\tau_j(x_0). \end{aligned}$$

□

t^* is an isolated quantization crossing of x if $\langle x, e'(t^*)\rangle \neq 0$. In particular, if $\langle x, e'(t^*)\rangle \neq 0$ for every quantization crossing t^*, then, x has only finitely many quantization crossings.

In general, if x has only finitely many quantization crossings, $e(.)$ leaves cuts every hyperplane $\{y : \langle x, y\rangle = k\delta\}$ at most at one point, i.e., if $\langle x, e(t^*)\rangle = k\delta$ for some integer k, then there is an $\eta > 0$ such that

i. either $\{e(t^* + t) : t \in (0, \eta)\}$ and $\{e(t^* - t) : t \in (0, \eta)\}$ are separated by the hyperplane $\{y : \langle x, y\rangle = k\delta\}$ (the case $\langle x, e'(t^*)\rangle \neq 0$),

ii. or $e(.)$ is tangent to the hyperplane (the case $\langle x, e'(t^*)\rangle = 0$).

Therefore,

Theorem 19. Let $x \in \mathbb{R}^d$, $\|x\| \leq 1$. If x has only finitely many quantization crossings. Then, Φ is continuous at x.

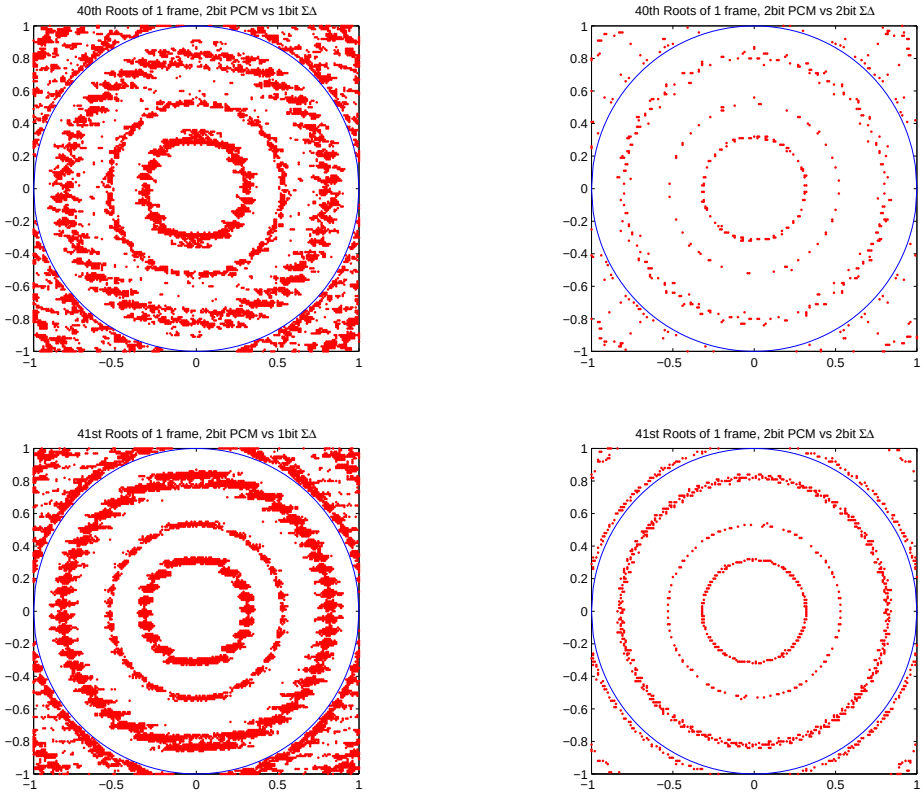

Figure 2.3: 40th and 41st roots of unity frames, 2-bit PCM vs. 1-bit and 2-bit Sigma-Delta. In the white area, the Sigma-Delta quantization error is less than the PCM quantization error.

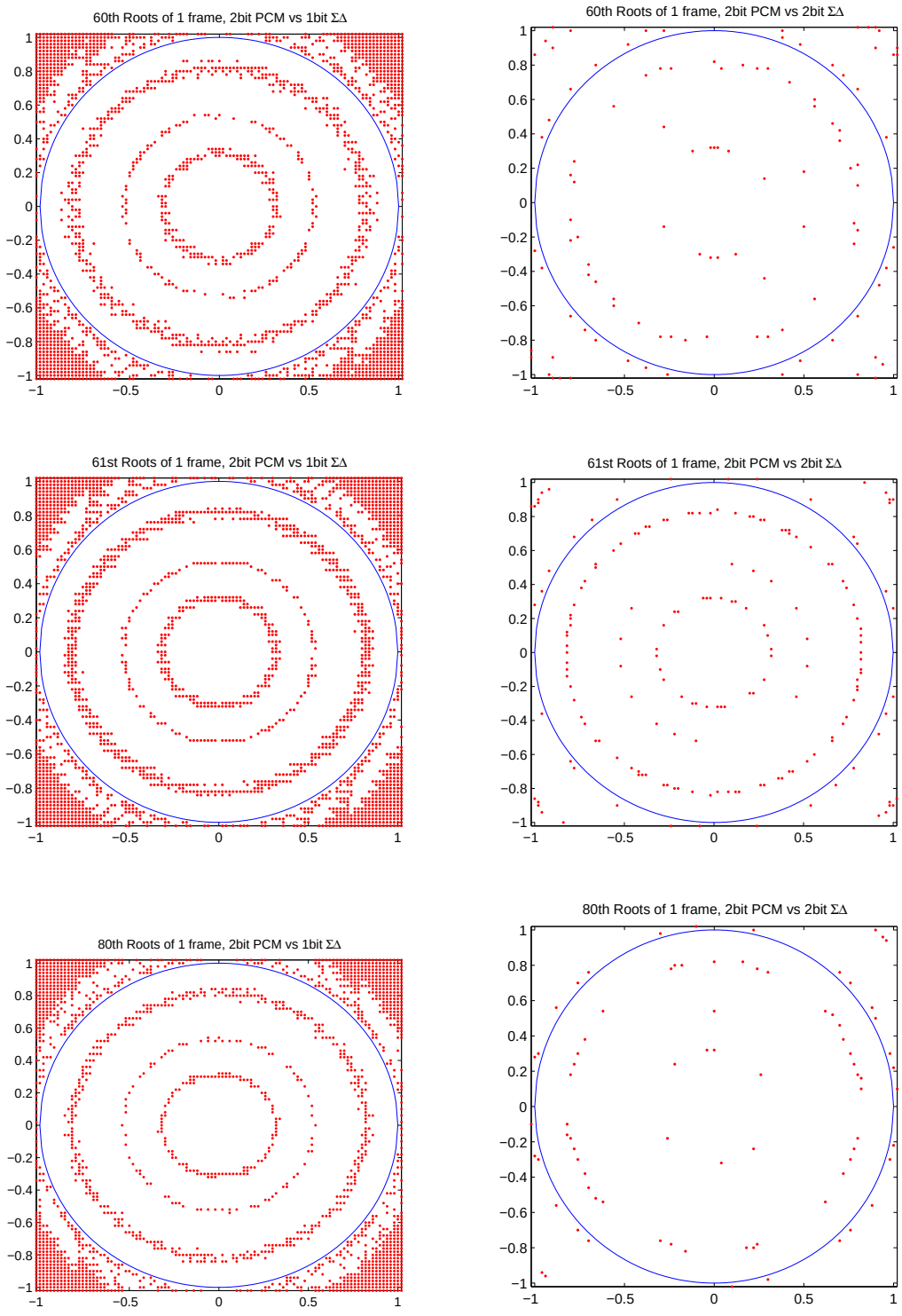

Figure 2.4: 60th, 61st and 80th roots of unity frames, 2-bit PCM vs. 1-bit and 2-bit Sigma-Delta. In the white area, the Sigma-Delta quantization error is less than the PCM quantization error.

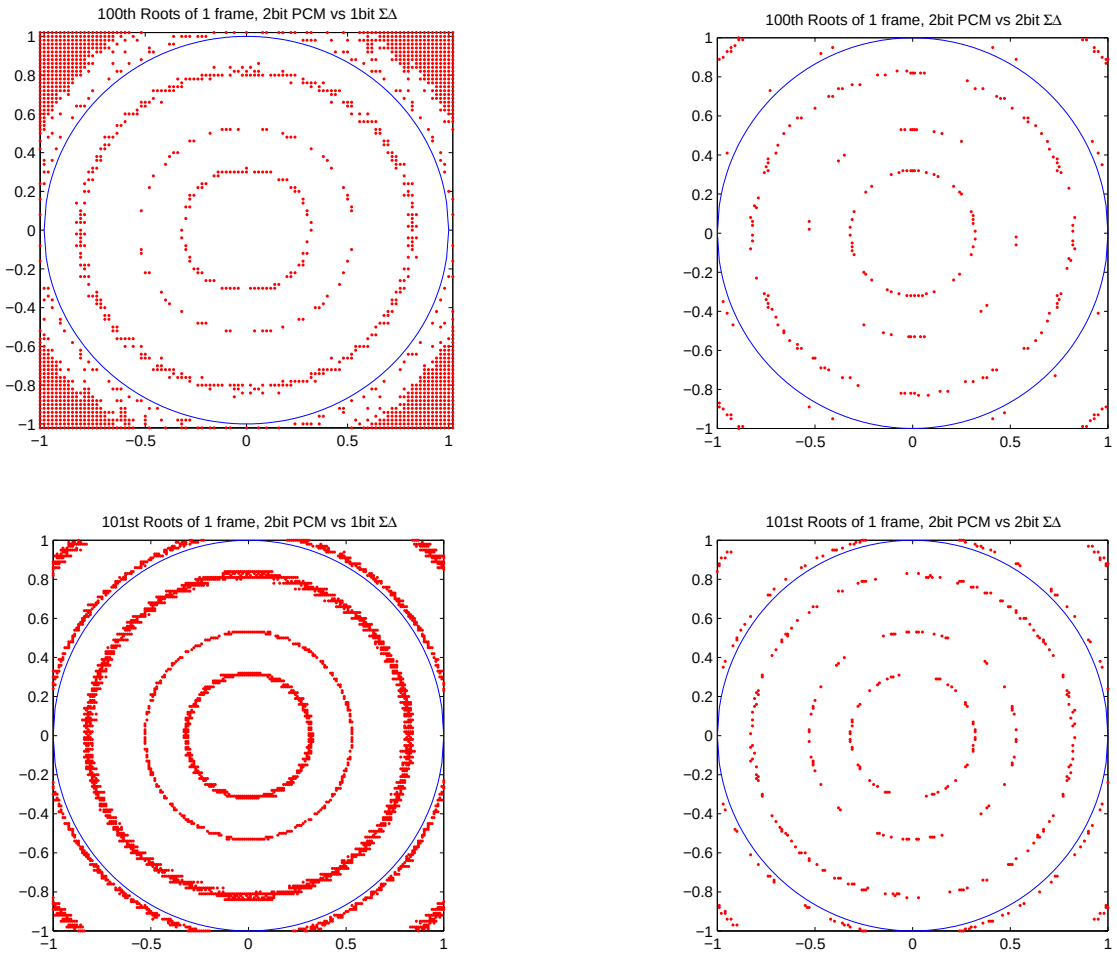

Figure 2.5: 100th and 101st roots of unity frames, 2-bit PCM vs. 1-bit and 2-bit Sigma-Delta. In the white area, the Sigma-Delta quantization error is less than the PCM quantization error.

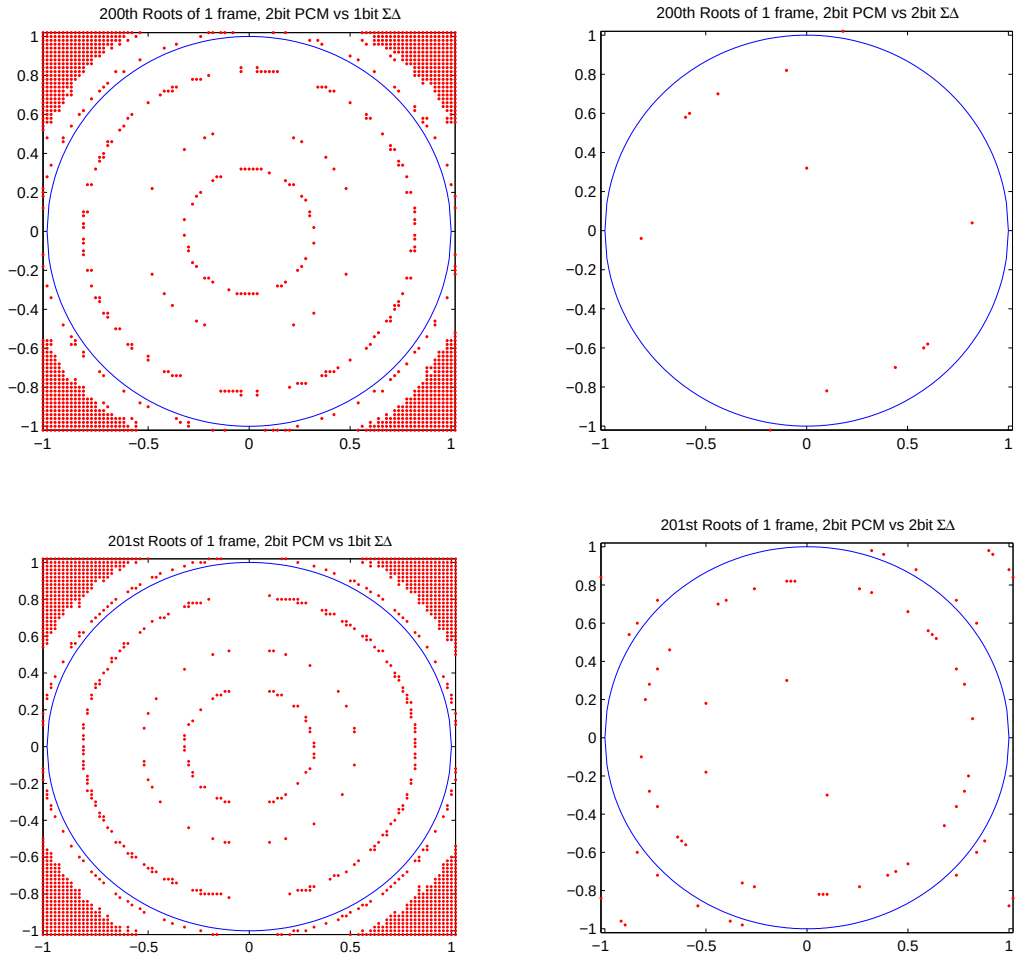

Figure 2.6: 200th and 201st roots of unity frames, 2-bit PCM vs. 1-bit and 2-bit Sigma-Delta. In the white area, the Sigma-Delta quantization error is less than the PCM quantization error.

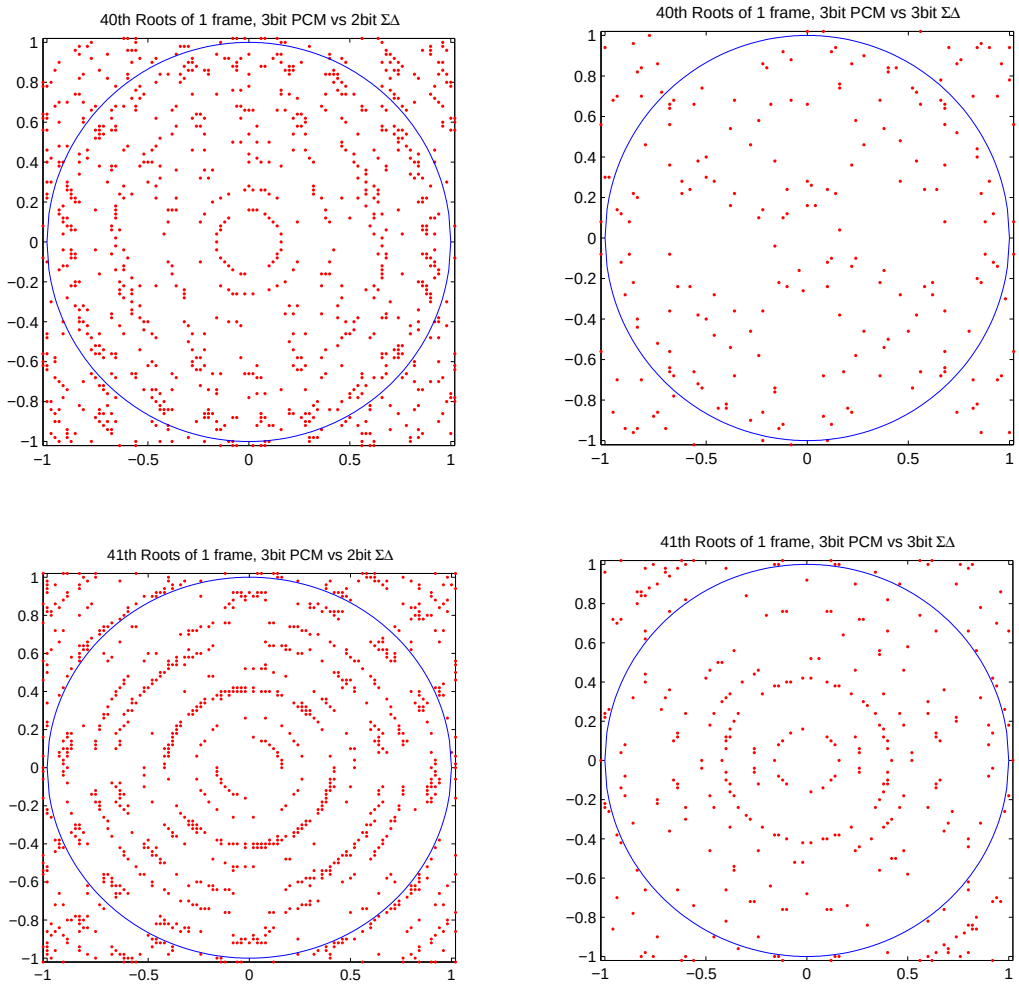

Figure 2.7: 40th and 41st roots of unity frames, 3-bit PCM vs. 2-bit and 3-bit Sigma-Delta. In the white area, the Sigma-Delta quantization error is less than the PCM quantization error.

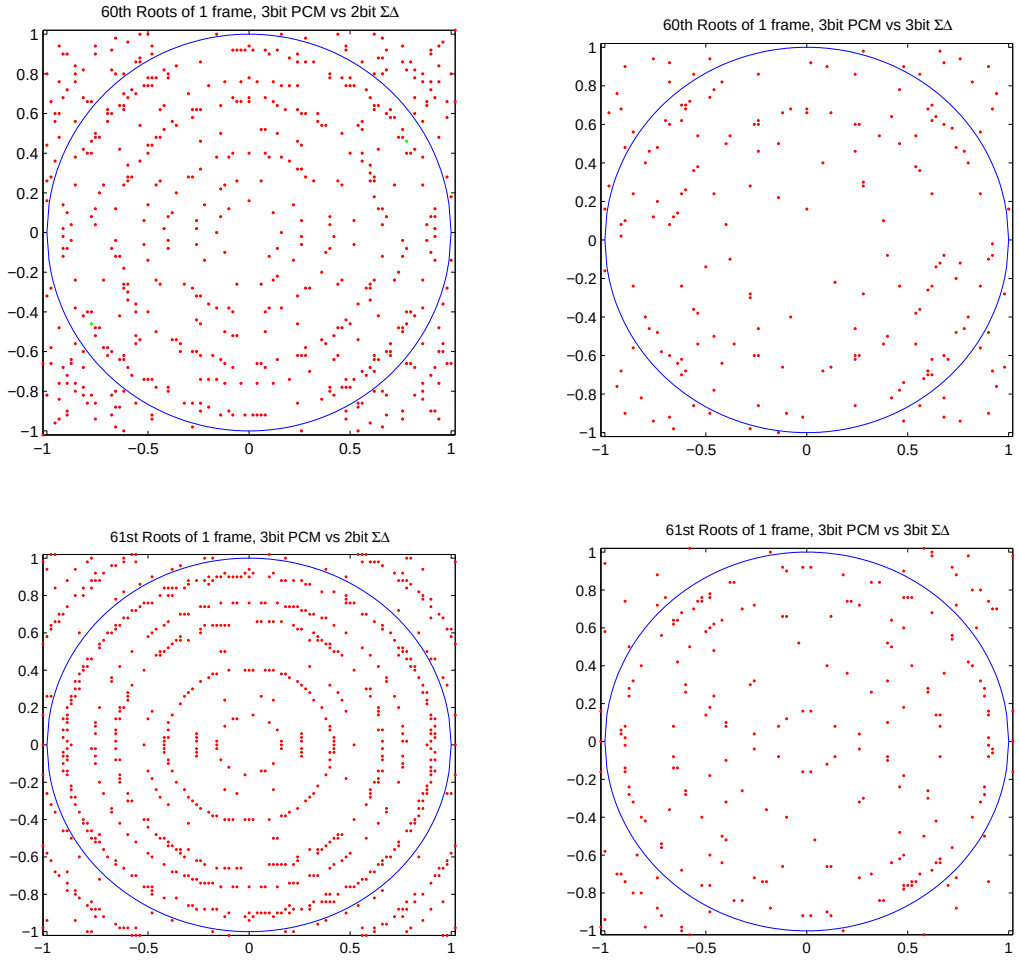

Figure 2.8: 60th and 61st roots of unity frames, 3-bit PCM vs. 2-bit and 3-bit Sigma-Delta. In the white area, the Sigma-Delta quantization error is less than the PCM quantization error.

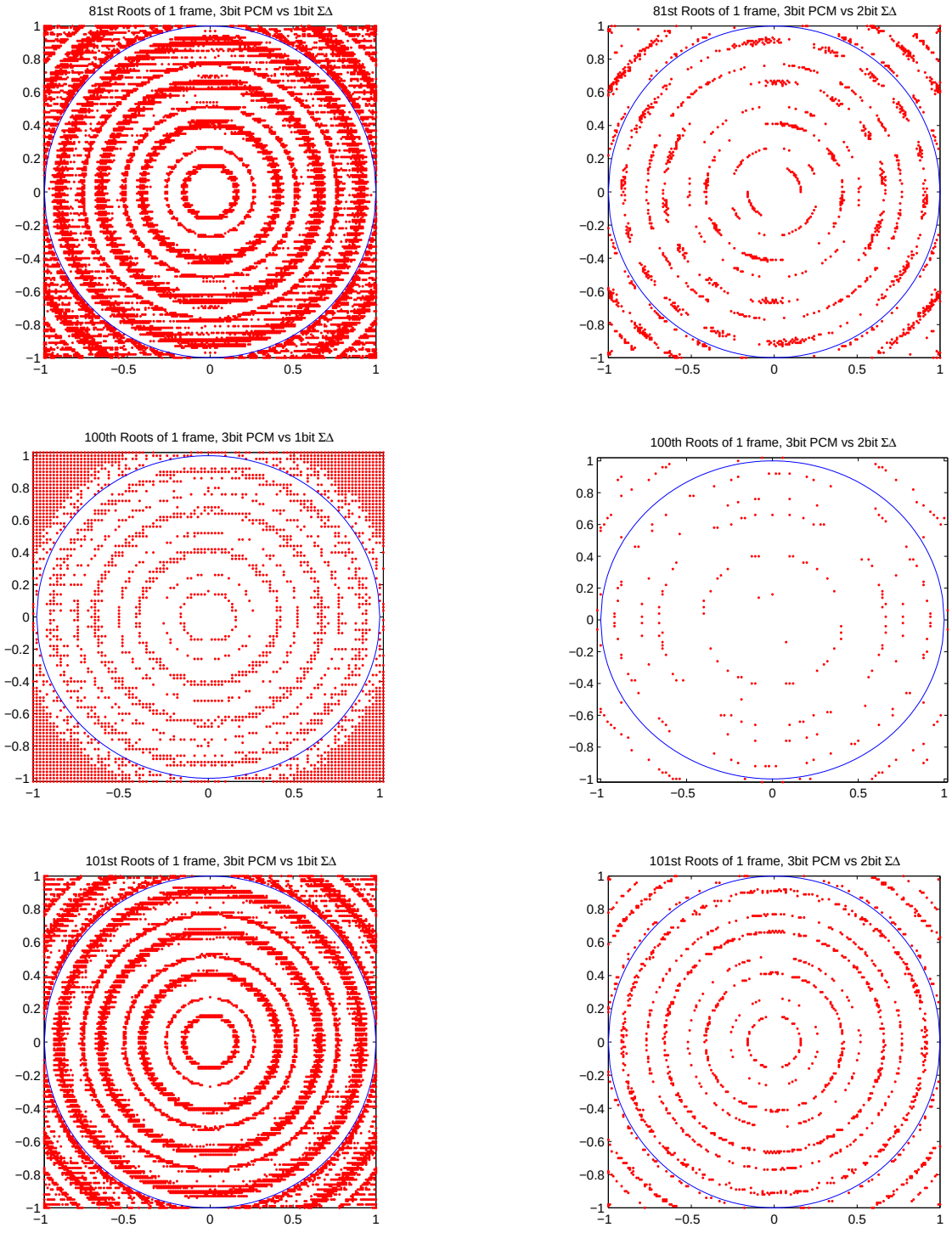

Figure 2.9: 81st, 100th and 101st roots of unity frames, 3-bit PCM vs. 1-bit and 2-bit Sigma-Delta. In the white area, the Sigma-Delta quantization error is less than the PCM quantization error.

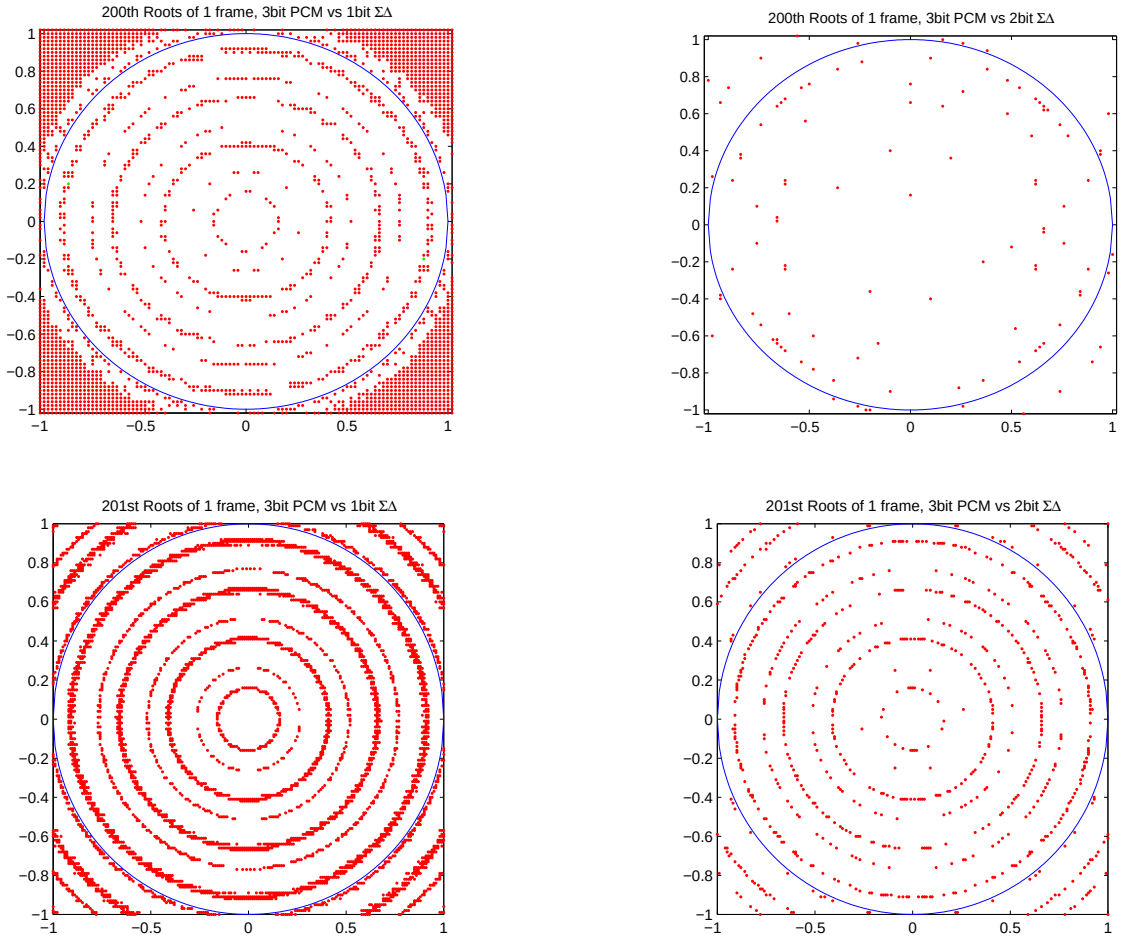

Figure 2.10: 200th and 201st roots of unity frames, 3-bit PCM vs. 1-bit and 2-bit Sigma-Delta. In the white area, the Sigma-Delta quantization error is less than the PCM quantization error.

Chapter 3

New Quantization Techniques

In this chapter, we shall consider the quantization problem in the finite frame setting. In the frame quantization setting, typically, a finite set of numbers, an *alphabet* is specified. The *midrise quantization alphabet* $\mathcal{A}_\delta$ (Definition 5) is an example of an alphabet with equally spaced numbers.

Given an x and a frame $\{e_n\}_{n=1}^N$ with the dual $\{\widetilde{e}_n\}_{n=1}^N$, the finite frame quantization problem is the problem of finding a linear combination of frame elements with coefficients coming from the pre-specified alphabet, which is close to x in some prescribed way. In other words, if ρ is a pre-specified metric, then we want to find q_n in the alphabet such that the quantity

$$\widetilde{x} = \sum_{i=1}^{N} q_n \widetilde{e}_n$$

makes the distance $\rho(x, \widetilde{x})$ sufficiently small.

The geometry of the *coefficient space* and the *signal space* give us a clearer picture of the quantization problem. The main objects in the coefficient space are the range $\mathcal{R}(L)$ of the analysis matrix L, the null space $\mathcal{N}(\widetilde{L}^*)$ of the synthesis matrix $\widetilde{L}^*$ of the dual frame, and the set

$$\mathcal{S} = \{q = (q_n)_{n=1}^N : \ q_n \text{ in the alphabet}\}$$

of *quantized coefficients*. If the numbers in the alphabet are equally spaced, then $\mathcal{S}$ is a rectangular grid. The matrix $L\widetilde{L}^*$ is the orthogonal projection onto $\mathcal{R}(L)$.

Using the definition of frames (Definition 1) we have

$$A\|x - \widetilde{L}^* q\|^2 \leq \|Lx - L\widetilde{L}^* q\|^2 \leq B\|x - \widetilde{L}^* q\|^2, \tag{3.1}$$

where A and B are the lower and upper frame bounds of $\{e_n\}_{n=1}^N$. In particular, when $\{e_n\}_{n=1}^N$ is a finite unit-norm tight frame, $\widetilde{L} = (d/N)L$ and

$$\|x - \frac{d}{N} L^* q\|^2 = \frac{d}{N}\|Lx - \frac{d}{N} LL^* q\|^2. \tag{3.2}$$

Having (3.1), (3.2) and the geometry of the coefficient space in mind, we reformulate the quantization problem as follows: "Given a frame $\{e_n\}_{n=1}^N$ and $x \in \mathbb{R}^d$, find a $q \in \mathcal{S}$ such that the projection of q onto $\mathcal{R}(L)$ is sufficiently close to Lx." We would like to note that

$$\|Lx - L\widetilde{L}^* q\| = \min\{\|q - \xi\| : \ \xi \in Lx + \mathcal{N}(\widetilde{L}^*)\}. \tag{3.3}$$

The main object in the signal space is the set

$$\widetilde{L}^*(\mathcal{S}) = \{\sum_{n=1}^N q_n \widetilde{e}_n : \ q_n \text{ in the alphabet}\}.$$

In particular, if the alphabet is $\mathcal{A}_\delta$, then

$$\widetilde{L}^*(\mathcal{S}) \quad \subseteq \quad \frac{\delta}{2}\widetilde{L}^*(2\mathbb{Z}^N + 1) \quad = \quad \frac{\delta}{2}\sum_{n=1}^N \widetilde{e}_n + \delta\widetilde{L}^*(\mathbb{Z}^N),$$

and $\widetilde{L}^*(\mathbb{Z}^N)$ is an additive subgroup of $\mathbb{R}^d$.

In Section 3.1, we give a general description of a *perfect quantizer*, and give a characterization of a perfect quantizer in this general setting. For the remainder of the chapter, we consider $x \in \mathbb{R}^d$, finite unit-norm tight frames for $\mathbb{R}^d$, Euclidean norm for metric, and the midrise quantization alphabet $\mathcal{A}_\delta$ only.

In Section 3.2 we shall talk about how we can use the geometry of the coefficient space. We shall consider the Sigma-Delta quantization in this context. We present a method to eliminate the *boundary terms* for the second order Sigma-Delta scheme. In Subsection 3.2.3 we shall give a description of the generalized Sigma-Delta schemes. In Section 3.3 we shall talk about how the geometry and the group structure of $\widetilde{L}^*(\mathbb{Z}^N)$ can be exploited. We shall use the generalized Sigma-Delta schemes, and the almost periodic solutions of those schemes in Section 3.3.

In Section 3.4, we shall provide a new 1-bit quantization method that uses minimization techniques. Frame quantization problem is inherently a combinatorial minimization problem. We replace the combinatorial constraint with a penalty term. We show that the solution of this new minimization problem are close to the constraint set $\{q \in \mathbb{R}^N : \ q_n = \pm 1\}$.

3.1 Perfect Quantizer

Throughout this section, (X, d) is a metric space, and $S \subseteq X$ is a finite of X. We call a map $p : X \to S$ a quantizer relative to S. Every quantizer induces an error function, which is defined by

$$\forall y \in X, \quad \mathtt{err}_p(y) = d(y, p(y)).$$

We use the notation $\overline{A}$ to denote the closure of a subset $A \subseteq X$.

Definition 10. Let (X, d) be a metric space, $S \subseteq X$ be a discrete subset. Let $x \in S$. The set of all points $y \in X$ that are closer to x than to any other point in

S, i.e.,

$$\mathcal{C}(x) := \{y \in X : d(y,x) < d(y,x'),\ \forall x' \in S - \{x\}\}$$

is the *Voronoi cell* or *Voronoi region* of x. In this case, x is the *center* of the Voronoi cell $\mathcal{C}(x)$.

Voronoi cells are disjoint, open subsets of X. Their union might not cover all of X since there might be points in X that are on the mutual boundary of two or more Voronoi cells. On the other hand, the union of the closures of all Voronoi cells is equal to X. (see Figure 3.1)

Definition 11. A quantizer $p : X \to S$ is a *perfect quantizer* if it maps every $y \in X$ to the center x of the Voronoi region that it belongs to. If y is on a mutual boundary of two or more Voronoi cells, then p maps y to the center of one of these cells. Equivalently, $p : X \to S$ is a perfect quantizer if it satisfies

$$\forall x \in S, \quad \mathcal{C}(x) \subseteq p^{-1}(\{x\}) \subseteq \overline{\mathcal{C}(x)}, \tag{3.4}$$

where $p^{-1}(\{x\}) = \{y \in X :\ p(y) = x\}$.

A perfect quantizer achieves the minimum possible quantization error, hence the name perfect quantizer.

We can define a perfect quantizer in many equivalent ways, which we summarize in Lemma 4.

Lemma 4. The following assertions are equivalent.

i. $\forall x \in S, \mathcal{C}(x) \subseteq p^{-1}(\{x\}) \subseteq \overline{\mathcal{C}(x)}$,

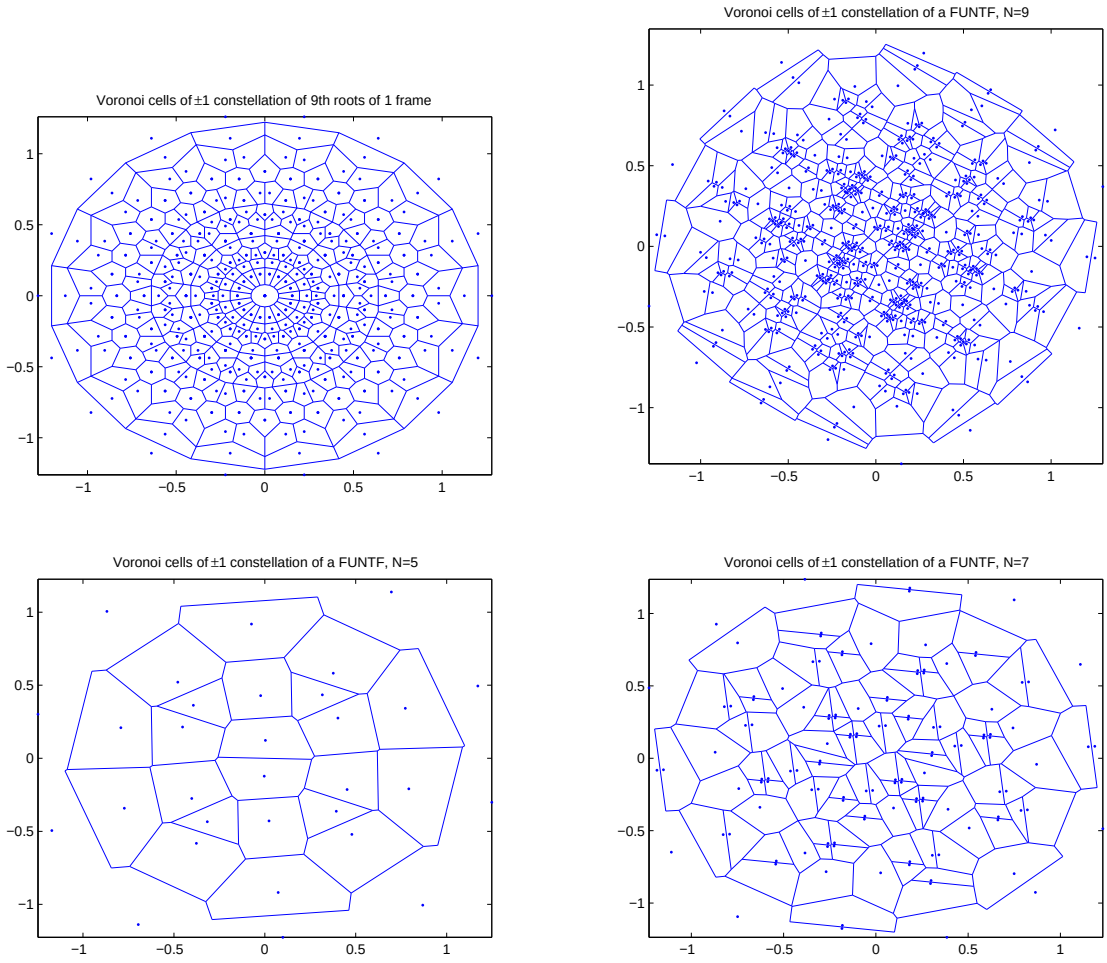

Figure 3.1: Voronoi regions for four tight frame constellation

ii. $\forall x \in S, \mathcal{C}(x) \subseteq \overline{p^{-1}(\{x\})} \subseteq \overline{\mathcal{C}(x)}$,

iii. $\forall x \in S, \mathcal{C}(x) \subseteq p^{-1}(\{x\})$,

iv. $\forall x \in S, \mathcal{C}(x) \subseteq \overline{p^{-1}(\{x\})}$.

A perfect quantizer p satisfies two nice properties. It fixes the elements of S. Also the error function $\mathtt{err}_p(x) = d(x, p(x))$ of a perfect quantizer is continuous in the metric. Theorem 20 shows that the converse is also true, i.e., these two properties are sufficient conditions for a perfect quantizer.

Theorem 20. p is a perfect quantizer if and only if

i. $\forall x \in S, p(x) = x$,

ii. $\mathtt{err}_p(y) = d(y, p(y))$ is continuous in the metric d.

Proof. For the forward implication, assume that p is a perfect quantizer. For every $x \in S$, $p(x) = x$ by Definition 11. Now, let (y_n) be a convergent sequence with $\lim_{n\to\infty} y_n = y$. If $y \in \mathcal{C}(x)$ for some $x \in S$, then for all but finitely many n, y_n is in $\mathcal{C}(x)$ since $\mathcal{C}(x)$ is open. Therefore,

$$\lim_{n\to\infty} \mathtt{err}_p(y_n) = \lim_{n\to\infty} d(y_n, x) = d(y, x) = \lim_{n\to\infty} \mathtt{err}_p(y).$$

If y lies in a mutual boundary of two or more Voronoi cells, say with centers $x_1, \ldots, x_r$, then

$$d(y, x_1) = \cdots = d(y, x_r),$$

and p maps y to one of those points, say $p(y) = x_1$. (We can always renumber

finitely many x_k to make $p(y) = x_1$.) Consider the subsequences

$$(y_n)_{n\in J_k}, \quad J_k = \{n : p(y_n) = x_k\}.$$

Renumber those subsequences so that, with no ambiguity, $(y_n)_{n\in J_k} = (y_n^{(k)})_{n\geq 1}$. It is enough to show that $\lim_{n\to\infty} \mathtt{err}_p(y_n^{(k)}) = \mathtt{err}_p(y)$. But, for any k,

$$\lim_{n\to\infty} \mathtt{err}_p(y_n^{(k)}) = \lim_{n\to\infty} d(y_n^{(k)}, x_k) = d(y, x_k) = d(y, x_1) = \lim_{n\to\infty} \mathtt{err}_p(y).$$

Hence, $\mathtt{err}_p$ is continuous.

For the converse, assume (i) and (ii). $p(x) = x$ by (i), so $p^{-1}(\{x\}) \cap \mathcal{C}(x) \neq \emptyset$. We want to show that $\mathcal{C}(x) \subseteq \overline{p^{-1}(\{x\})}$. For a contradiction, assume not. Then, there exist a $y \in \overline{p^{-1}(\{x\})} \cap \mathcal{C}(x)$, and a sequence (y_n) in the open set $\mathcal{C}(x) \backslash \overline{p^{-1}(\{x\})}$ such that $\lim_{n\to\infty} y_n = y$.

Since $y \in \overline{p^{-1}(\{x\})}$, there exist $x_n \in p^{-1}(\{x\})$ such that $\lim_{n\to\infty} x_n = y$. Since $\mathtt{err}_p$ is continuous by (ii), we have

$$d(y, x) = \lim_{n\to\infty} d(x_n, x) = \lim_{n\to\infty} \mathtt{err}_p(x_n) = \mathtt{err}_p(y) = d(y, p(y)). \tag{3.5}$$

Then, by (3.5), we have

$$\lim_{n\to\infty} \mathtt{err}_p(y_n) = \mathtt{err}_p(y) = d(y, p(y)) = d(y, x). \tag{3.6}$$

For every n, $p(y_n) \neq x$. Then, there is an $x' \in S\backslash\{x\}$, and a subsequence, (y_{n_l}) such that $\lim_{l\to\infty} p(y_{n_l}) = x'$. Then,

$$0 = \lim_{n\to\infty} (\mathtt{err}(y_n) - d(y, x)) = \lim_{l\to\infty} (d(y_{n_l}, p(y_{n_l})) - d(y_{n_l}, x)) = d(y, x') - d(y, x) \tag{3.7}$$

which implies that $d(y, x') = d(y, x)$. However, $y \in \mathcal{C}(x)$, so we must have that $d(y, x) < d(y, x')$. Contradiction.

By contradiction,

$$\forall x \in S, \quad \mathcal{C}(x) \subseteq \overline{p^{-1}(\{x\})}.$$

Hence the result follows by Lemma 4. □

3.2 Sparse Matrices and Periodic Solutions

In this section, we shall consider the geometry of the coefficient space in the frame quantization setting. With (3.3) in mind, given $x \in \mathbb{R}^d$ and a FUNTF $\{x_n\}_{n=1}^N$ with analysis matrix L, we would like to find a $q = (q_n)$ such that $q - Lx$ is sufficiently close to $\mathcal{N}(L^*)$.

One approach is to find a basis, or more generally a spanning set for $\mathcal{N}(L^*)$. Given a spanning set $\{b_1, \ldots, b_r\}$, we form a matrix B, whose $k-th$ column is b_k. Then, $L^*B \equiv 0$, and for any $u \in \mathbb{R}^r$, we have

$$\|q - Lx - Bu\|^2 \geq \frac{d}{N}\|LL^*(q - Lx - Bu)\|^2 = \frac{d}{N}\|L^*(Lx + Bu - q)\|^2 = \|x - \frac{d}{N}L^*q\|^2, \tag{3.8}$$

since $(d/N)LL^*$ is an orthogonal projection. Therefore, one might want to find a u and a quantized sequence q such that $\|q - Lx - Bu\|$ is smaller than a prescribed tolerance.

For a fast and memory efficient numerical algorithm, one might want to choose a sparse spanning set for $\mathcal{N}(L^*)$. Since $\{x_n\}_{n=1}^N$ is a FUNTF for $x \in \mathbb{R}^d$, any $d+1$ element subset of $\{x_n\}_{n=1}^N$ is linearly dependent. Therefore, we can choose $b_l \in \mathbb{R}^N$ such that

$$x_k + \sum_{l=1}^{d} b_l(k) x_{k+l} = 0, \quad \forall k = 1, \ldots, N-d,$$

and $b_l(k) = 0$ otherwise. Then, $\{b_1, \ldots, b_{N-d}\}$ gives a basis for $\mathcal{N}(L^*)$. Furthermore, each b_l has at most $d+1$ nonzero entries.

However, one might want to use sparser vectors. Also we might want to impose certain restrictions on the entries of b_l, for instance, for numerical stability. In this case, we might want to *approximate* $\mathcal{N}(L^*)$ with a sparse set of vectors. By "approximating $\mathcal{N}(L^*)$", we mean finding a sparse set of vectors $\{b_1, \ldots, b_r\}$ such that the span of those vectors is close to $\mathcal{N}(L^*)$ in the sense that the quantity

$$\|L^*B\|_{a,b} = \sup_{\|u\|_b=1} \|L^*Bu\|_a$$

is small, where $\|.\|_a$ and $\|.\|_b$ are two norms on $\mathbb{R}^d$ and $\mathbb{R}^r$, respectively. (This distance, in fact, is ambiguous, because it depends on the choice of the basis. $\|L^*B\|_{a,b}/\|B\|_{b,r}$ is a better distance measure, however for the practical purposes in this section, we use $\|L^*B\|_{a,b}$.) We shall show that $\|L^*B\|_{a,b}$ is closely related to the *frame variation* [10, 9] in the coming subsections.

Similar to (3.8), for any $u \in \mathbb{R}^r$, we have

$$\begin{aligned} \|x - \frac{d}{N}L^*q\|_a &= \frac{d}{N}\|L^*(Lx-q)\|_a \\ &\leq \frac{d}{N}\|L^*(Lx+Bu-q)\|_a + \frac{d}{N}\|L^*Bu\|_a \\ &\leq C\|q-Lx-Bu\|_a + \frac{d}{N}\|L^*B\|_{a,b}\|u\|_b \end{aligned} \tag{3.9}$$

where $C > 0$ is a constant depending on $\|.\|_a$ and L. We can choose $C = \sqrt{d/N}$ for the usual Euclidean norm.

Notation 1. We intend to use the notation $\|.\|_a$ and $\|.\|_b$ for arbitrary norms defined on an m-dimensional Euclidean space $\mathbb{R}^m$. However, we reserve the notation $\|.\|_p$ to

denote the p-norm ($1 \leq p < \infty$)

$$\forall v \in \mathbb{R}^m, \quad \|v\|_p = \left(\sum_{k=1}^{m} |v(k)|^p \right)^{1/p},$$

and the notation $\|.\|_\infty$ to denote the infinity-norm

$$\forall v \in \mathbb{R}^m, \quad \|v\|_\infty = \max\{|v(k)| : k = 1, \ldots, m\}.$$

When $p = 2$, we drop the subscript.

In the remainder of this section, for notational convenience, we sometimes index frames with $\mathbb{Z}_N$. In this case, without mentioning, we view every $v \in \mathbb{R}^m$ as a real valued function $v : \mathbb{Z}_m \to \mathbb{R}$, i.e.,

$$\forall t \in \mathbb{Z}, \quad \forall k = 1, \ldots, m \quad v(tm + k) = v(k).$$

3.2.1 First Order Sigma-Delta Scheme

Let $x \in \mathbb{R}^d$, $\{e_n\}_{n=1}^N$ a FUNTF for $\mathbb{R}^d$. First order Sigma-Delta scheme for finite frames is defined by the iteration

$$\begin{aligned} q(n) &= Q_\delta(u(n-1) + Lx(n)) \\ u(n) &= u(n-1) + Lx(n) - q(n) \end{aligned} \tag{3.10}$$

for $n = 1, \ldots, N$, with the initial condition $u(0)$, and the input sequence Lx. Q_δ is the uniform quantizer with step size δ, defined in Definition 5.

(3.10) gives rise to the matrix equation $q = Lx - Bu + \eta$, where $u \in \mathbb{R}^{N-1}$ and

B is the $N \times (N-1)$ matrix

$$B = \begin{pmatrix} 1 & & & \\ -1 & 1 & & \\ & -1 & \ddots & \\ & & \ddots & 1 \\ & & & -1 \end{pmatrix}, \quad \eta = \begin{pmatrix} u(0) \\ 0 \\ \vdots \\ 0 \\ u(N) \end{pmatrix}$$

i.e., for every k, $b(k,k) = 1$, $b(k+1,k) = -1$, and all the other entries of B are equal to zero. In other words, B is defined by

$$\forall n = 2, \ldots, N-1, \quad \forall u \in \mathbb{R}^{N-1}, \quad (Bu)(n) = u(n) - u(n-1),$$

and $(Bu)(1) = u(1)$, $(Bu)(N) = -u(N-1)$.

The columns of the matrix B spans an $N-1$ dimensional subspace of $\mathbb{R}^N$. If we use the regular Euclidean norm for $\|.\|_a$, and $\|.\|_\infty$ for $\|.\|_b$, then the quantity $\|L^*B\|_{a,b}$ is less than or equal to the frame variation [10]

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

for the identity permutation p, $p(k) = k$. We prove this result in Lemma 6 using Lemma 5.

Lemma 5. Let A be an $M \times r$ matrix with columns $\{a_1, \ldots, a_r\}$. Then,

$$\|A\|_{p,\infty} = \sup_{\|u\|_\infty = 1} \|Au\|_p \leq \sum_{k=1}^{r} \|a_k\|_p.$$

Proof. $Au = \sum u(k) a_k$. Therefore, it is it is enough to show

$$\|Au\|_p = \|\sum_{k=1}^{r} u(k) a_k\|_p \leq \sum_{k=1}^{r} |u(k)| \ \|a_k\|_p \leq \|u\|_\infty \sum_{k=1}^{r} \|a_k\|_p.$$

□

Lemma 6. Let $\|.\|_a = \|.\|$, $\|.\|_b = \|.\|_\infty$, and let the matrix B be as above. For a permutation p, let L be the analysis matrix of the permuted frame $\{e_{p(n)}\}_{n=1}^N$. Then,

$$\|L^*B\|_{2,\infty} \leq \sigma(\{e_n\}_{n=1}^N, p) = \sum_{n=1}^{N-1} \|e_{p(n)} - e_{p(n+1)}\|.$$

Proof. The nth column of L^*B is $e_{p(n)} - e_{p(n+1)}$. Therefore, the inequality is a direct result of Lemma 5. □

Benedetto, Powell and Yilmaz [10] proved Theorem 21. We shall give an alternative proof using (3.9).

Theorem 21. Let $\{e_n\}_{n=1}^N$ be a FUNTF for $\mathbb{R}^d$, let p be a permutation of $\{1, \ldots, N\}$, let $|u(0)| \leq \delta/2$, and let $x \in \mathbb{R}^d$ satisfy $\|x\| \leq 1$. Let $\widetilde{x}$ denote the 1st order Sigma-Delta estimate of x. Then,

$$\|x - \widetilde{x}\| \leq \frac{d}{N}\Big(\frac{\delta}{2}\sigma(\{e_n\}_{n=1}^N, p) + |u(0)| + |u(N)|\Big).$$

Proof. By (3.10) if $|u(0)| \leq \delta/2$, then it is not hard to show that

$$|u(n)| \leq |(u(n-1) - Lx(n)) - Q_\delta(u(n-1) - Lx(n))| \leq \delta/2.$$

Let p be a permutation, and let L denote the analysis matrix of the frame $\{e_{p(n)}\}_{n=1}^N$. Then

$$\|L^*\eta\| = \|u(0)e_{p(1)} - u(N)e_{p(N)}\| \leq |u(0)| + |u(N)|.$$

Therefore, by (3.9), we have

$$\|x - \widetilde{x}\| \leq \frac{d}{N}\|L^*\eta\| + \frac{d}{N}\|L^*B\|_{2,\infty}\|u\|_\infty \leq \frac{d}{N}\left(\frac{\delta}{2}\|L^*B\|_{2,\infty} + |u(0)| + |u(N)|\right).$$

The result follows using Lemma 6. □

We could have chosen a slightly different B, and an η accordingly. Namely, if

$$B = \begin{pmatrix} 1 & & & & -1 \\ -1 & 1 & & & \\ & -1 & \ddots & & \\ & & \ddots & 1 & \\ & & & -1 & 1 \end{pmatrix}, \quad \eta = \begin{pmatrix} u(0) - u(N) \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix},$$

then we would have a slightly different version of Theorem 21. In this case, we have the inequality

$$\|x - \widetilde{x}\| \leq \frac{d}{N}\Big(\frac{\delta}{2} \sum_{n \in \mathbb{Z}_N} \|e_{p(n)} - e_{p(n+1)}\| + |u(0) - u(N)|\Big). \tag{3.11}$$

The proof of (3.11) is very similar to the proof of Theorem 21, so we shall not provide a separate proof.

3.2.2 Second Order Sigma-Delta Scheme

Let $x \in \mathbb{R}^d$, $\{e_n\}_{n=1}^N$ a FUNTF for $\mathbb{R}^d$. Second order Sigma-Delta scheme for finite frames is defined by the iteration

$$\begin{aligned} q(n) &= Q_\delta(2u(n-1) - u(n-2) + Lx(n)) \\ u(n) &= 2u(n-1) - u(n-2) + Lx(n) - q(n) \end{aligned} \tag{3.12}$$

for $n = 1, \ldots, N$, with the initial conditions $u(-1)$ and $u(0)$, and the input sequence Lx.

(3.12) gives rise to a matrix equation of the form $q = Lx - Bu + \eta$ for the following choices of B and η.

$$B = \begin{pmatrix} 1 & & & & \\ -2 & 1 & & & \\ 1 & -2 & \ddots & & \\ & 1 & \ddots & 1 & \\ & & \ddots & -2 & \\ & & & 1 & \end{pmatrix}, \quad \eta = \begin{pmatrix} u(-1) - 2u(0) \\ u(0) \\ 0 \\ \vdots \\ 0 \\ u(N-1) \\ u(N) - 2u(N-1) \end{pmatrix}$$

$$B = \begin{pmatrix} 1 & & & & & 1 & -2 \\ -2 & 1 & & & & & 1 \\ 1 & -2 & \ddots & & & & \\ & 1 & \ddots & 1 & & & \\ & & \ddots & -2 & 1 & & \\ & & & 1 & -2 & 1 & \end{pmatrix}, \quad \eta = \begin{pmatrix} (u(-1) - u(N-1)) - 2(u(0) - u(N)) \\ u(0) - u(N) \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}$$

We shall focus on the second choice. B is a cyclical convolution matrix, which can also be defined as

$$\forall u \in \mathbb{R}^d, \quad \forall n \in \mathbb{Z}_N, \quad (Bu)(n) = u(n) - 2u(n-1) + u(n-2).$$

The quantity $\|L^* B\|_{2,\infty}$ is closely related to the *second frame variation* [9]. We establish this relation in Lemma 7.

Lemma 7. Let $\|.\|_a = \|.\|$, $\|.\|_b = \|.\|_\infty$, and let the matrix B be as above. For a permutation p, let L be the analysis matrix of the permuted frame $\{e_{p(n)}\}_{n=1}^N$. Then,

$$\|L^* B\|_{2,\infty} \leq \sum_{n \in \mathbb{Z}_N} \|e_{p(n)} - 2e_{p(n+1)} + e_{p(n+2)}\|.$$

Proof. The nth column of L^*B is $e_{p(n)} - 2e_{p(n+1)} + e_{p(n+2)}$. Therefore, the inequality is a direct result of Lemma 5. □

If we used the first choice for B in this subsection, then we would have

$$\|L^*B\|_{2,\infty} \leq \sum_{n=1}^{N-2} \|e_{p(n)} - 2e_{p(n+1)} + e_{p(n+2)}\| = \sigma_2(\{e_n\}_{n=1}^N, p),$$

where $\sigma_2(\{e_n\}_{n=1}^N, p)$ is the second frame variation.

Benedetto, Powell and Yilmaz [9] proved the following upper bound for the 1-bit second order Sigma-Delta scheme, with a slight change in notation:

$$\|x-\widetilde{x}\| \leq \frac{d}{N}\Big(\|u\|_\infty \sigma_2(\{e_n\}_{n=1}^N, p) + |u(N-1)|\ \|e_{p(N-1)} - e_{p(N)}\| + |u(N) - u(N-1)|\Big). \tag{3.13}$$

A slightly different version of (3.13) is in Theorem 22.

Theorem 22. Let $\{e_n\}_{n=1}^N$ be a FUNTF for $\mathbb{R}^d$, let p be a permutation of $\{1, \ldots, N\}$, and let $x \in \mathbb{R}^d$ satisfy $\|x\| \leq 1$. Let $\widetilde{x}$ denote the 1-bit second order Sigma-Delta estimate of x. Then,

$$\begin{aligned} \|x - \widetilde{x}\| \ &\leq\ \frac{d}{N}\Big(\|u\|_\infty \sum_{n\in\mathbb{Z}_N} \|e_{p(n)} - 2e_{p(n+1)} + e_{p(n+2)}\| + |u(N) - u(0)|\ \|e_{p(1)} - e_{p(2)}\| \\ &+\ |u(N) - u(N-1) - u(0) + u(-1)|\Big). \end{aligned} \tag{3.14}$$

Proof. Let p be a permutation, and let L denote the analysis matrix of the frame $\{e_{p(n)}\}_{n=1}^N$. Then

$$\|L^*\eta\| = \|(u(N) - u(0))(e_{p(1)} - e_{p(2)}) + (u(N-1) - u(-1) - u(N) + u(0))e_{p(1)}\|.$$

Therefore, by (3.9), we have

$$\begin{aligned} \|x-\widetilde{x}\| &\leq \frac{d}{N}\|L^*\eta\| + \frac{d}{N}\|L^*B\|_{2,\infty}\|u\|_\infty \\ &\leq \frac{d}{N}\Big(\|L^*B\|_{2,\infty}\|u\|_\infty + |u(N)-u(0)|\ \|e_{p(1)}-e_{p(2)}\| \\ &\quad +|u(N)-u(N-1)-u(0)+u(-1)|\Big). \end{aligned} \tag{3.15}$$

The result follows using Lemma 7. □

In certain cases, for example, for the family of Harmonic frames F_N, the second frame variation satisfies

$$\sigma_2(F_N,p) \leq \frac{C}{N},$$

for some constant $C > 0$ [9]. However, since we have extra terms in (3.15), the upper bound has a decay rate of N^{-1}. If we could eliminate these extra terms, then we could have an error decay rate of N^{-2}.

Having (3.12) in hand, we have

$$\begin{pmatrix} 1 & 1 & \dots & 1 \\ N & N-1 & \dots & 1 \end{pmatrix} (Lx-q) = \begin{pmatrix} 1 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} u(N)-u(0) \\ u(N-1)-u(-1) \end{pmatrix}. \tag{3.16}$$

Conversely, if Lx and q, $u(N), u(N-1), u(0)$ and $u(-1)$ were given that satisfies (3.15), then we could set

$$u(n) = u(0) + (N-n)(u(-1)-u(0)) + \sum_{k=1}^{n}(n-k+1)(Lx(k)-q(k)), \tag{3.17}$$

for $n = 1, \dots, N-2$, and this u would satisfy the first equation in (3.12).

The extra terms in (3.15) vanish if and only if the right hand side of (3.16) is zero.

The quantity $\sum_{n=1}^{N} q(n)$ is an integer, and all of the values that it can get lie in the set

$$\{-N + 2k : \ k = 0, \ldots, N\}. \tag{3.18}$$

Therefore, the best possible value that $\sum_{n=1}^{N} q(n)$ can get is the integer $\alpha(x)$ in this set that is closest to $\sum_{n=1}^{N} \langle x, e_n \rangle$. Given a q, we can always generate a $\widetilde{q}$ by switching the signs of some of the entries of q such that $\alpha(x) = \sum_{n=1}^{N} \widetilde{q}(n)$.

The quantity $\sum_{n=1}^{N} (N - n + 1) q(n)$ is also an integer. Moreover, all of the values that this quantity can get lie in the set

$$\{2k - N(N+1)/2 : \ k = 0, \ldots, N(N+1)\}. \tag{3.19}$$

Given a q that satisfies $\alpha(x) = \sum_{n=1}^{N} q(n)$, switching the signs of two entries with opposite signs leaves the value of $\sum_{n=1}^{N} q(n)$ intact. For instance, if $q(k) = -1$ and $q(k+1) = 1$ are two successive entries, then switching the signs of these two entries does not affect the value of $\sum_{n=1}^{N} q(n)$. However, same operation increases the value of $\sum_{n=1}^{N} (N - n + 1) q(n)$ by 2. Therefore, if $\beta(x)$ is the closest integer in the set in (3.19) to the quantity $\sum_{n=1}^{N} (N - n + 1) \langle x, e_n \rangle$, we can find a q that simultaneously satisfies

$$\alpha(x) \ = \ \sum_{n=1}^{N} q(n) \tag{3.20}$$

$$\beta(x) \ = \ \sum_{n=1}^{N} (N - n + 1) q(n). \tag{3.21}$$

Having this q in hand, (3.17) gives us a u that satisfies

$$\begin{pmatrix} 1 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} u(N)-u(0) \\ u(N-1)-u(-1) \end{pmatrix} = \begin{pmatrix} \sum_{n=1}^{N}\langle x, e_n\rangle - \alpha(x) \\ \sum_{n=1}^{N}(N-n+1)\langle x, e_n\rangle - \beta(x) \end{pmatrix}. \tag{3.22}$$

In particular,

$$|u(N)-u(0)-u(N-1)+u(-1)| = |\sum_{n=1}^{N}\langle x, e_n\rangle - \alpha(x)|.$$

With (3.20) and (3.22) in hand, we have a direct corollary of Theorem 22.

Theorem 23. Let N be an even integer, and let $\{e_n\}_{n=1}^{N}$ be a zero sum FUNTF. With u and q given by (3.20) and (3.22), we have the upper bound

$$\|x-\widetilde{x}\| \leq \frac{d}{N}\Big(\|u\|_\infty \sum_{n\in\mathbb{Z}_N} \|e_n - 2e_{n+1} + e_{n+2}\| + \|e_{p(1)} - e_{p(2)}\|.\Big)$$

3.2.3 Generalized Sigma-Delta Schemes

All of the Sigma-Delta schemes can be expressed as a convolution equation

$$\begin{aligned} q(n) &= Q_\delta((h*u)(n) + Lx(n)) \\ u(n) &= (h*u)(n) + Lx(n) - q(n) \end{aligned} \tag{3.23}$$

for $n \geq 1$, where $h \in \mathbb{R}^N$. For instance, if $h(1) = 1$ and $h(k) = 0$ otherwise, we obtain the first order Sigma-Delta scheme. If $h(1) = 2$, $h(2) = -1$ and $h(k) = 0$ otherwise, we obtain the second order Sigma-Delta scheme. If we choose

$$h(k) = (-1)^{r-k+1}\frac{r!}{k!\ (r-k)!}, \quad k = 1, \ldots, r,$$

and $h(k) = 0$ otherwise, we obtain the rth order Sigma-Delta scheme [28, 46, 9].

In the bandlimited setting, Güntürk constructed a family of finitely supported filters $(h^{(n)})$, and showed that the limit of the generalized Sigma-Delta error upper bounds decrease exponentially in the oversampling rate, as the oversampling ratio tends to infinity [46].

We consider (3.23) for the finite frame setting with the input sequence Lx. For this setting, we choose $h \in \mathbb{R}^N$ such that $h(r+1), \ldots, h(N) = 0$ for some index $r < N$. In this case, we need r initial conditions $u(0), u(-1), \ldots, u(-r+1)$ in order to be able to define (3.23) for every $1 \leq n \leq N$.

A natural choice for B is

$$B = \begin{pmatrix} 1 & & & h(r) & \ldots & h(1) \\ h(1) & \ddots & & & \ddots & \vdots \\ \vdots & \ddots & 1 & & & h(r) \\ h(r) & & h(1) & 1 & & \\ & \ddots & \vdots & \vdots & \ddots & \\ & & h(r) & h(r-1) & \ldots & 1 \end{pmatrix},$$

i.e., B is defined by

$$\forall w \in \mathbb{R}^N, \quad \forall n \in \mathbb{Z}_N \quad (Bw)(n) = w(n) + (h * w)(n) = w(n) + \sum_{k \in \mathbb{Z}_N} h(n-k)w(k).$$

Accordingly, given u, we define

$$\forall n = 1, \ldots, r \quad \eta(n) = \sum_{k=0}^{r-n} h(n+k)(u(-k) - u(N-k)),$$

and $\eta(n) = 0$ otherwise.

A further generalization to the system (3.23) is given in

$$\begin{aligned} q(n) &= Q_\delta(\sum_{k=n-d}^{n-1} b(n,k)u(k) + Lx(n)) \\ u(n) &= \sum_{k=n-d}^{n-1} b(n,k)u(k) + Lx(n) - q(n) \end{aligned} \tag{3.24}$$

for $n \geq 1$, where $b : \mathbb{Z}_N^2 \to \mathbb{R}$ function such that $b(n,n) = 1$, and $b(n,k) = 0$ if $1 \leq k < n-d$ or $n < k \leq N$. Again, we need d initial conditions $u(0), u(-1), \ldots, u(-d+1)$ in order to be able to define (3.24) for every $1 \leq n \leq N$. In fact, we can define (3.24) for every $n \geq 1$ if we consider Lx an N-periodic sequence. A candidate for B is

$$B = \begin{pmatrix} 1 & & & & & b(1,N-d+1) & \ldots & b(1,N) \\ b(2,1) & 1 & & & & & \ddots & \vdots \\ \vdots & b(3,2) & \ddots & & & & & b(d,N) \\ b(d+1,1) & \vdots & & 1 & & & & \\ & b(d+2,2) & & & 1 & & & \\ & & \ddots & \vdots & \vdots & & \ddots & \\ & & & b(N,N-d) & b(N,N-d+1) & \ldots & 1 & \end{pmatrix},$$

i.e., B is defined by

$$\forall w \in \mathbb{R}^N, \quad \forall n \in \mathbb{Z}_N \quad (Bw)(n) = \sum_{k=n-d}^{n} b(n,k)w(k).$$

Accordingly, given u, we define

$$\forall n = 1, \ldots, d \quad \eta(n) = \sum_{k=0}^{d-n} b(n, N-k)(u(-k) - u(N-k)),$$

and $\eta(n) = 0$ otherwise.

If $\{x_n\}_{n\in\mathbb{Z}_N}$ is a FUNTF for $\mathbb{R}^d$, then, any $d+1$ element subset of this FUNTF is linearly dependent. Then, we can choose $b \in \mathbb{Z}_N^2$ in such a way that

$$\forall k = 1, \ldots, N-d, \quad x_k + \sum_{l=1}^{d} b(k+l,k)x_{k+l} = 0,$$

set $b(n,n) = 1$, and set $b(n,k) = 0$ otherwise. Then, $L^*B \equiv 0$. Since $q = Lx+Bu+\eta$, we have

$$\|x - \frac{d}{N}L^*q\|_a \leq \frac{d}{N}\|L^*\eta\|_a$$

by (3.9). We want to note that $\eta = 0$ if and only if u, defined for all $n \geq 1$ is N-periodic.

For this particular example, it is unrealistic to assume that we can find a u for every x, since this would imply that $x = (d/N)L^*q$. On the other hand, we might hope to find a u that satisfies the condition

$$\forall \varepsilon > 0, \quad \exists t, M > 0, \quad \forall n \geq M, \quad |u(n+tN) - u(n)| \leq \varepsilon. \tag{3.25}$$

The condition (3.25) is closely related to the concept of *almost periodicity*. We give the definition of almost periodic sequences in Definition 12, and prove that every almost periodic sequence satisfy (3.25) in Theorem 24

Definition 12. A sequence $u : \mathbb{Z} \to \mathbb{C}$ is *almost periodic* if it is in the uniform closure of the linear span of the set

$$\{e_\gamma : \mathbb{Z} \to \mathbb{C} \ : \ 0 \leq \gamma < 1\},$$

where $e_\gamma(n) = e^{2\pi i n\gamma}$. In other words, $u : \mathbb{Z} \to \mathbb{C}$ is *almost periodic* if for every

$\varepsilon > 0$, there are $c_1, \ldots, c_r \in \mathbb{C}$ and $\gamma_1, \ldots, \gamma_r \in [0, 1)$ such that

$$\forall n \in \mathbb{Z}, \quad |u(n) - \sum_{k=1}^{r} c_k e^{2\pi i n \gamma_k}| \leq \varepsilon.$$

Theorem 24. If a sequence $u : \mathbb{Z} \to \mathbb{C}$ is *almost periodic* then

$$\forall \varepsilon > 0, \quad \exists m \in \mathbb{N}, \quad \forall n \in \mathbb{Z}, \quad |u(m+n) - u(n)| \leq \varepsilon.$$

Proof. By Definition 12, for every $\varepsilon > 0$, there are $c_1, \ldots, c_r \in \mathbb{C}$ and $\gamma_1, \ldots, \gamma_r \in [0, 1)$ such that

$$\forall n \in \mathbb{Z}, \quad |u(n) - \sum_{k=1}^{r} c_k e^{2\pi i n \gamma_k}| \leq \varepsilon/3.$$

Let α be such that $10^\alpha \sum_{k=1}^{r} |c_k| \varepsilon > 3$. Then, for every k, there are p_k such that

$$|\gamma_k - \frac{p_k}{10^\alpha}| \leq 10^{-\alpha} \varepsilon/3 \sum_{k=1}^{r} |c_k|.$$

Let $m = 10^\alpha$. Then,

$$\begin{aligned}
|u(n) - u(m+n)| &\leq |u(n) - \sum_{k=1}^{r} c_k e^{2\pi i t \gamma_k}| + |\sum_{k=1}^{r} c_k e^{2\pi i n \gamma_k}(1 - e^{2\pi i m \gamma_k})| \\
&\quad + |u(m+n) - \sum_{k=1}^{r} c_k e^{2\pi i (m+n) \gamma_k}| \\
&\leq \varepsilon/3 + \sum_{k=1}^{r} |c_k| |(1 - e^{2\pi i m \gamma_k})| + \varepsilon/3 \\
&\leq 2\varepsilon/3 + \sum_{k=1}^{r} |c_k| |m\gamma_k - p_k| \\
&\leq \varepsilon
\end{aligned}$$

□

Therefore, if u is almost periodic, then u satisfies (3.25) by Theorem 24.

Theorem 25. Let u be a solution of the 1-bit generalized Sigma-Delta system (3.24) that satisfies the condition (3.25). Then, for every $\varepsilon > 0$, there exists a $\widetilde{q} \in \mathbb{R}^N$ with integer entries, and a constant $C > 0$ depending only on b such that

$$\|x - \frac{d}{tN} L^* \widetilde{q}\|_a \leq \frac{d}{N} \|L^* B\|_{a,b} \|u_t\|_b + \frac{d}{Nt} C\varepsilon, \tag{3.26}$$

where $u_t \in \mathbb{R}^N$ is defined by the quantity

$$\forall n = 1, \ldots, N, \quad u_t(n) = \frac{1}{t} \sum_{k=1}^{t} u(kN + n + M).$$

Proof. Let q be the output of (3.24), and let

$$\forall n = 1, \ldots, N, \quad \widetilde{q}(n) = \sum_{k=1}^{t} q(kN + n + M).$$

Also, let $\eta_t \in \mathbb{R}^N$ be defined by

$$\forall n = 1, \ldots, d \quad \eta_t(n) = \sum_{k=0}^{d-n} b(n, N-k)(u(N-k+M) - u((t+1)N - k + M)),$$

and let $\eta_t(n) = 0$ otherwise. Then,

$$\|\eta_t\|_b \leq \sum_{k=0}^{N-1} \|b(., N-k)\|_b |(u(N-k+M) - u((t+1)N-k+M)| \leq \varepsilon \sum_{k=1}^{N} \|b(., k)\|_b.$$

Let

$$C = \|L^*\|_{a,b} \sum_{k=1}^{N} \|b(., k)\|_b.$$

Moreover, a straightforward (and long) calculation shows that $\widetilde{q} = tLx + tBu_t + \eta_t$, where $B = [b(n,k)]_{n,k=1}^{N}$. Therefore, we have

$$\begin{aligned} \|x - \frac{d}{tN} L^* \widetilde{q}\|_a &\leq \frac{d}{N} \|L^* B\|_{a,b} \|u_t\|_b + \frac{d}{Nt} \|L^*\|_{a,b} \|\eta\|_b. \\ &\leq \frac{d}{N} \|L^* B\|_{a,b} \|u_t\|_b + \frac{d}{Nt} C\varepsilon. \end{aligned}$$

□

Almost periodic sequences are bounded. However, the generalized Sigma-Delta schemes does not always give a bounded $u : \mathbb{N} \to \mathbb{R}$. The reason is that the range of the quantizer Q_δ is restricted to a finite range. In the next section, we show that both u and q can be made bounded if we choose a sufficiently wider range for the quantizer. We shall replace the quantizer Q_δ in (3.24) with

$$round(x) = argmin\{|n - x| : n \in \mathbb{Z}\}.$$

This corresponds to increasing the range of the quantizer Q_δ for the variable-bit quantization scheme we shall describe in the next section.

We conclude this section with a few examples, for which the system (3.24) could find an almost periodic u. In the following examples, $\{e_n\}_{n=1}^7$ is 7th roots-of unity frame for $\mathbb{R}^2$, and $h = (1.247, -1, 0, \ldots, 0)$ so that for every $n \in \mathbb{Z}_7$, $e_n = h(1)e_{n+1} + h(2)e_{n+2}$.

Example 3. $x = (2/7)L^*(1, -1, 1, -1, -1, 1, -1)$. u is 7-periodic in $\{n : n \geq 4N\}$, exact reconstruction. There is a plot of u in Figure 3.2.

Example 4. $x = (2/7)L^*(1, -1, 1, 1, -1, 1, 1)$. u is almost periodic in $\{n : n \geq 6N\}$ with $t = 10$. There is a plot of u in Figure 3.3.

$\widetilde{q} = \frac{1}{10}(2, 2, -4, -4, 2, 2, 0)$, error= 0.0025.

Example 5. $x = (-0.1020, 0.4468)$. u is almost periodic for $\{n : n \geq 2N\}$ with $t = 6$. There is a plot of u in Figure 3.4.

$\widetilde{q} = \frac{1}{6}(0, 0, 6, 0, 0, -2, -2)$, error= 0.00092273.

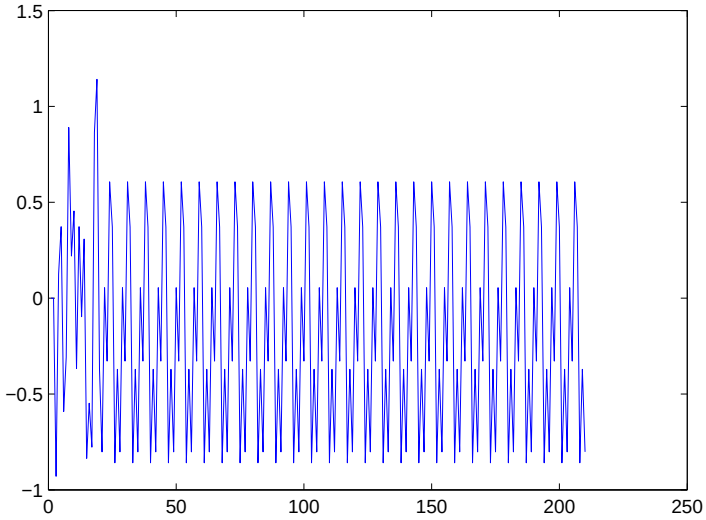

Figure 3.2: Plot of u given in Example 3

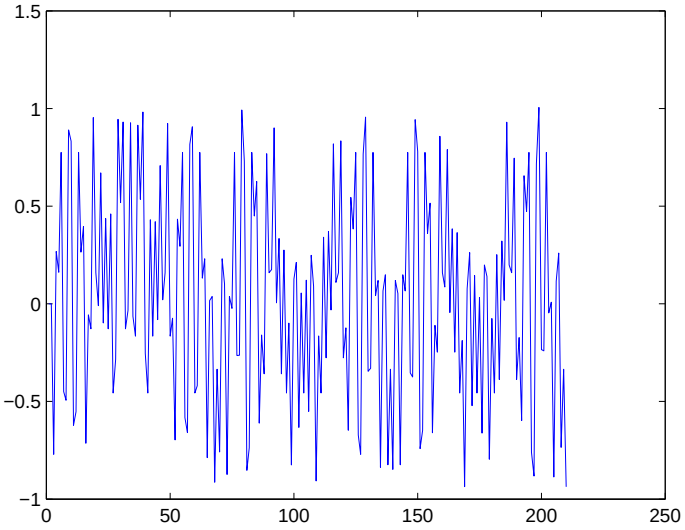

Figure 3.3: Plot of u given in Example 4

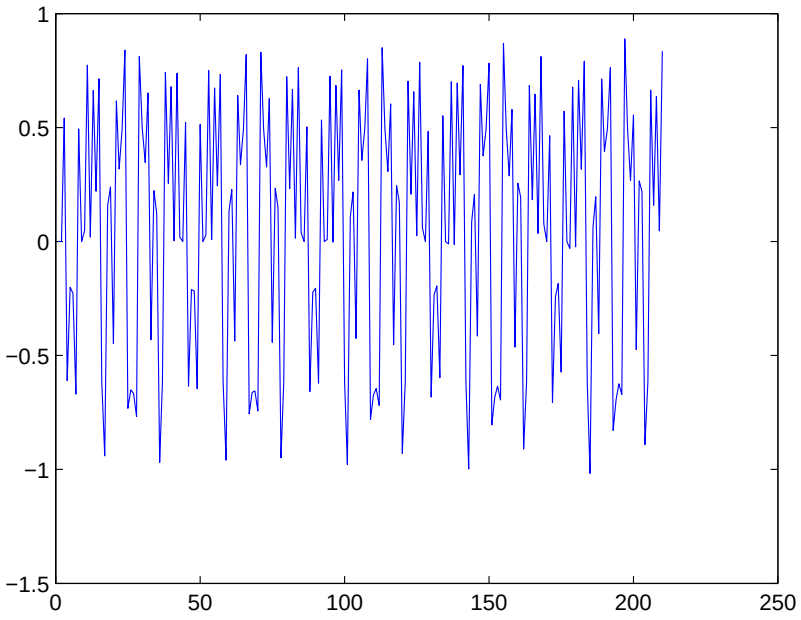

Figure 3.4: Plot of u given in Example 5

3.3 $\mathbb{Z}$-span of Frames and a Variable-bit Quantization

Let $b \geq 1$ be an integer, $\delta = 2^{1-b}$ and $K\delta = 1$. Then, the mid-rise quantization alphabet $\mathcal{A}_\delta$ for real sequences $\{y \in \mathbb{R}^d : \|y\|_\infty \leq 1\}$ is of the form

$$\mathcal{A}_\delta = \{-1 + \frac{\delta}{2} + k\delta : \ k = 0, 1, \ldots, 2K-1\}.$$

Let $\{e_n\}_{n=1}^N$ be a FUNTF for $\mathbb{R}^d$ with the analysis matrix L. Let $\mathcal{S}_\delta$ be the set of all linear combinations of frame vectors with coefficients coming from $\mathcal{A}_\delta$, i.e., let

$$\mathcal{S}_\delta := \frac{d}{N} L^*(\mathcal{A}_\delta \times \mathcal{A}_\delta \times \cdots \times \mathcal{A}_\delta) = \{\frac{d}{N} \sum_{k=1}^N q_n e_n : \ q_n \in \mathcal{A}_\delta\}.$$

Given $x \in \mathbb{R}^d$, the b-bit quantization problem concerns finding an element in $\mathcal{S}_\delta$ that is sufficiently close (closest, if possible) to x.

Any element in the alphabet $\mathcal{A}_\delta$, multiplied by the number $(2/\delta)$ is an odd integer. In fact,

$$(2/\delta)\mathcal{A}_\delta = \{(2a/\delta) : a \in \mathcal{A}_\delta\} = \{2^b + 2k + 1 : k = 0, \ldots, 2K-1\} \subseteq 2\mathbb{Z} + 1.$$

Then,

$$\mathcal{S}_\delta \quad \subseteq \quad \frac{d\delta}{2N} L^*(2\mathbb{Z}^N + 1) \quad = \quad \frac{d\delta}{2N} \sum_{n=1}^N e_n + \frac{d\delta}{N} L^*(\mathbb{Z}^N), \tag{3.27}$$

or,

$$\frac{N}{d\delta}\mathcal{S}_\delta - \frac{1}{2}\sum_{n=1}^N e_n \quad \subseteq \quad L^*(\mathbb{Z}^N). \tag{3.28}$$

Therefore, one considers approximating x by a y in this intermediate set in (3.27), and approximating y by an element $\widetilde{x}$ in $\mathcal{S}_\delta$. This double approximation process would double the difficulty level of the problem if the structure of the intermediate set is not *nicer* than the structure of $\mathcal{S}_\delta$.

The structure of the second set is determined by the group structure of $L^*(\mathbb{Z}^N)$. $L^*(\mathbb{Z}^N)$ is an additive subgroup of $\mathbb{R}^d$, so is its closure $\overline{L^*(\mathbb{Z}^N)}$. Using Theorem 26, which is known as the structure theorem for locally compact Abelian groups, or Van Kampen's theorem, we describe the geometric structure of $\overline{L^*(\mathbb{Z}^N)}$. The proof of Theorem 26 can be found in [67].

Theorem 26. Every locally compact Abelian group (LCAG) G has a subgroup G_0, which is isomorphic to a direct sum of a compact group K and an Euclidean space $\mathbb{R}^n$. Moreover, the factor group G/G_0 is a discrete group.

Theorem 27. Every closed additive subgroup of $\mathbb{R}^d$ is direct a sum of a subspace and a discrete lattice.

Proof. Let G be a closed additive subgroup of $\mathbb{R}^d$. Then, G is a LCAG. Let G_0 be a subgroup of G, let K be a compact group, and let $\phi : K \oplus \mathbb{R}^n \to G$ be the isomorphism as described in Theorem 26. Let $x \in \mathbb{K}$. Then, the closure of the subgroup of K generated by x, $\overline{< x >}$ is also compact. Therefore,

$$\phi(\overline{< x >}) = \overline{\{n\phi(x) : n \in \mathbb{Z}\}}$$

is a compact subgroup of $G_0 \subseteq \mathbb{R}^d$. But, every compact subset of $\mathbb{R}^d$ is bounded, so we must have that $\phi(x) = 0$. Thus, since ϕ is an isomorphism, K can only have one element. Therefore, G_0 is isomorphic to $\mathbb{R}^n$, so G_0 is a subspace of $\mathbb{R}^d$.

Next, G/G_0 is a subgroup of $\mathbb{R}^d/G_0$. Since G_0 is a subspace, $\mathbb{R}^d/G_0$ is isomorphic to the orthogonal complement of G_0 in $\mathbb{R}^d$. But, any nontrivial discrete additive subgroup of a linear space is a discrete lattice.

Hence, $G = V \oplus D$, where V is a subspace of $\mathbb{R}^d$ and D is a discrete lattice. G is a subspace, if D is trivial, and G is a discrete lattice if $V = \{0\}$. $\square$

Example 6. Let $e_1 = (1, 0)$, $e_2 = (-1/2, \sqrt{3}/2)$ and $e_3 = (-1/2, -\sqrt{3}/2)$. Then, $\{e_1, e_2, e_3\}$ is a FUNTF for $\mathbb{R}^2$. This frame satisfies $e_1 + e_2 + e_3 = 0$. Let L be the analysis matrix of this FUNTF. Then, $L^*(\mathbb{Z}^3)$ is a discrete lattice, generated by any of the two frame elements, i.e,

$$L^*(\mathbb{Z}^3) = \{me_1 + ne_2 : \ m, n \in \mathbb{Z}\}.$$

Example 7. Let $e_1 = (1, 0)$, $e_2 = (0, 1)$ and $e_3 = (1, 1)$. Then, $\{e_1, e_2, e_3\}$ is a frame for $\mathbb{R}^2$. $\overline{L^*(\mathbb{Z}^3)}$ is a direct sum of the discrete lattice generated by e_1 and e_2, and the line generated by e_3, i.e.,

$$\overline{L^*(\mathbb{Z}^3)} = \{me_1 + ne_2 + te_3 : \ m, n \in \mathbb{Z}, \ t \in \mathbb{R}\}.$$

Example 8. Let $e_n = (\cos(2\pi n/N), \sin(2\pi n/N))$. $\{e_n\}_{n=1}^N$ is the Nth roots of unity frame for $\mathbb{R}^2$, and it is a FUNTF for $\mathbb{R}^2$. Then, $\overline{L^*(\mathbb{Z}^N)}$ is equal to $\mathbb{R}^2$ for $N \geq 7$.

In fact, let

$$\begin{aligned} e_n^{(1)} &= e_{n-1} - 2e_n + e_{n+1} = 2\sin^2(\pi/N)e_n, \\ e_n^{(k+1)} &= e_{n-1}^{(k)} - 2e_n^{(k)} + e_{n+1}^{(k)} = (2\sin^2(\pi/N))^{k+1} e_n^{(k)}. \end{aligned}$$

Then, $e_n^{(k)} \in L^*(\mathbb{Z}^N)$ for every n and k. Moreover, $2\sin^2(\pi/N) < 1$ if and only if $N \geq 7$.

If $L^*(\mathbb{Z}^N)$ were not dense in $\mathbb{R}^2$, then there would exist an $x \in \mathbb{R}^2$, and an $\varepsilon > 0$ such that the intersection of $\overline{L^*(\mathbb{Z}^N)}$ and the ε ball $N_\varepsilon(x)$ centered at x is

empty. Choose the biggest possible $\varepsilon > 0$ such that there is an element z of $\overline{L^*(\mathbb{Z}^N)}$ on the boundary of $N_\varepsilon(x)$. Then, $N_\varepsilon(x-z) \cap \overline{L^*(\mathbb{Z}^N)}$ is empty, and 0 lies on the boundary of $N_\varepsilon(x-z)$. However, such a ball must intersect with $\{e_n^{(k)}\}_{n=1}^N$ for some k. Therefore, $L^*(\mathbb{Z}^N)$ must be dense in $\mathbb{R}^2$.

If $L^*(\mathbb{Z}^N)$ is a discrete lattice in $\mathbb{R}^d$, the frame $\{e_n\}_{n=1}^N$ includes a basis B for $\mathbb{R}^d$, and all the other frame vectors can be expressed as a linear combination of this basis elements with integer coefficients. B cannot be less than a basis, since a frame is a spanning set. B cannot be more than a basis, either. Since otherwise, $L^*(\mathbb{Z}^N)$ would have an accumulation point by Theorem 28, and so, it would not be a discrete lattice.

Given $y \in \mathbb{R}^d$, we calculate the B basis coefficients of y, and round them to the nearest integer. The basis expansion $\widetilde{y}$ with these integer coefficients $c = (c_k)$ is usually the closest point in $L^*(\mathbb{Z}^N)$ to y. $\widetilde{y}$ might not be the closest point if the determinant of B is too close to zero. On the other hand, this expansion is not the only way to express $\widetilde{y}$ as a linear combination of frame elements with integer coefficients. If we let

$$\mathcal{N}_{\mathbb{Z}}(L) = \{z = (z_k)_{k=1}^N \in \mathbb{Z}^N : \sum_{k=1}^N z_k e_k = 0\},$$

then any $(\widetilde{c}_k)$ that satisfy

$$\widetilde{y} = \sum_{k=1}^N \widetilde{c}_k e_k$$

can be expressed as a sum $\widetilde{c} = c + z$ for some $z \in \mathcal{N}_{\mathbb{Z}}(L)$. Therefore, we have a wide

variety of choices for coefficients.

By (3.28) we have

$$\frac{N}{d\delta}\mathcal{S}_\delta - \frac{1}{2}\sum_{n=1}^{N} e_n = \{\sum_{n=1}^{N} z_n e_n : \ z_n = -K, \ldots, K-1\}.$$

Therefore, we want each coefficient c_k to fall in the range $-K, \ldots, K-1$. If not, we need to find another coefficient sequence $\widetilde{c}$ with entries falling in the required range such that $\widetilde{c} = c + z$ for some $z \in \mathcal{N}_{\mathbb{Z}}(L)$ if possible. Otherwise, we want $\widetilde{c}$ minimize the quantity

$$\| \sum_{k=1}^{N} c_k e_k - \sum_{k=1}^{N} \widetilde{c}_k e_k \|.$$

Hence, given $x \in \mathbb{R}^d$, to quantize x relative to $\mathcal{A}_\delta$, we set

$$y = \frac{N}{d\delta}x - \frac{1}{2}\sum_{n=1}^{N} e_n,$$

then find the corresponding $\widetilde{c}$, and then let

$$\widetilde{x} = \frac{d\delta}{N}\sum_{k=1}^{N}\widetilde{c}e_k + \frac{d\delta}{2N}\sum_{n=1}^{N} e_n = \frac{d}{N}\sum_{k=1}^{N}\delta\left(\widetilde{c}_k + \frac{1}{2}\right)e_k.$$

However, finding such $(\widetilde{c}_k)$ is a difficult problem. Instead, we refer to another method, with which we can make (c_k) fall into the desired range. We describe this method in (3.29). Theorem 30 shows how we can guarantee to make (c_k) to fall in a desired range with a proper choice of b for (3.29).

If $L^*(\mathbb{Z}^N)$ is not a discrete lattice, then the problem of approximating a $y \in \mathbb{R}^d$ by a $\widetilde{y}$ in $L^*(\mathbb{Z}^N)$ is a relatively difficult problem using V and D. $L^*(\mathbb{Z}^N)$ is not a discrete lattice if and only if the frame includes a $d+1$ element $\mathbb{Z}$-independent subset in the sense of Definition 13.

Definition 13. $x_1, \ldots, x_N \in \mathbb{R}^d$ is $\mathbb{Z}$-*independent* if for every $c_1, \ldots, c_N \in \mathbb{Z}$

$$\sum_{i=1}^{N} c_i x_i = 0 \quad \Rightarrow \quad c_i = 0, \quad \forall i = 1, \ldots, N.$$

Definition 14. Let $a = (a_1, a_2, \ldots, a_d) \in \mathbb{R}^d$. We define the *floor* of a as

$$\lfloor a \rfloor = (\lfloor a_1 \rfloor, \ldots, \lfloor a_d \rfloor).$$

The following theorem is a direct result of Theorem 27. We also provide an alternative proof.

Theorem 28. Let $\{x_i\}_{i=1}^{d+1}$ be a $\mathbb{Z}$-independent subset of $\mathbb{R}^d$. Then, 0 is an accumulation point of the additive group

$$\mathbb{Z}[x_1, \ldots, x_{d+1}] := \left\{ \sum_{i=1}^{d+1} c_i x_i : c_i \in \mathbb{Z} \right\}.$$

Proof. Let L be the matrix with kth row is equal to x_k. Then, $\mathbb{Z}[x_1, \ldots, x_{d+1}] = L^*(\mathbb{Z}^{d+1})$. For a contradiction, assume that 0 is not an accumulation point. Then,

$$\exists \varepsilon_0 > 0 : \ N_{\varepsilon_0}(0) \cap Z[x_1, \ldots, x_{d+1}] = \{0\}.$$

For any $\varepsilon < \varepsilon_0$, let

$$\mathcal{C}^{\varepsilon} = \{y \in \mathbb{R}^{d+1} : ||L^* y|| < \varepsilon\}.$$

Then,

$$\forall \varepsilon \leq \varepsilon_0, \ \mathcal{C}^{\varepsilon} \cap \mathbb{Z}^{d+1} = \{0\}.$$

In fact, if there existed $y \neq 0$, $y \in \mathcal{C}^{\varepsilon} \cap \mathbb{Z}^{d+1}$, then $L^*(y) \in N_{\varepsilon_0}(0) \cap \mathbb{Z}[\{x_i\}] = \{0\}$, so $L^* y = 0$. But then, $\{x_i\}_{i=1}^{d+1}$ would not be $\mathbb{Z}$-independent.

Now, for any $z \in \mathbb{Z}^{d+1}$, $\varepsilon \leq \varepsilon_0$, let

$$\mathcal{C}_z^{\varepsilon} := \{a - \lfloor a \rfloor : a \in \mathcal{C}^{\varepsilon}, \lfloor a \rfloor = z\}.$$

Clearly, each nonempty $\mathcal{C}_z^\varepsilon$ lies in the unit cube $\{a = (a_1, \ldots, a_{d+1}) : 0 \leq a_i < 1\}$. Moreover,

$$\forall z_1 \neq z_2 \in \mathbb{Z}^{d+1}, \quad \forall \varepsilon \leq \frac{\varepsilon_0}{2}, \quad \mathcal{C}_{z_1}^\varepsilon \cap \mathcal{C}_{z_2}^\varepsilon = \emptyset.$$

In fact, if there existed $\alpha \in \mathcal{C}_{z_1}^\varepsilon \cap \mathcal{C}_{z_2}^\varepsilon$, then there would exist $a, b \in \mathcal{C}^\varepsilon$ such that $\alpha = a - z_1$, $\alpha = b - z_2$. But, $||L^*(a-b)|| \leq ||L^*(a)|| + ||L^*(b)|| \leq 2\varepsilon < \varepsilon_0$, and so $z_1 - z_2 = a - b \in \mathcal{C}^{\varepsilon_0}$. Then, since $\mathcal{C}^{\varepsilon_0} \cap \mathbb{Z}^{d+1} = \{0\}$, we would have $z_1 = z_2$.

Hence, for $\varepsilon \leq \varepsilon_0/2$, $\{\mathcal{C}_z^\varepsilon\}_{z \in \mathbb{Z}^{d+1}}$ is a countable, disjoint family of sets, satisfying

- $\bigcup (z + \mathcal{C}_z^\varepsilon) = \mathcal{C}^\varepsilon$, and
- $\forall z \in \mathbb{Z}^{d+1}, \mathcal{C}_z^\varepsilon \subseteq \{a = (a_1, \ldots, a_{d+1}) : 0 \leq a_i < 1\}$.

Let $\mathcal{L}^n$ denote the usual Lebesgue measure on $\mathbb{R}^n$. Then,

$$\mathcal{L}^{d+1}(\mathcal{C}^\varepsilon) = \sum_z \mathcal{L}^{d+1}(z + \mathcal{C}_z^\varepsilon) = \mathcal{L}^{d+1}(\bigcup \mathcal{C}_z^\varepsilon) \leq \mathcal{L}^{d+1}(\{a = (a_1, \ldots, a_{d+1}) : 0 \leq a_i < 1\}) = 1.$$

But, $\mathcal{L}^{d+1}(\mathcal{C}^\varepsilon) = \infty$. Contradiction.

By contradiction, 0 must be an accumulation point of the group $\mathbb{Z}[x_1, \ldots, x_{d+1}]$.

□

If we have a prior knowledge about the structure of V and D, then given $y \in \mathbb{R}^d$, we first project y onto the closest lattice shift of V. Let this point be y_0. This affine subspace can be expressed as $d + V$ for some $d \in D$. Then, we approximate $y_0 - d$ in $V \cap L^*(\mathbb{Z}^N)$. One might want to choose a suitable basis $B \subseteq L^*(\mathbb{Z}^N)$ for V, and find integer coefficients such that B basis expansion is close to $y_0 - d$. This B basis expansion is, in turn, a frame expansion with integer coefficients c_k.

Alternatively, we use the following system, which is a modified version of (3.24)

$$\begin{aligned} q(n) &= round(-\sum_{k=n-d}^{n-1} b(n,k)u(k) + Lx(n)) \\ u(n) &= -\sum_{k=n-d}^{n-1} b(n,k)u(k) + Lx(n) - q(n) \end{aligned} \tag{3.29}$$

for $n \geq 1$ with the initial conditions $u(0), u(-1), \ldots, u(-d+1)$, and a $b : \mathbb{Z}_N^2 \to \mathbb{R}$ such that $b(n,n) = 1$, and $b(n,k) = 0$ if $1 \leq k < n-d$ or $n < k \leq N$.

System (3.29) is equivalent to (3.24) with a slight change in the quantizer. This relation is stated in Theorem 29.

Theorem 29. Let $\delta > 0$, and let $\widetilde{Q}_\delta(a) = \delta round(\delta^{-1}a)$. Then, the system (3.29) with the input sequence Lx and the initial conditions $u(0), \ldots, u(-d+1)$ is equivalent to the generalized Sigma-Delta system (3.24) with the quantizer $\widetilde{Q}_\delta$, the input sequence δLx and the initial conditions $\delta u(0), \ldots, \delta u(-d+1)$, in the sense that u and q are outputs of (3.29) if and only if δu and δq are outputs of (3.24).

The following theorem shows that the output sequences q and u of (3.29) are always bounded.

Theorem 30. Given $x \in \mathbb{R}^d$, let q and u be the output sequences of the system (3.29). Then, $\|u\|_\infty \leq 1/2$, and

$$\forall n \geq 1, \quad |q(n)| \leq \frac{1}{2}\sum_{k=1}^{N} |b(n,k)| + |Lx(n)|.$$

Proof. If $\beta_n = -\sum_{k=n-d}^{n-1} b(n,k)u(k) + Lx(n)$, then $|u(n)| \leq |\beta_n - round(\beta_n)| \leq 1/2$.

Second,

$$|q(n)| \leq |\sum_{k=n-d}^{n} b(n,k)u(k)| + |Lx(n)| \leq \|u\|_\infty \sum_{k=1}^{N} |b(n,k)| + |Lx(n)|.$$

Hence, the result follows. □

By Theorem 30 and Theorem 29, we can show that if the range of the quantizer Q_δ we use for the generalized Sigma-Delta system (3.24) is wide enough to include

$$\pm\delta \max_{1\le n\le N}\left(\frac{1}{2}\sum_{k=1}^{N}|b(n,k)|+1\right),$$

then the output sequences of (3.24) are bounded.

Theorem 31. Let u be a solution of the system (3.29) that satisfies the condition (3.25). Let $B=[b(n,k)]_{n,k=1}^{N}$. Then, there exists a constant $C>0$ depending only on b, and for every $\varepsilon>0$, there is a positive integer t, and a $\widetilde{q}\in\mathbb{R}^N$ with integer entries such that

$$|\widetilde{q}(n)|\le t(\frac{C}{2}+|Lx(n)|).$$

Furthermore,

$$\|x-\frac{d}{tN}L^*\widetilde{q}\|\le\frac{d}{2N}\|L^*B\|_{2,\infty}+\frac{d}{Nt}C\varepsilon. \tag{3.30}$$

Proof. First part follows from Theorem 30 with any

$$C>C_1:=\max_{1\le n\le N}\sum_{k=1}^{N}|b(n,k)|.$$

Let q be the output of (3.29), and let

$$\forall n=1,\dots,N,\quad \widetilde{q}(n)=\sum_{k=1}^{t}q(kN+n+M).$$

Also, let $\eta_t\in\mathbb{R}^N$ be defined by

$$\forall n=1,\dots,d\quad \eta_t(n)=\sum_{k=0}^{d-n}b(n,N-k)(u(N-k+M)-u((t+1)N-k+M)),$$

and let $\eta_t(n) = 0$ otherwise. Then,

$$\begin{aligned}\|\eta_t\|_\infty &\leq \max_{1\leq n\leq N}\sum_{k=0}^{n-d}|b(n,N-k)|\ |(u(N-k+M)-u((t+1)N-k+M)|\\ &\leq \varepsilon\max_{1\leq n\leq N}\sum_{k=1}^{N}|b(n,k)|\\ &= \varepsilon C_1.\end{aligned}$$

Let $u_t \in \mathbb{R}^N$ is defined by the quantity

$$\forall n = 1,\ldots,N, \quad u_t(n) = \frac{1}{t}\sum_{k=1}^{t} u(kN+n+M).$$

Then, $\widetilde{q} = tLx + tBu_t + \eta_t$. Therefore, we have

$$\begin{aligned}\|x - \frac{d}{tN}L^*\widetilde{q}\| &\leq \frac{d}{N}\|L^*B\|_{2,\infty}\|u_t\|_\infty + \frac{d}{Nt}\|L^*\|_{2,\infty}\|\eta_t\|_\infty.\\ &\leq \frac{d}{N}\|L^*B\|_{2,\infty}\|u_t\|_\infty + \frac{d}{Nt}\|L^*\|_{2,\infty}C_1\varepsilon.\end{aligned}$$

The result follows with $\|u_t\|_\infty \leq \|u\|_\infty \leq 1/2$, and $C = \max\{\|L^*\|_{2,\infty}, 1\}C_1$ □

We would like to note that, even if u does not satisfy (3.25), for any integer $t > 0$, we still have that

$$\|\eta_t\|_\infty \leq 2\|u\|_\infty\frac{d}{Nt}\|L^*\|_{2,\infty}C_1 \leq \frac{d}{Nt}\|L^*\|_{2,\infty}C_1.$$

Therefore, if $L^*B = 0$, then

$$\|x - \frac{d}{tN}L^*\widetilde{q}\| = \mathcal{O}(\frac{1}{t}) \quad \text{as} \quad t\to\infty.$$

If we think of $t^{-1}\widetilde{q}$ as a $log_2(t)$-bit quantized sequence of x, then, the quantization system (3.25) attains the best error decay rate in the bit rate.

Definition 15. Let $\{e_n\}_{n=1}^N$ be a given frame for $\mathbb{R}^d$, and let $b : \mathbb{Z}_2^N \to \mathbb{R}$ with the corresponding matrix $B = [b(n,k)]_{n,k=1}^N$ satisfy $L^*B = 0$. We call the system (3.29) with this choice of B a *variable-bit rate generalized Sigma-Delta quantization scheme.*

Using a variable-bit rate quantization scheme is more advantageous than a fixed b-bit scheme. If there is no solution u of the variable-bit rate scheme, then, both quantization schemes have the same error decay rate as $b \to \infty$. However, if for some given $\varepsilon > 0$, there is a t satisfying $t < \varepsilon 2^b$, and if there is a solution u the variable-bit system satisfying (3.25) for this $\varepsilon > 0$, then we can achieve the same quantization error with a lower bit number $log_2(t)$.

3.4 1-bit Quantization by Minimization

Frame quantization problem is inherently a combinatorial problem. Given a frame $\{e_n\}_{n=1}^N$ for $\mathbb{R}^d$, and an $x \in \mathbb{R}^d$, the frame quantization problem concerns minimizing the quantity

$$f_x(q) = \|x - \frac{d}{N} \sum q_n e_n\|,$$

subject to the constraint

$$q \in \mathcal{S}_\pm := \{q \in \mathbb{R}^N : q_n = \pm 1\}.$$

In this section, we relax this constraint by means of adding a penalty term to the objective function. This way, we replace the combinatorial problem with an analytic problem.

In general, one might consider to construct functions $F_x : \mathbb{R}^N \to \mathbb{R}$, $F_x \geq 0$ such that a minimizer y of F_x makes the quantity

$$\|x - \frac{d}{N} L^* y\|$$

sufficiently small, while y is in or close to the set $\mathcal{S}_\pm$. In this section, we consider functions F_x of the form

$$F_x(y) = \lambda \|x - \frac{d}{N} L^* y\| + P(y),$$

where $P : \mathbb{R}^N \to \mathbb{R}$, $P \geq 0$ is a penalty term that has small values if y is in or close to $\mathcal{S}_\pm$ and gets bigger values as y moves away from $\mathcal{S}_\pm$. $\lambda > 0$ is a tuning parameter, with which we adjust the weight of the penalty term on the functional F_x.

We shall consider a penalty term of the form

$$P(y) = \sum_{k=1}^{N} f(y_k),$$

where $f \geq 0$, and $f \in \mathcal{C}^2(\mathbb{R})$. In particular, we shall investigate the nature of the local/global minimizers of the functional

$$F_x(y) = \lambda \|x - \frac{d}{N} L^* y\|_2^2 + \sum_{k=1}^{N} f(y_k), \tag{3.31}$$

for

$$f(t) = (1 - t^n)^2 + c(1 - t^m),$$

where m, n are even positive integers with $m \leq n$, and $c \geq 0$.

Lemma 8. Let F_x be given as in (3.31). Then, F_x is a $\mathcal{C}^2$-function. Moreover, the

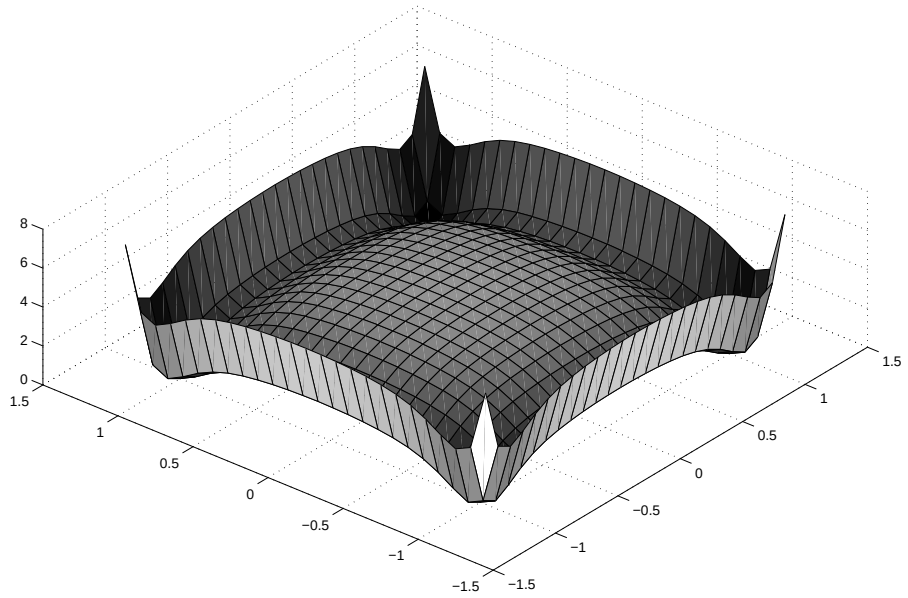

Figure 3.5: $P(y) = f(y_1) + f(y_2)$ with $n = 20$, $c = 1$, $m = 2$.

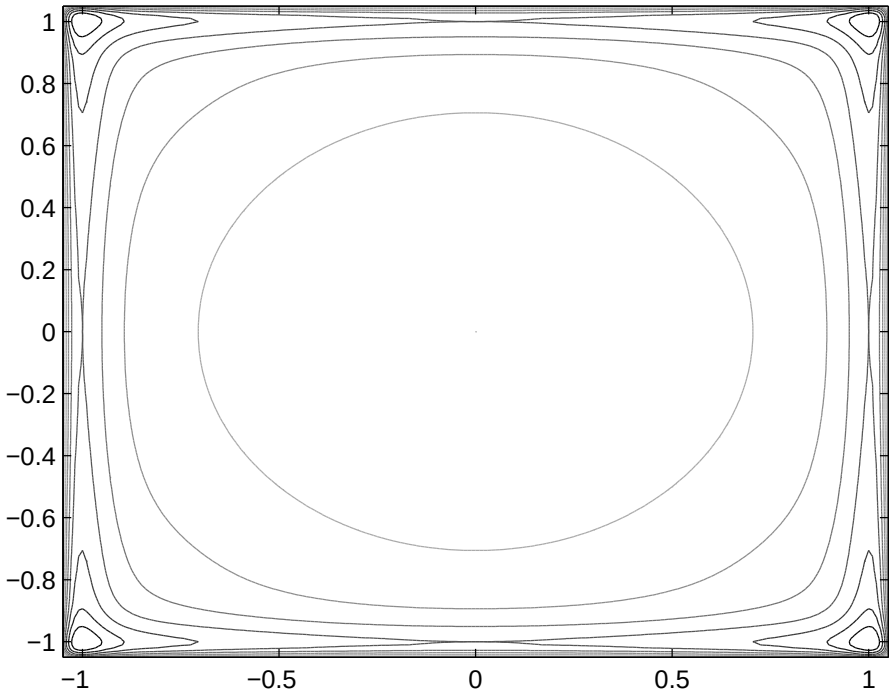

Figure 3.6: Level curves of P in Figure 3.5

gradient DF_x and the Hessian D^2F_x of F_x are given by

$$\begin{aligned} DF_x(y) &= \frac{2\lambda d}{N}\left(\frac{d}{N}LL^*y - Lx\right) + (f'(y_1), \ldots, f'(y_N)), \\ D^2F_x(y) &= \frac{2\lambda d^2}{N^2}LL^* + diag(f''(y_1), \ldots, f''(y_N)), \end{aligned}$$

where $diag(f''(y_1), \ldots, f''(y_N))$ is a diagonal matrix with diagonal entries $f''(y_1), \ldots, f''(y_N)$.

The penalty term P with the choice $f(t) = (1-t^n)^2 + c(1-t^m)$, i.e.,

$$P(y) = \sum_{k=1}^{N}(1-y_k^n)^2 + c(1-y_k^m)$$

effectively works as a barrier. By this we mean, P gets relatively large values off of the unit cube, and this property of P guarantees that the minimizers of F_x are in or close to the unit cube (Theorem 33.a).

We need the following theorem for the proof of Theorem 33. Theorem 32 and a proof can be found in [62].

Theorem 32. Let $G : \mathbb{R}^N \to \mathbb{R}$ be a $\mathcal{C}^2$ function satisfying $G \geq 0$. Then, every nonconstant bounded solution of the ordinary differential equation

$$\dot{\gamma}(t) = -DG(\gamma(t))$$

converges to a local minimum of G. Also, every isolated local minimum y of G is asymptotically stable, i.e.,

$$\exists \varepsilon > 0 \quad \text{such that} \quad \|y - \gamma(0)\| < \varepsilon \Rightarrow \lim_{t\to\infty} \gamma(t) = y.$$

Theorem 33. Let $c \geq 0$, m, n positive even integers with $m \leq n$, and let

$$F_x(y) = \lambda \|x - \frac{d}{N} L^* y\|_2^2 + \sum_{k=1}^{N} (1 - y_k^n)^2 + c(1 - y_k^m).$$

a. Let $\varepsilon > 0$. Every solution of the ordinary differential equation $\dot{\gamma}(t) = -DF_x(\gamma(t))$ enters the bounded set

$$\mathcal{B}^{\varepsilon}_{n,m,\lambda} := \left\{ y \in \mathbb{R}^N : \sum_{k=1}^{N} \left(y_k^n - \frac{1}{2} \right)^2 - \frac{cm}{2n} \sum_{k=1}^{N} y_k^m \leq \frac{N}{4} + \frac{\lambda}{4n} \|x\|^2 + \varepsilon \right\}$$

and stays there.

b. If y is a local minimum of F_x, then y is in the set

$$\mathcal{B}_{n,m,\lambda} := \left\{ y \in \mathbb{R}^N : \sum_{k=1}^{N} \left(y_k^n - \frac{1}{2} \right)^2 - \frac{cm}{2n} \sum_{k=1}^{N} y_k^m \leq \frac{N}{4} + \frac{\lambda}{4n} \|x\|^2 \right\}.$$

In particular $\|y\|_\infty \leq R$, where R is the positive root of the polynomial

$$\Pi(\rho) = \rho^{2n} - \rho^n - \frac{cmN}{2n} \rho^m - \frac{\lambda}{4n} - \frac{N-1}{4}$$

that depends on n, m, c and λ. Moreover, $R = 1 + \mathcal{O}(n^{-1})$ as $n \to \infty$.

c. If y is a local minimum of F_x, then $|y_k| \geq r$ for at least $N - d$ indices, where r is the positive root of the polynomial

$$\pi(\rho) = \rho^{n-m}((2n-1)\rho^n - (n-1)) - \frac{cm(m-1)}{2n}$$

that depends on n, m and c. Moreover, $r = 1 - \mathcal{O}(n^{-1})$ as $n \to \infty$.

Proof. a. By Lemma 8,

$$DF_x(y) = \frac{2\lambda d}{N} \left(\frac{d}{N} L L^* y - L x \right) + 2n(y^{2n-1} - y^{n-1}) - cmy^{m-1},$$

where $y^r := (y_1^r, \ldots, y_N^r)$ with the abuse of notation. Let $\gamma = (\gamma_1, \ldots, \gamma_N)$ be a solution of the system $\dot{\gamma}(t) = -DF_x(\gamma(t))$. Then,

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\|\gamma(t)\|_2^2 = -\langle DF_x(\gamma(t)), \gamma(t)\rangle \\
= \;& -2\lambda\left(\left\|\frac{d}{N}L^*\gamma(t)\right\|^2 - \langle x, \frac{d}{N}L^*\gamma(t)\rangle\right) - 2n\sum_{k=1}^{N}\left(\gamma_k^{2n}(t) - \gamma_k^n(t)\right) + cm\sum_{k=1}^{N}\gamma_k^m(t) \\
= \;& -2\lambda\left\|\frac{d}{N}L^*\gamma(t) - \frac{1}{2}x\right\|^2 + \frac{\lambda}{2}\|x\|^2 - 2n\sum_{k=1}^{N}\left(\gamma_k^n(t) - \frac{1}{2}\right)^2 + cm\sum_{k=1}^{N}\gamma_k^m(t) + \frac{2nN}{4}.
\end{aligned}$$

Hence,

$$\forall \varepsilon > 0, \quad \frac{d}{dt}\|\gamma(t)\|_2^2 \leq -\varepsilon$$

if $\gamma(t)$ is outside of the bounded set $\mathcal{B}^{\varepsilon}_{n,m,\lambda}$. Therefore, γ is bounded, and so it converges to a local minimum of F_x by Theorem 32. Then,

$$\lim_{t\to\infty}\frac{d}{dt}\|\gamma(t)\|_2^2 = 0.$$

Thus, γ must enter and stay in each $\mathcal{B}^{\varepsilon}_{n,m,\lambda}$.

b. If y is a local minimum of F_x, there is a nonconstant solution to $\dot{\gamma} = -DF_x(\gamma)$ with $\lim_{t\to\infty}\gamma(t) = y$ by Theorem 32. By part (a), for every $\varepsilon > 0$, $y \in \mathcal{B}^{\varepsilon}_{n,m,\lambda}$. Therefore, $y \in \mathcal{B}_{n,m,\lambda}$.

Second, since $y \in \mathcal{B}_{n,m,\lambda}$,

$$\left(\|y\|_\infty^n - \frac{1}{2}\right)^2 \leq \frac{cmN}{2n}\|y\|_\infty^m + \frac{N}{4} + \frac{\lambda}{4n}\|x\|^2.$$

Hence,

$$\|y\|_\infty^{2n} - \|y\|_\infty^n - \frac{cmN}{2n}\|y\|_\infty^m - \frac{\lambda}{4n} - \frac{N-1}{4} \leq 0.$$

From this, part (b) follows.

c. Since F_x is a $\mathcal{C}^2$-function, if y is a local minimum, then the Hessian matrix $D^2F_x(y)$ is positive definite. By Lemma 8,

$$D^2F_x(y) = \frac{2\lambda d^2}{N^2}LL^* + diag(f''(y_1), \ldots, f''(y_N)),$$

where $f''(t) = 2nt^{n-2}((2n-1)t^n - (n-1)) - cm(m-1)t^{m-2}$. Then,

$$\forall z \in \mathbb{R}^N, \quad 0 < \langle D^2F_x(y)z, z\rangle = \sum_{k=1}^{N} f''(y_k)|z_k|^2 + \frac{2\lambda d^2}{N^2}\|L^*z\|^2. \tag{3.32}$$

Let $J = \{k : f''(y_k) \leq 0\}$. If $|J| > d$, since $KerL^*$ is $N-d$ dimensional, we could find a $z \in KerL^*$ such that $z_k = 0$ if $k \notin J$. But, this would contradict (3.32). Thus, $|J| \leq d$. Hence,

$$\forall k \notin J, \quad 0 < f''(y_k) = 2ny_k^{n-2}((2n-1)y_k^n - (n-1)) - cm(m-1)y_k^{m-2}.$$

Then, $y_k^{n-m}((2n-1)y_k^n - (n-1)) - cm(m-1)/2n > 0$, and from this, part (c) follows. □

Example 9. If we choose $n = 10$, $c = 0$, $N = 500$ and $\lambda = 1000$, then

$$r = 0.928 \quad \text{and} \quad R = 1.273.$$

If we choose $n = 100$, $m = 2$, $c = 2$, $N = 256$, $d = 16$ and $\lambda = 2^{N/d}$, then

$$r = 0.993 \quad \text{and} \quad R = 1.028.$$

Theorem 34. Let y be a local minimizer of F_x, let $q = (round(y_1), \ldots, round(y_N))$, and let $J \subseteq \{1, \ldots, N\}$ be the set of indices where $|y_k| < r$. Then, we obtain the

decomposition

$$x - \frac{d}{N}L^*q = x_\lambda + x_{ns} + x_J \tag{3.33}$$

where

$$\begin{aligned} x_\lambda &= x - \frac{d}{N}L^*y, \quad \text{and} \quad \|x_\lambda\| = \mathcal{O}(\lambda^{-1}) \quad \text{as} \quad \lambda \to \infty, \\ x_{ns} &= \frac{d}{N}\sum_{k\notin J}(y_k - q_k)e_k, \quad \text{and} \quad \|x_{ns}\| = \mathcal{O}(n^{-1}) \quad \text{as} \quad n \to \infty, \\ x_J &= \frac{d}{N}\sum_{k\in J}(y_k - q_k)e_k, \quad \text{where} \quad |J| \le d. \end{aligned}$$

Proof. For a fixed N, since y is a local minimum,

$$DF_x(y) = \frac{2\lambda d}{N}\left(\frac{d}{N}LL^*y - Lx\right) + (f'(y_1), \ldots, f'(y_N)) = 0.$$

Then,

$$\|x_\lambda\|^2 = \|x - \frac{d}{N}L^*y\|^2 = \frac{d}{N}\|Lx - \frac{d}{N}LL^*y\|^2 = \frac{1}{2\lambda}\sum_{k=1}^{N}|f'(y_k)|^2 = \mathcal{O}(\lambda^{-1}) \quad \text{as} \quad \lambda \to \infty.$$

By Theorem 33, there is a set of indices J, $|J| \le d$, such that if $k \in \{1, \ldots, N\}\backslash J$, then

$$|y(k) - q(k)| = |y(k)| - 1 = \mathcal{O}(n^{-1}) \quad \text{as} \quad n \to \infty.$$

Therefore,

$$\|x_{ns}\| = \|\frac{d}{N}\sum_{k\notin J}(y_k - q_k)e_k\| \le \frac{d}{N}\sum_{k\notin J}|y_k - q_k| \le \max_{k\notin J}|y_k - q_k| = \mathcal{O}(n^{-1}) \quad \text{as} \quad n \to \infty.$$

□

We finish this section with a few examples. In the following examples, we used the simple MATLAB code to minimize F_x. For each example, the minimization starts at the point $y_0 = Lx$.

```
n=100; c=1; m=4; [N d]=size(L); lambda=2^(N/d); y0=Lx;
optns=optimset(Display,on,Largescale,on,...
       MaxIter,1e+4, MaxFunEvals,1e+5,...
       Gradobj,on,Hessian,on);
y=fminunc(@(y)F_x(y,x,L,lambda,n,c,m),y0,optns); q=round(y);
```

Example 10. For this example, we used the real Harmonic frame for $\mathbb{R}^4$. The real harmonic frames H_N^d with N elements for $\mathbb{R}^d$ are defined by

- If $d = 2k$,

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

- If $d = 2k + 1$,

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

We quantized the vector $x \in \mathbb{R}^{16}$

$$\begin{aligned} x \;=\; & (-0.33778, 0.008157, 0.12914, 0.53439, 0.55974, -0.031804, \\ & 0.60443, -0.057976, -0.59448, 0.159230, 0.333, 0.35353, \\ & 0.88502, 0.5403, 0.47481, 0.73252), \end{aligned}$$

and calculated the quantization error for various values of N, which can be seen in Figure 3.8. The parameters we used are $\lambda = 2^{N/d}$, $n = 100$, $c = 1$, $m = 4$.

For $N = 216$, we obtained the following values for each of the components of the decomposition given in Theorem 34. Figure 3.7 shows how the sequence

$(y_k - q_k)_{k=1}^N$ behaves. It can be seen that

$$J = \{15, 37, 46, 72, 87, 101, 107, 138, 144, 167, 190, 214\}.$$

$$\begin{aligned} \|x_\lambda\| &= 4.1062e-005 \\ \|x_{ns}\| &= 0.0012903 \\ \|x_J\| &= 0.085473 \end{aligned}$$

$$card\{k \in J : q_k = 0\} = 4,$$

$$\|x - (d/N)L^* q\| = 0.085565$$

Example 11. For this example, we used the real Harmonic frame for $\mathbb{R}^4$. The other parameters we used are $\lambda = 2^{N/d}$, $n = 100$, $c = 1$, $m = 4$. We quantized the vector $x \in \mathbb{R}^4$

$$x = (-0.046816, 0.96742, 0.8447, 0.12239)$$

for every $N = 20, 21, \ldots, 120$. Figure 3.9 shows how the quantization error behaves as N increases.

For $N = 120$, we obtained the following values for each of the components of the decomposition given in Theorem 34. Figure 3.10 shows how the sequence $(y_k - q_k)_{k=1}^N$ behaves.

$$\begin{aligned} \|x_\lambda\| &= 5.12241e-009 \\ \|x_{ns}\| &= 3.3921e-004 \\ \|x_J\| &= 0.021658 \end{aligned}$$

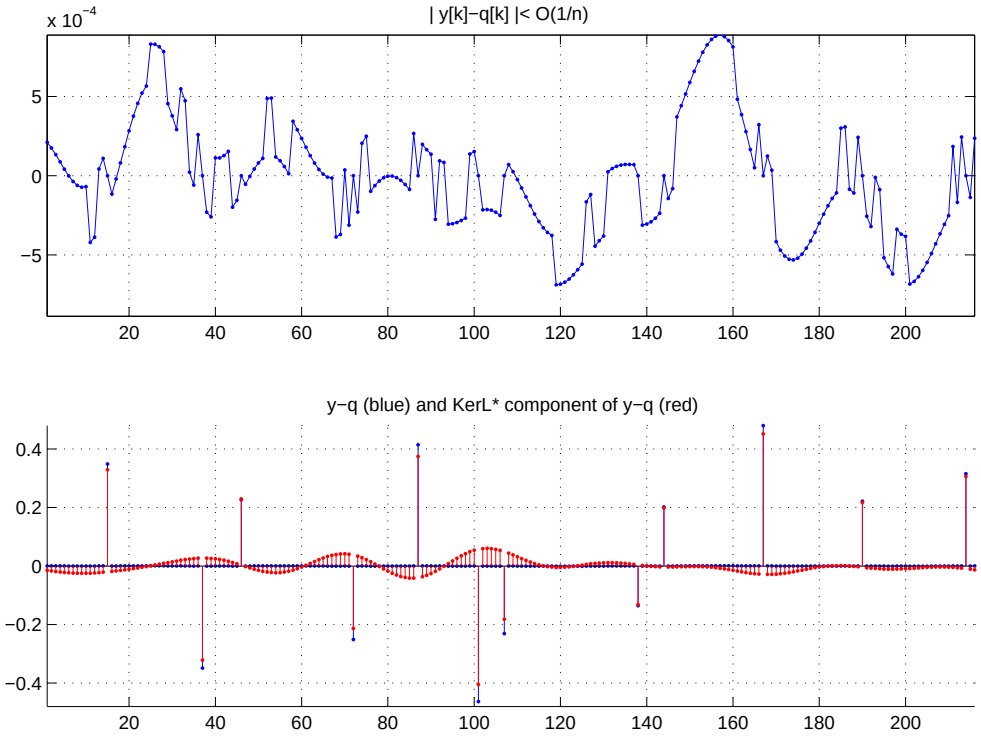

Figure 3.7: $N = 216$ in Example 10. $|J| = 12$

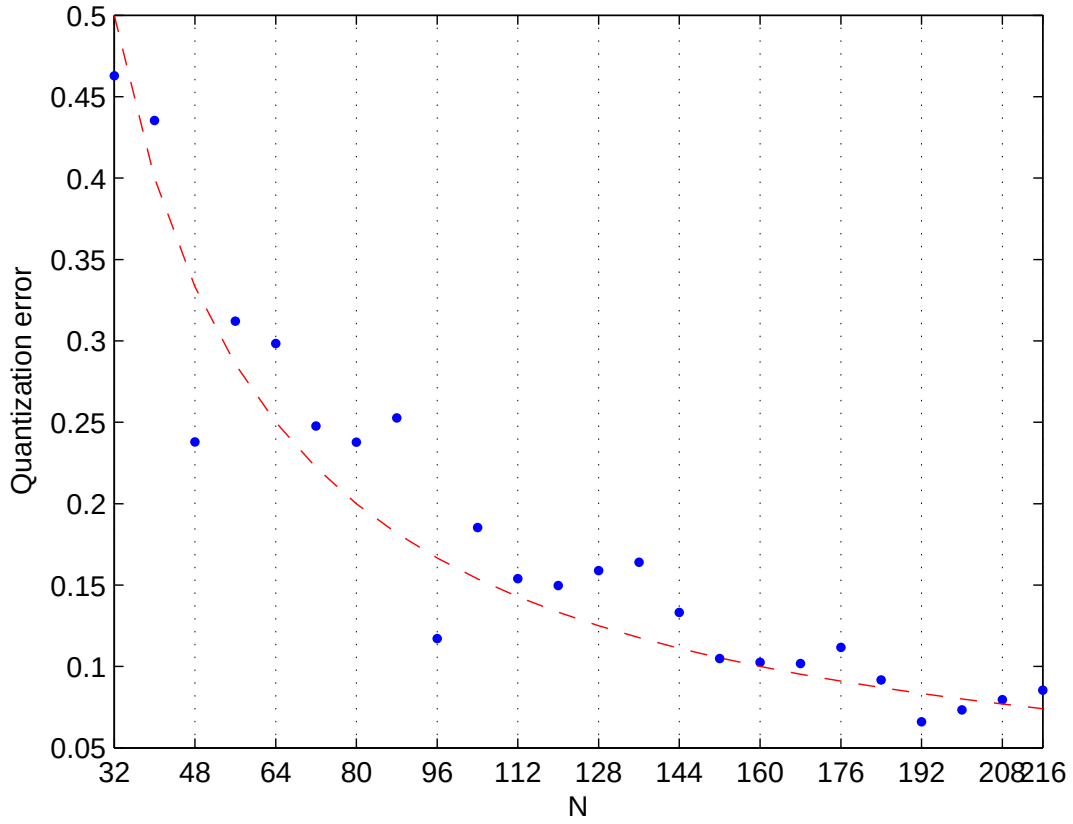

Figure 3.8: The quantization error for various values of N in Example 10. Dots represent the values of quantization error, and the dashed line is the curve $y = d/N$.

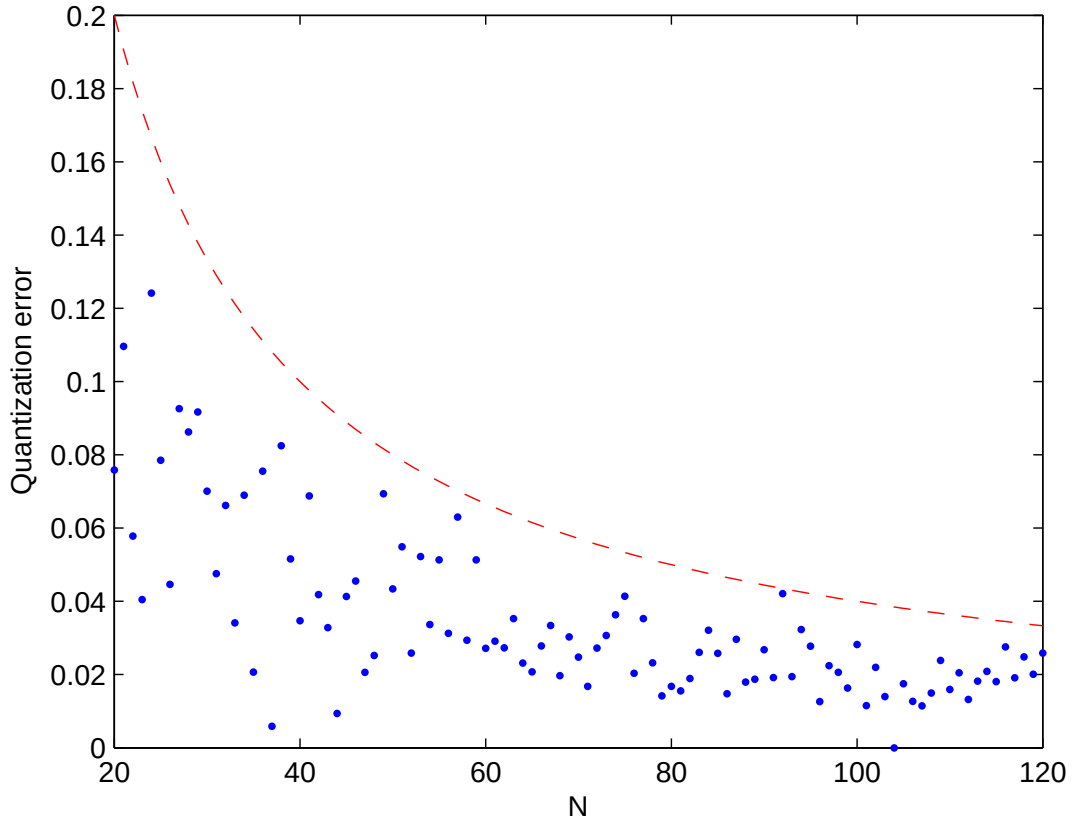

Figure 3.9: The quantization error for various values of N in Example 11. Dots represent the values of quantizatin error, and the dashed line is the curve $y = d/N$.

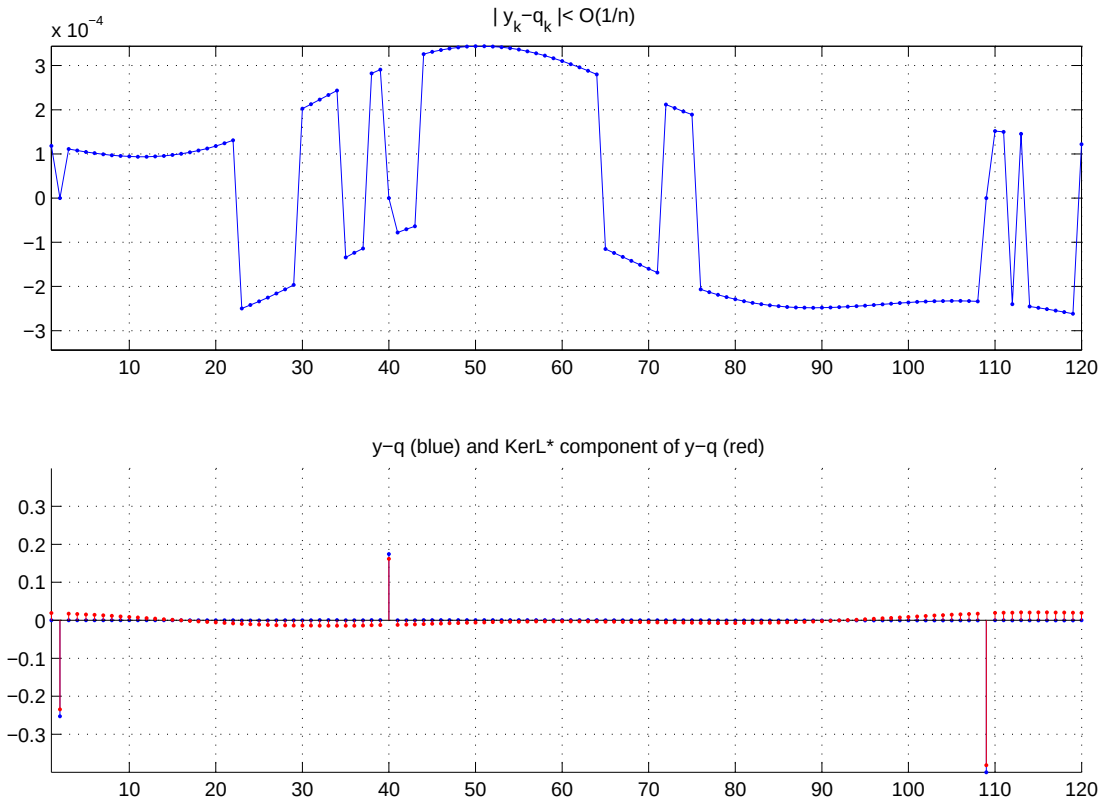

Figure 3.10: $N = 120$ in Example 11. $J = \{2, 40, 109\}$

$$card\{k \in J : q_k = 0\} = 0,$$

$$\|x - (d/N)L^*q\| = 0.0217534$$

For $N = 70$, we obtained the following values for each of the components of the decomposition given in Theorem 34. Figure 3.11 shows how the sequence $(y_k - q_k)_{k=1}^N$ behaves.

$$\|x_\lambda\| = 2.4368e - 006$$

$$\|x_{ns}\| = 9.8816e - 004$$

$$\|x_J\| = 0.024926$$

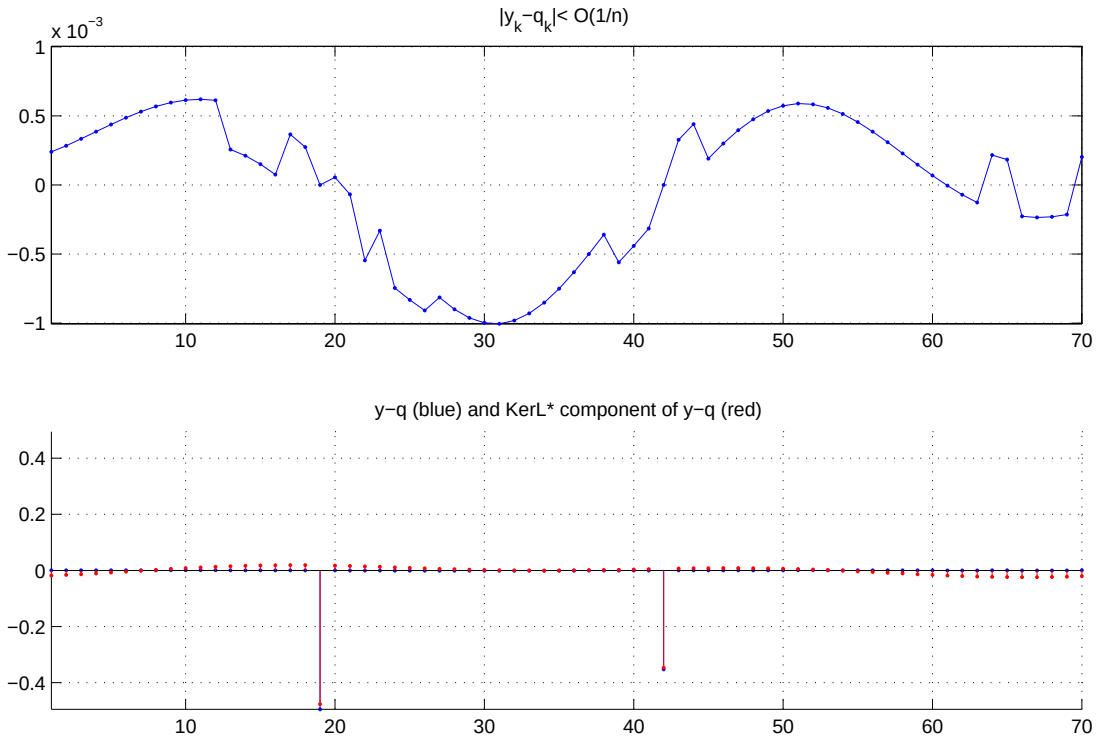

Figure 3.11: $N = 70$ in Example 11. $J = \{19, 42\}$

$$card\{k \in J : q_k = 0\} = 1,$$

$$\|x - (d/N)L^*q\| = 0.0249428$$

Example 12. For this example, we used the eleventh roots of unity frame for $\mathbb{R}^2$. $d = 2$, $N = 11$, and the parameters we used are $\lambda = 2^{N/d}$, $n = 100$, $c = 1$, $m = 4$.

We quantized each point in the regular grid

$$G = \{x = (x_1, x_2) : x_1, x_2 = -1, -0.9, \ldots, 0.9, 1\}$$

In Figure 3.12 shows the quantization error for every point in the grid. Considering G as a matrix, we stacked the columns of G, and plotted G versus the quantization

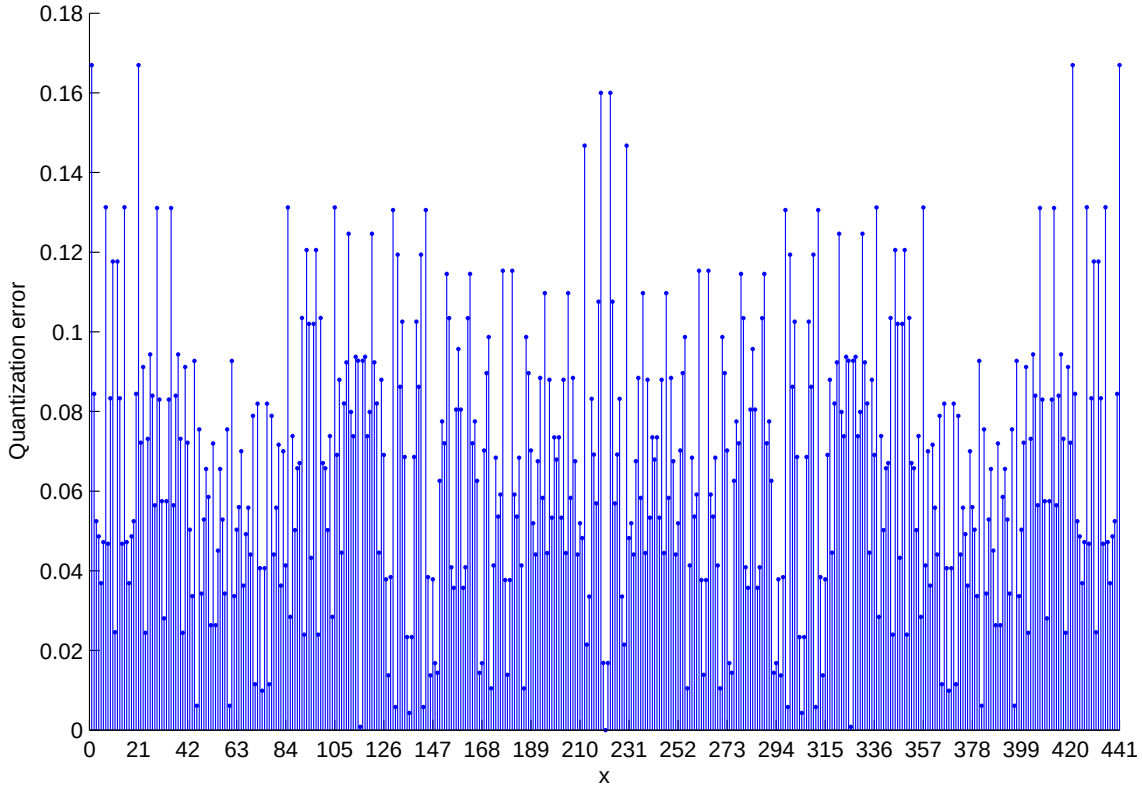

Figure 3.12: Plot of quantization errors for 441 points given in Example 12. The Average Noise is equal to 0.0665, and the Average Noise-Squared is equal to 0.0056

error. Therefore (x_1, x_2) in Figure 3.12 corresponds to $k = 11(10+10x_1)+11+10x_2$ on the horizontal axis.

Example 13. $d = 2$, $N = 10$, $\lambda = 2^{N/d}$, $n = 100$, $c = 1$, $m = 4$. The rows of the matrix L constitute a unit-norm tight frame for $\mathbb{R}^2$:

$$L = \begin{pmatrix} -0.26753 & -0.96355 \\ -0.25355 & -0.96732 \\ -0.67101 & -0.74145 \\ -0.81442 & -0.58028 \\ -0.97042 & -0.24142 \\ -0.99797 & 0.06367 \\ -0.8892 & 0.45752 \\ -0.73249 & 0.68078 \\ -0.64949 & 0.76037 \\ -0.25279 & 0.96752 \end{pmatrix}$$

We quantized each point in the regular grid

$$\{x = (x_1, x_2) : x_1, x_2 = -1, -0.9, \ldots, 0.9, 1\}$$

Figure 3.13 shows the quantization error for every point in the grid.

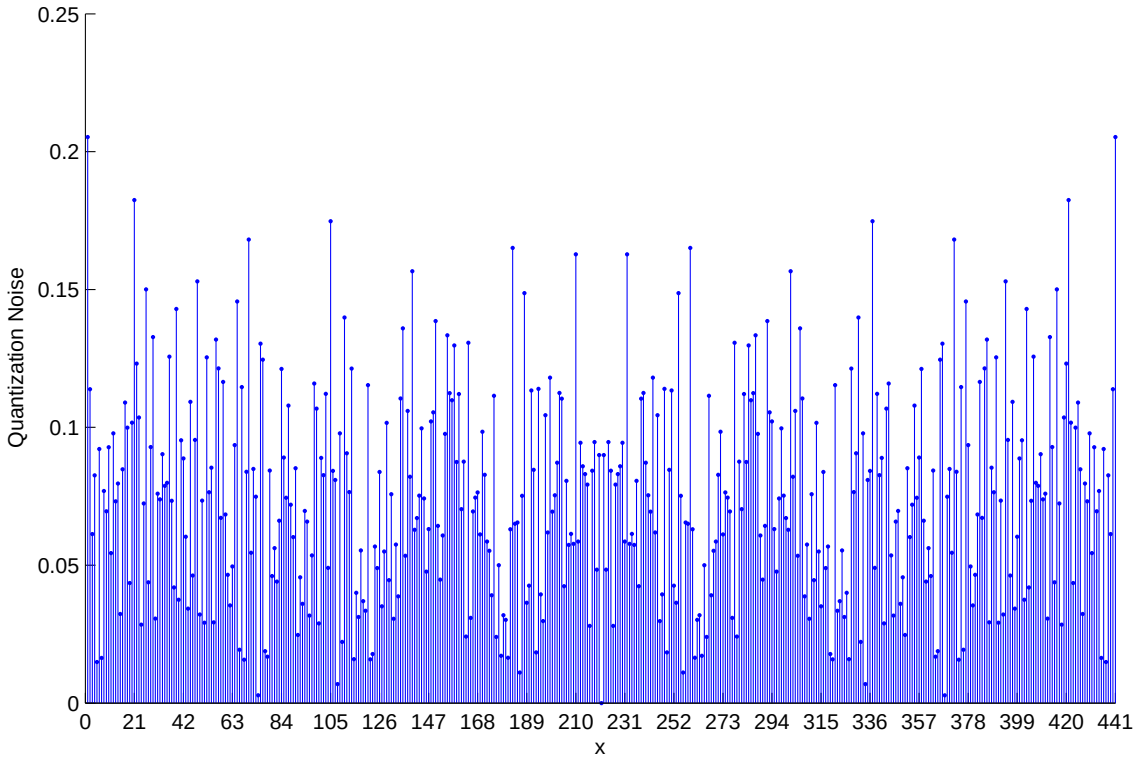

Figure 3.13: Plot of quantization errors for 441 points given in Example 13. The Average Noise is equal to 0.0757, and the Average Noise-Squared is equal to 0.0072.

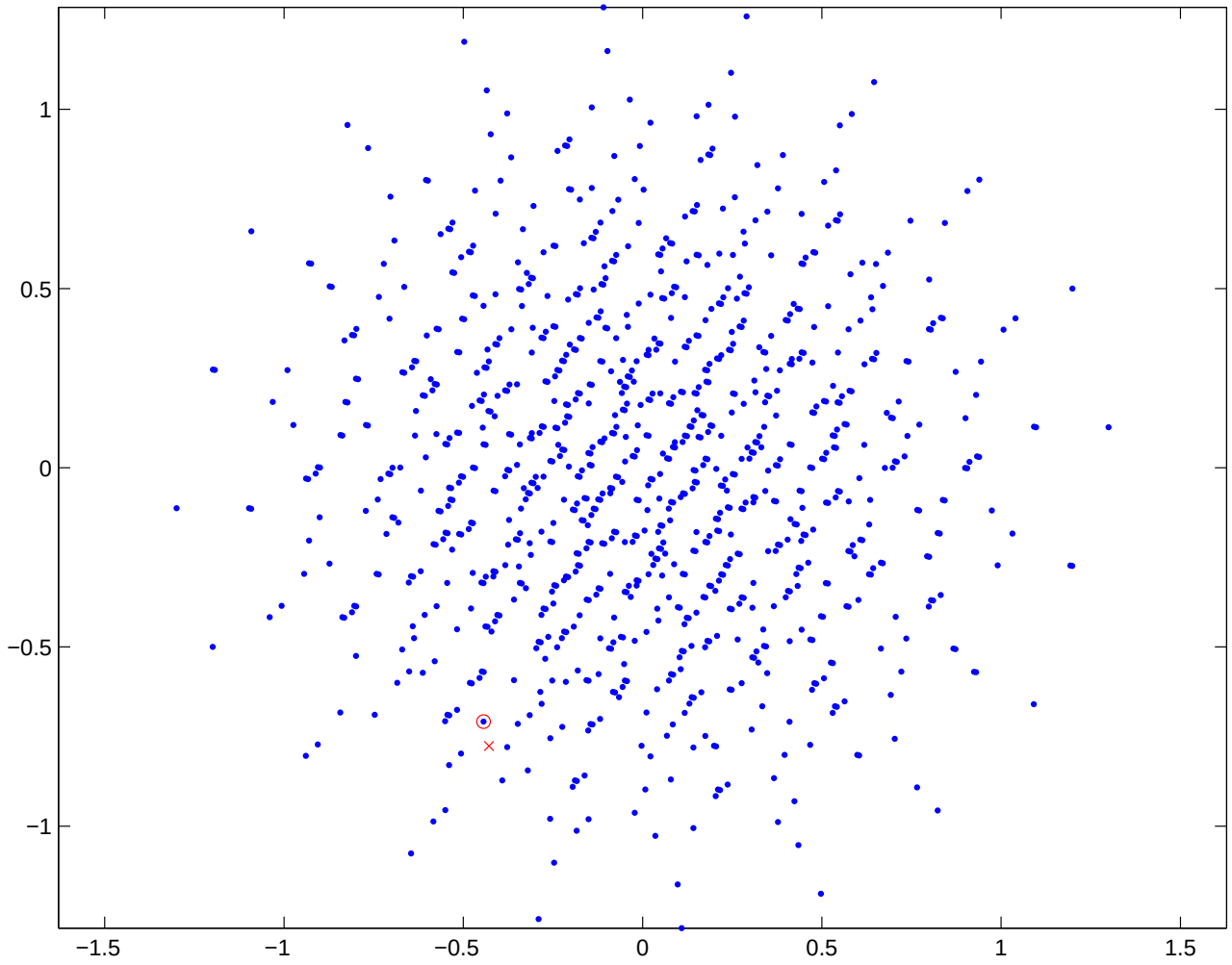

Figure 3.14: Plot of all linear combinations of the frame given in Example 13 with ± 1 coefficients. "o" represents the quantized estimate of "x".

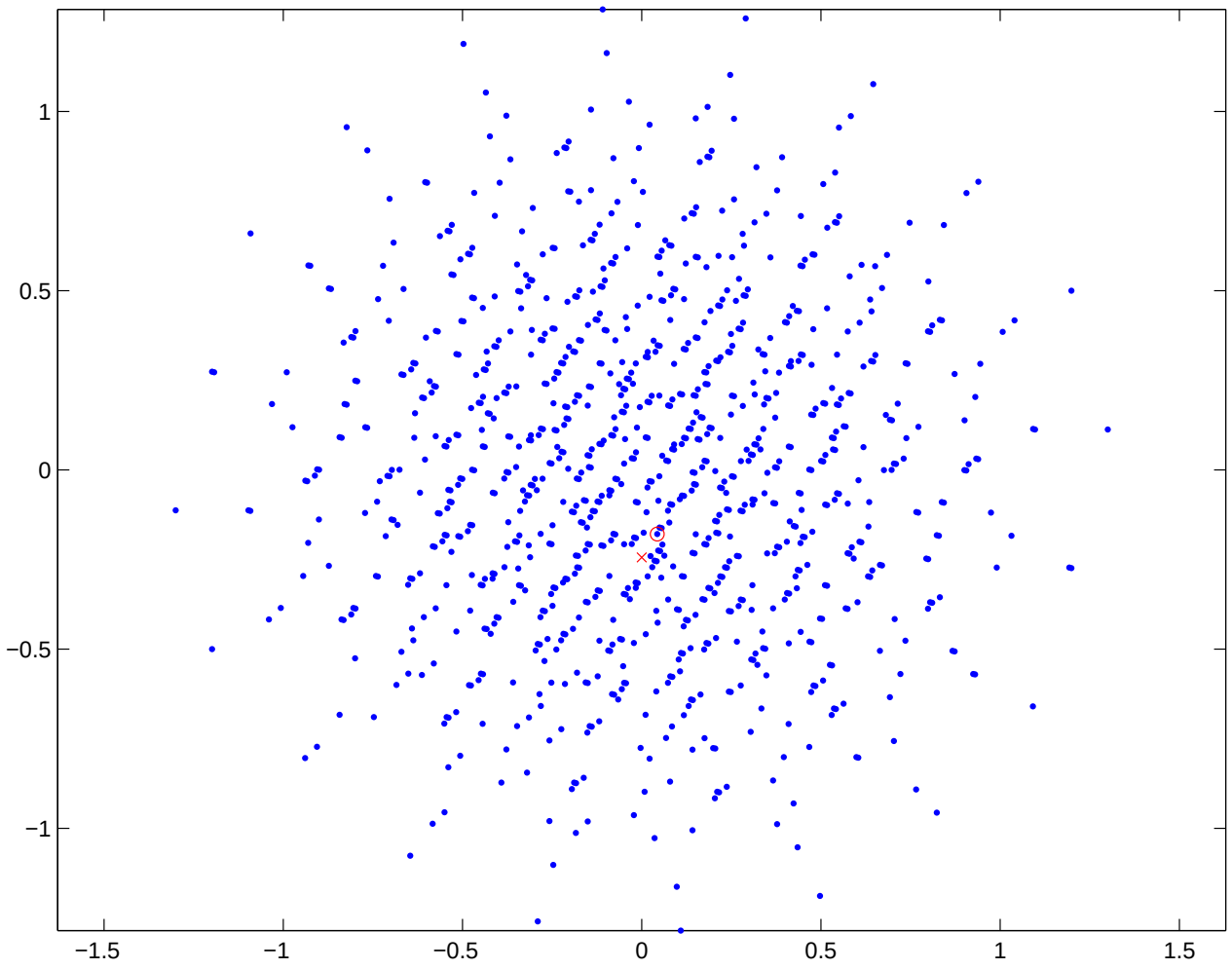

Figure 3.15: Plot of all linear combinations of the frame given in Example 13 with ± 1 coefficients. "o" represents the quantized estimate of "x".

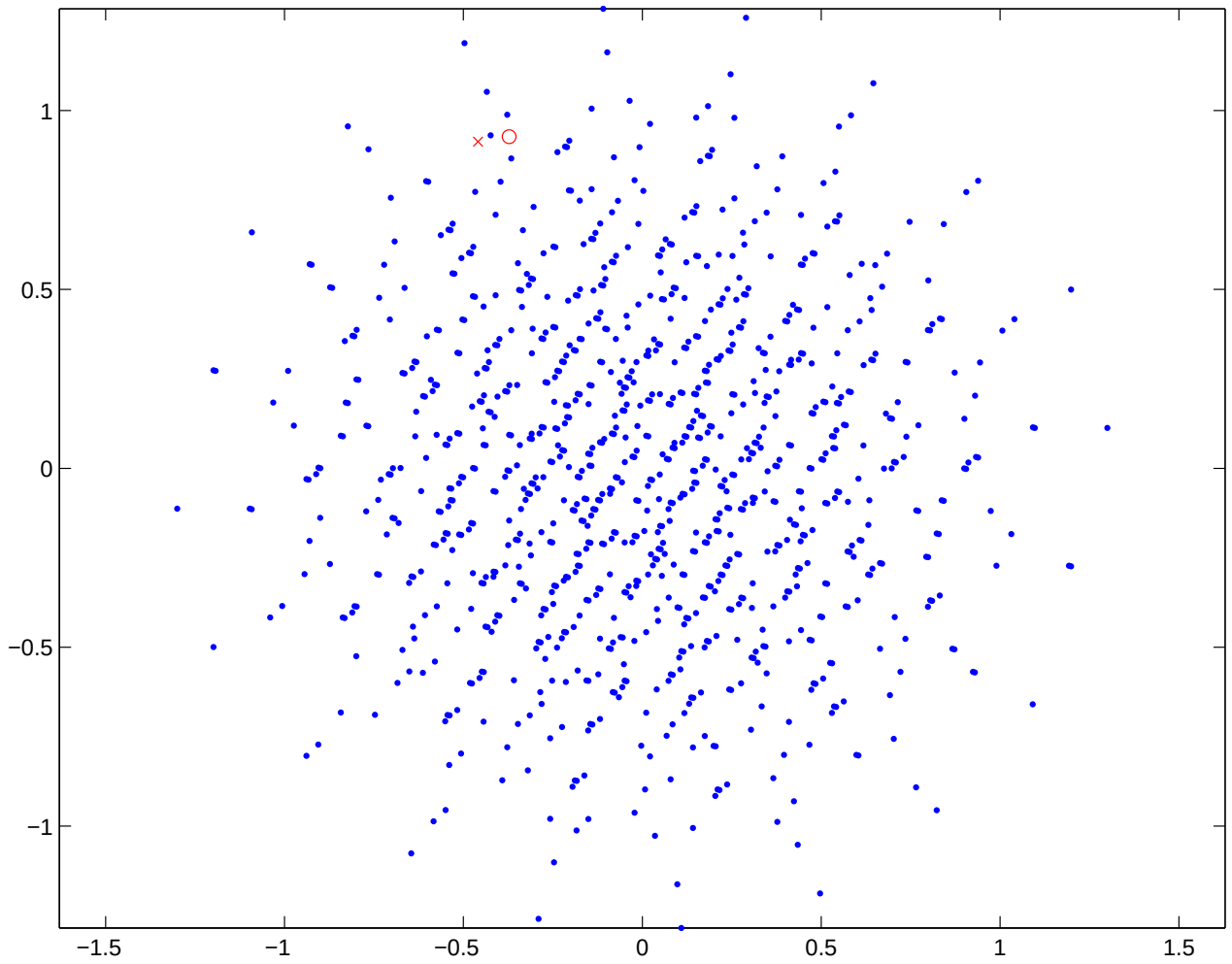

Figure 3.16: Plot of all linear combinations of the frame given in Example 13 with ± 1 coefficients. "o" represents the quantized estimate of "x".

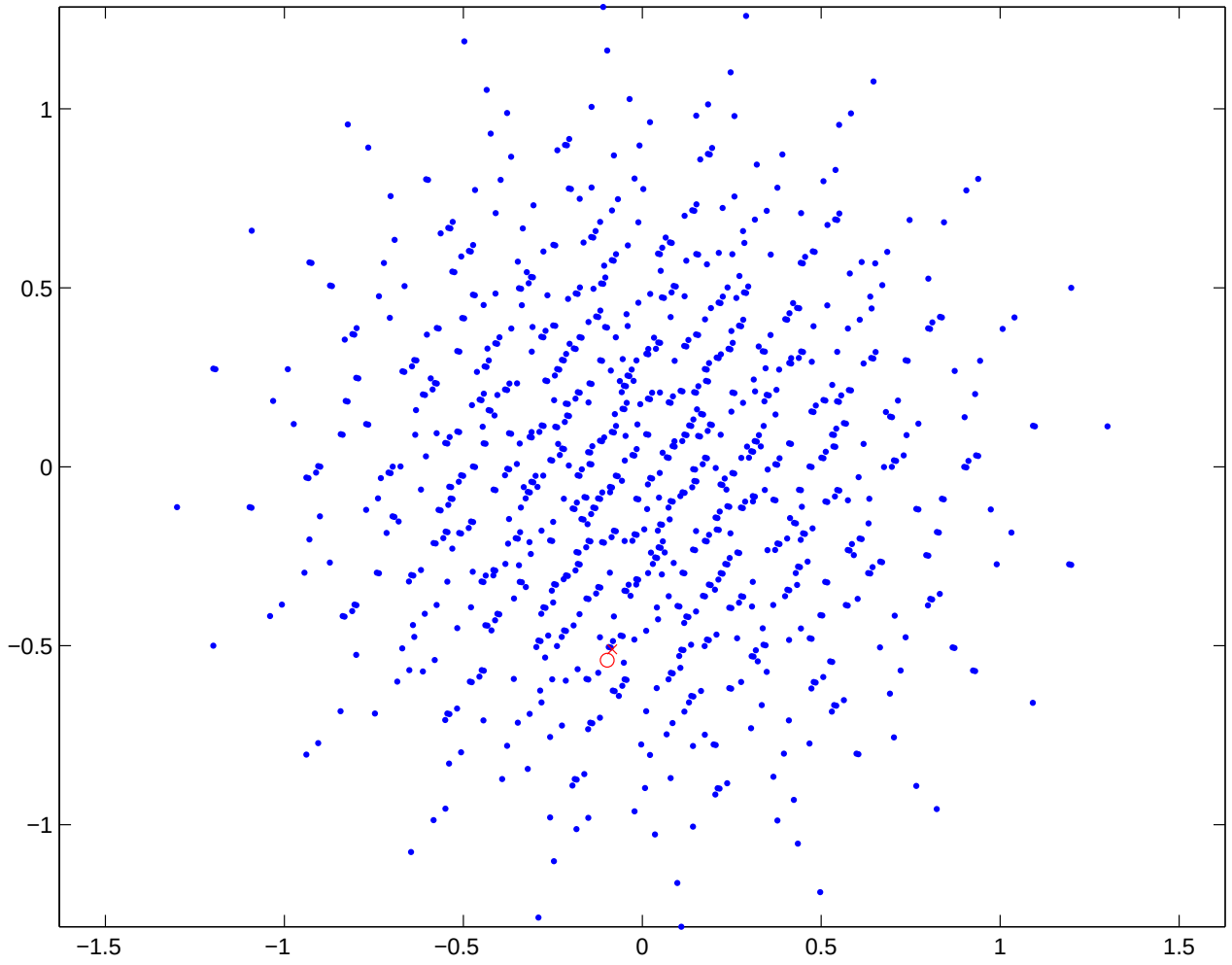

Figure 3.17: Plot of all linear combinations of the frame given in Example 13 with ± 1 coefficients. "o" represents the quantized estimate of "x".

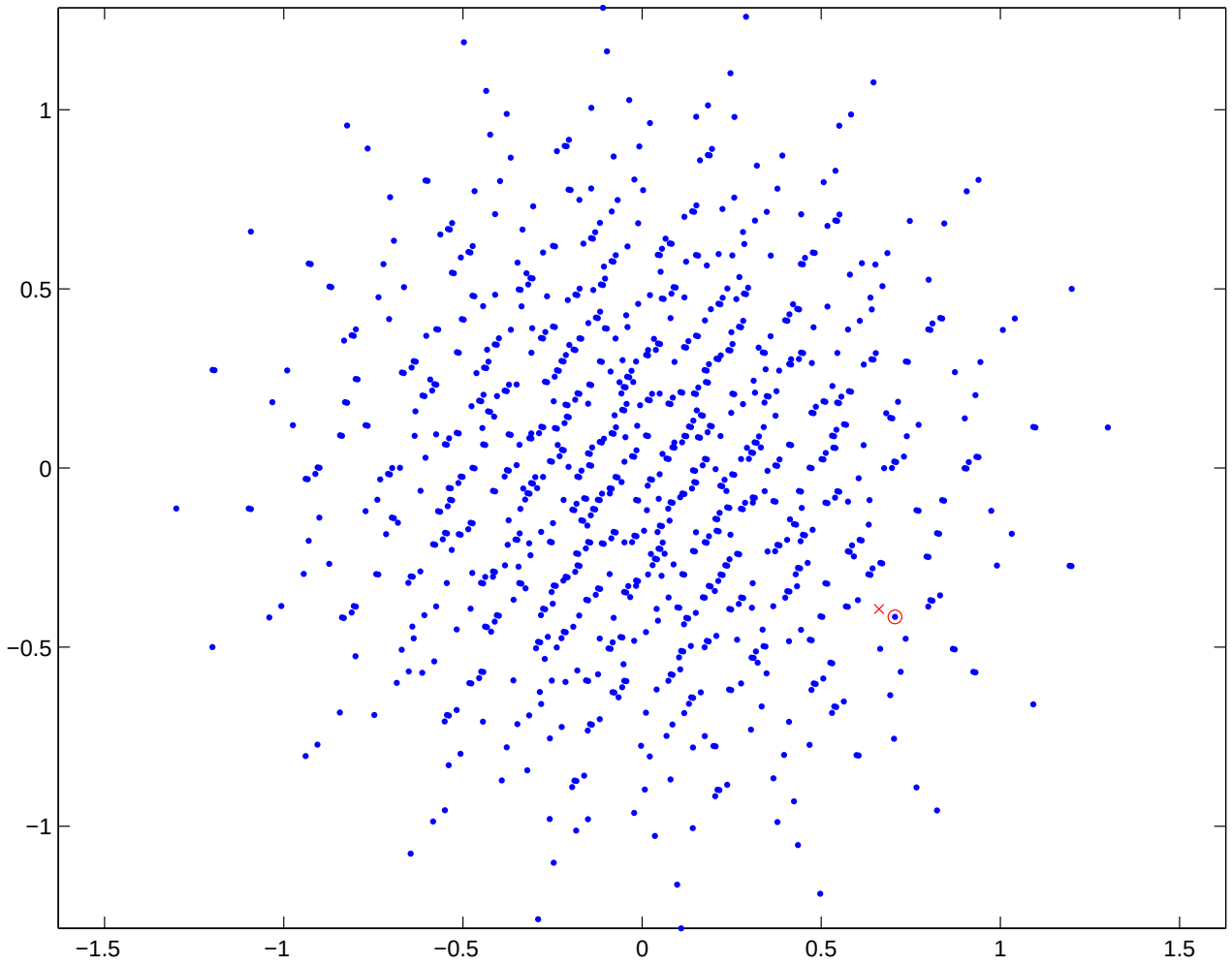

Figure 3.18: Plot of all linear combinations of the frame given in Example 13 with ± 1 coefficients. "o" represents the quantized estimate of "x".

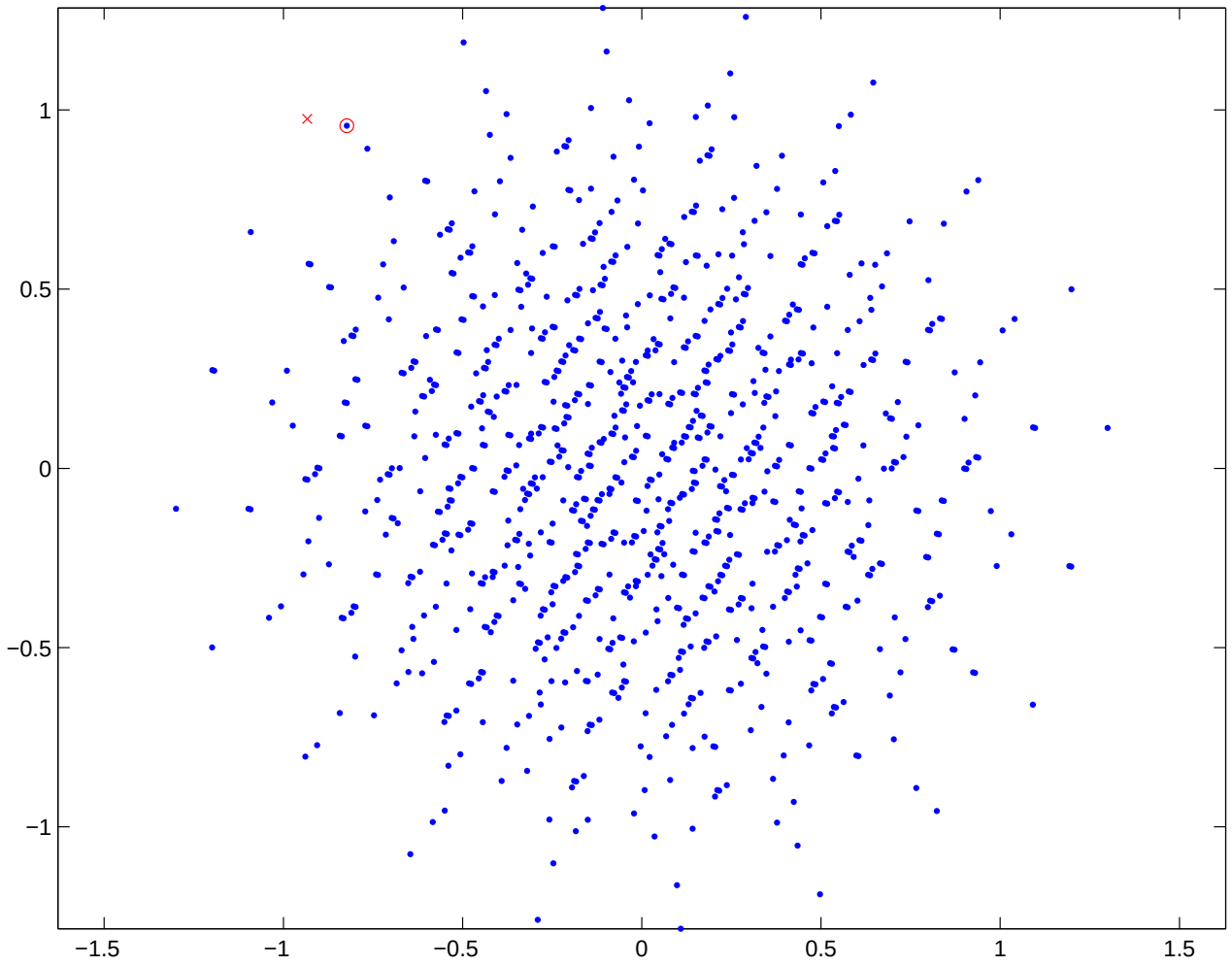

Figure 3.19: Plot of all linear combinations of the frame given in Example 13 with ± 1 coefficients. "o" represents the quantized estimate of "x".

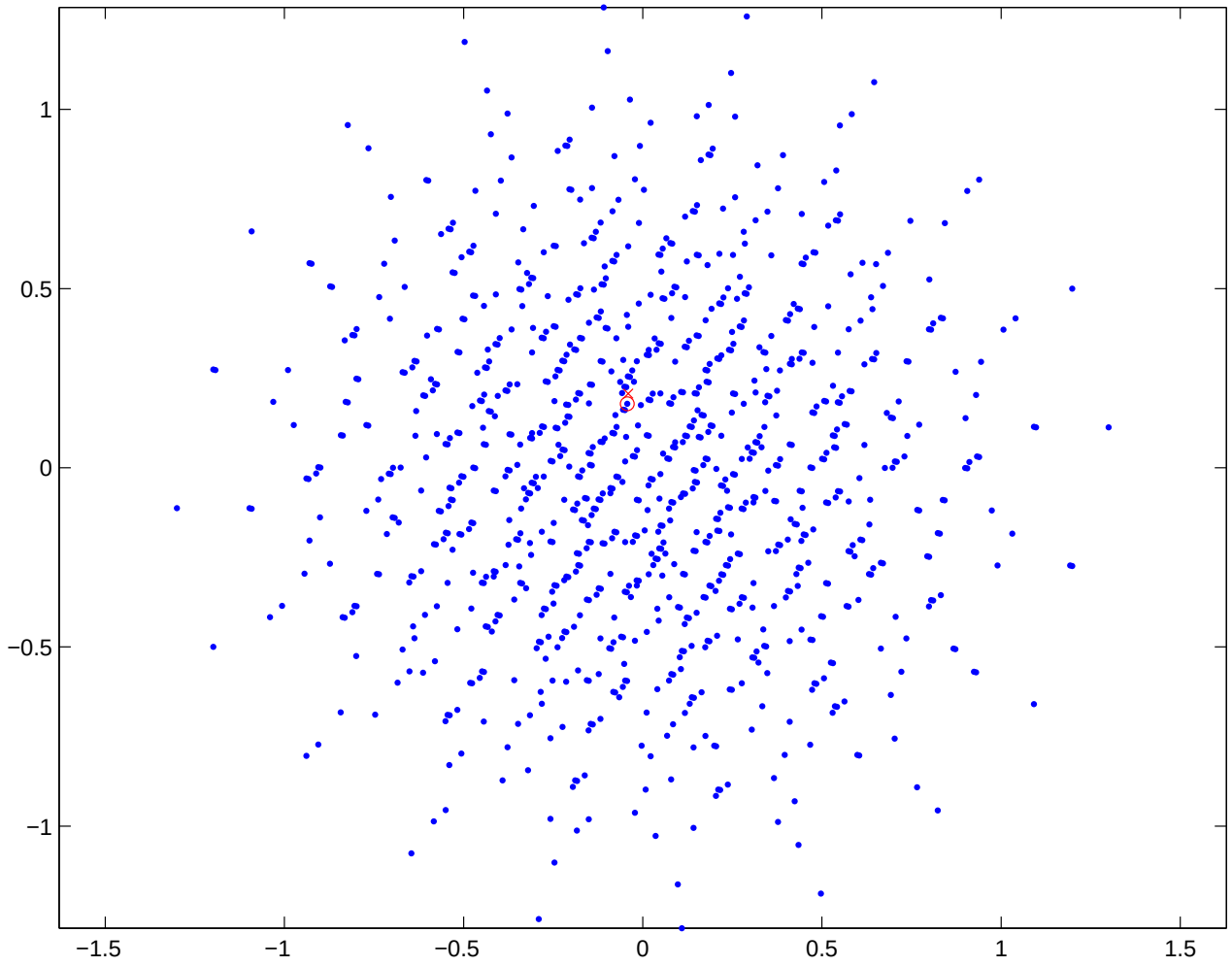

Figure 3.20: Plot of all linear combinations of the frame given in Example 13 with ± 1 coefficients. "o" represents the quantized estimate of "x".

Chapter 4

Equiangular Tight Frames

Equiangular tight frames have arisen in different areas of pure and applied mathematics. For instance, equiangular tight frames are shown to be optimal configurations for the Grassmanian line packing problem [18, 19]. It was shown in [65] and [49] that equiangular tight frames are spherical designs, which are used for fast numerical integration of certain polynomials on the sphere. Holmes and Paulsen [49] and Heath and Strohmer [69] proved that equiangular tight frames minimize the error due to erasures in communications. Tropp, Dhillon, Heath and Strohmer [72, 71] showed that equiangular tight frames have potential application for CDMA systems in wireless communications, and they provided an algorithm to design such equiangular signature sequences.

Definition 16. A unit norm frame (not necessarily tight) $\{x_i\}_{i=1}^N \subseteq \mathbb{F}^d$ ($\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$) is *equiangular* if

$$\exists c > 0 \quad \text{such that} \quad \forall i \neq j \quad |\langle x_i, x_j \rangle| = c.$$

If, in addition, $\{x_i\}_{i=1}^N$ is tight, then it is called an *equiangular tight frame*.

It is known that equiangular tight frames always exist when $N = d$ and $N = d + 1$. When $N = d$, equiangular tight frames are orthonormal bases. When $N = d+1$, an equiangular tight frame is given by the vertices of the regular simplex

in $\mathbb{R}^d$. This particular frame can be characterized by its Grammian matrix LL^* as

$$LL^* = I + \frac{1}{d}(I - \mathbb{1})$$

where $\mathbb{1}$ is the matrix of all ones. Other than these trivial cases, Paulsen and Bodmann [13] showed that Hadamard matrices induce equiangular tight frames for $\mathbb{R}^d$. They also show that we can construct equiangular tight frames for $\mathbb{R}^d$ using graphs with certain special structure, which are *regular two-graphs*.

Equiangular tight frames with a prescribed redundancy do not always exist. The problem of determining for which values of N an equiangular tight frame $\{e_n\}_{n=1}^N$ exists for $\mathbb{F}^d$ is still an unsolved problem. There are many necessary conditions in the literature for the existence of equiangular tight frames for a given pair (d, N). However, no sufficient condition has been established, yet.

Table 4.1, which is taken from [73] shows whether an equiangular frame exists for several pairs (d, N). Also, a list of (d, d^2) equiangular tight frames for several values of d can be found in [65].

In Section 4.1.1, we discuss the nonexistence results in the literature. In Section 4.1.2 we shall describe the numerical method presented in [73], and in Section 4.1.3 we shall talk about the theorem given in [65], which establishes a connection between the spherical designs and the equiangular tight frames. And then, we shall briefly talk about how equiangular tight frames are related to the erasure and the line packing problems in Section 4.1.4 and Section 4.1.6, respectively. Finally, we shall present our results in Section 4.2. Also, we shall talk about the relation between the new results given in Section 4.2.2 and the main result of [14].

4.1 Known Results in the Literature, and Relations to Other Problems

4.1.1 Nonexistence Results

Equiangular tight frames do not exist for certain pairs (d, N). The following theorems rule out many pairs (d, N), and provide necessary conditions for equiangular tight frames to exist.

Theorem 35. Equiangular tight frames with N elements for $\mathbb{F}^d$ can exist only if

$$\begin{aligned} N &\leq d(d+1)/2 \quad \text{if} \quad \mathbb{F} = \mathbb{R}, \\ N &\leq d^2 \quad \text{if} \quad \mathbb{F} = \mathbb{C}. \end{aligned}$$

Different proofs of this theorem can be found in [70] and [19], as well as in the proof of Theorem 51.

Theorem 36. Let $\{x_i\}_{i=1}^N$ be a unit norm tight frame for $\mathbb{F}^d$, and let L be the associated Bessel map. Then,

$$\max_{i \neq j} |\langle x_i, x_j \rangle| \geq \sqrt{\frac{N-d}{d(N-1)}} =: c_{N,d}.$$

Moreover, this lower bound is attained if and only if $\{x_i\}_{i=1}^N$ is equiangular.

Proof. The result follows from

$$\sum_{i \neq j} |\langle x_i, x_j \rangle|^2 = trace((LL^*)^2) - N = \frac{N}{d} trace(LL^*) - N = \frac{N^2}{d} - N = \frac{N(N-d)}{d}.$$

□

Theorem 37. If $\{x_i\}_{i=1}^N$ is an equiangular tight frame for $\mathbb{R}^d$, and L is the associated Bessel map, then

$$Q = \frac{1}{c_{N,d}} \left(LL^* - \frac{N}{d} I \right)$$

is a matrix with zeros on the diagonal and ± 1 on the off diagonal entries, and has exactly two eigenvalues

$$\lambda_1 = -\sqrt{\frac{d(N-1)}{N-d}},\ \lambda_2 = \sqrt{\frac{(N-d)(N-1)}{d}}$$

with multiplicities $N-d$ and d respectively. Moreover, if $N \neq 2d$ and $N \neq d+1$, then λ_1, λ_2 are odd integers.

Proof of Theorem 37 can be found in [73].

The cases listed in Table 4.1, for which an equiangular tight frame does not exist, can actually be verified by the Theorems 35, 36 and 37.

4.1.2 Numerical Computation

Tropp, Dhillon, Heath and Strohmer [73] developed a numerical method, with which they computed equiangular tight frames for $\mathbb{F}^d$ for several values of d and N. They translate the problem of finding equiangular tight frames to an inverse eigenvalue problem. They construct $N \times N$ matrices G subject to the constraints

i. *Structural constraint:* $G = (g_{ij})$ is a self adjoint matrix that has 1s on the diagonal entries, and

$$|g_{ij}| \leq \sqrt{\frac{N-d}{d(N-1)}} \quad \forall i \neq j.$$

ii. *Spectral constraint:* G has precisely two distinct eigenvalues, N/d with multiplicity d, and 0 with multiplicity $N - d$.

If G satisfies both of the constraints above, then it is the Grammian of an equiangular tight frame. In fact, let

$$G = UDU^*$$

be a singular value decomposition of G, where U is unitary, and D is a diagonal matrix with first d diagonal entries are 1 and the rest are zero. Then, if we take first d columns of U and form a new $N \times d$ matrix L, the rows of L gives an equiangular tight frame. Moreover, we would have $LL^* = G$.

These two constraints above induce two sets, per se,

$$\begin{aligned} A &= \{G : \text{Structural constraint}\}, \\ B &= \{G : \text{Spectral constraint}\}. \end{aligned}$$

Both sets are compact, with respect to the Frobenius norm in the space of all $N \times N$ matrices. Also, A is convex, while B is not.

The projections onto A and B are defined in Theorem 38 and Theorem 39, respectively. Proofs of these theorems are in [73].

Theorem 38. Let $Z = (z_{ij})$ be an $N \times N$ self adjoint matrix. Then, the closest unique matrix $H = (h_{ij}) \in A$ to Z in Frobenius norm is given by $h_{ii} = 1$, and

$$h_{ij} = \begin{cases} z_{ij}, & \text{if} \quad |z_{ij}| \leq \sqrt{\frac{N-d}{d(N-1)}}, \\ \frac{N-d}{d(N-1)} \frac{z_{ij}}{|z_{ij}|}, & \text{otherwise.} \end{cases}$$

Theorem 39. Let Z be an $N \times N$ self adjoint matrix with a unitary factorization UDU^{-1} where the entries of D are arranged in a non-increasing order. Let L be the $N \times d$ matrix formed using the first d columns of U. Then, $\frac{N}{d}LL^*$ is the closest matrix in B to Z with respect to the Frobenius norm. This matrix is unique if the eigenvalues of Z are strictly decreasing.

When both sets are convex, the alternating projections converge. Theorem 40 summarizes this result. However, if one of the sets is not convex, the alternating projections algorithm may fail to converge. Theorem 41 is taken from [73], and it describes scenarios that can take place for this particular alternating projections problem we consider in this subsection.

Theorem 40. Let A, B be two compact convex subsets of a Hilbert space $\mathcal{H}$. Define the projections $P_A(x) = argmin_{a \in A}||x - a||$, $P_B(x) = argmin_{b \in B}||x - b||$. For any starting point x_1, define the sequences

$$x_{i+1} = P_A(y_i), \quad y_i = P_B(x_i).$$

Then,

$$\exists x, y \in \mathcal{H}, \quad \text{such that} \quad x_i \to x, \quad \text{and} \quad y_i \to y,$$

and

- if $A \cap B = \emptyset$, then $\min\{||a - b|| : a \in A, b \in B\} = ||x - y||$
- if $A \cap B \neq \emptyset$, then $x = y \in A \cap B$.

Theorem 41. Assume that the Alternating Projection between $\mathbb{A}$ and $\mathbb{B}$ generates a sequence of iterates (G_i, H_i), and suppose there is $J \in \mathbb{N}$ such that $||G_J - H_J||_F <$

$N/(d\sqrt{2})$. The sequence of iterates possesses at least one accumulation point $(\bar{G}, \bar{H})$. Then,

i. Every accumulation point lies in $A \times B$.

ii. $(\bar{G}, \bar{H})$ is a fixed point of the alternating projection algorithm, i.e., if we start with $(\bar{G}, \bar{H})$, then, every iterate would be equal to $(\bar{G}, \bar{H})$.

iii. Every accumulation point satisfies

$$||\bar{G} - \bar{H}||_F = \lim_{j \to \infty} ||G_j - H_j||_F.$$

iv. The component sequences are asymptotically regular, i.e.,

$$||G_{j+1} - G_j||_F \to 0 \quad \text{and} \quad ||H_{j+1} - H_j||_F \to 0$$

v. Either the component sequences both converge,

$$||G_j - \bar{G}||_F \to 0 \quad \text{and} \quad ||H_j - \bar{H}||_F \to 0,$$

or the set of accumulation points forms a continuum.

4.1.3 Spherical t-designs

Spherical designs are the spherical analogues of Gaussian quadrature sequences. A set $\{x_i\}_{i=1}^N$ of unit norm vectors in $\mathbb{F}^d$ is a *spherical t-design* if the integral on the surface of the unit sphere S^{d-1} of any polynomial f of degree at most t is equal to the average value of the polynomial evaluated at $\{x_i\}_{i=1}^N$, i.e.,

$$\frac{1}{|S^{d-1}|} \int_{S^{d-1}} f d\sigma = \frac{1}{N} \sum_{i=1}^{N} f(x_i).$$

	d				
N	2	3	4	5	6
3	$\mathbb{R}$	$\mathbb{R}$	..	..	..
4	$\mathbb{C}$	$\mathbb{R}$	$\mathbb{R}$	..	..
5	..	.	$\mathbb{R}$	$\mathbb{R}$	..
6	..	$\mathbb{R}$	.	$\mathbb{R}$	$\mathbb{R}$
7	..	$\mathbb{C}$	$\mathbb{C}$	.	$\mathbb{R}$
8	..	.	$\mathbb{C}$	.	.
9	..	$\mathbb{C}$	.	.	$\mathbb{C}$
10	..	..	.	$\mathbb{R}$	.
11	..	..	.	$\mathbb{C}$	$\mathbb{C}$
12	..	..	.	.	$\mathbb{C}$
13	..	..	$\mathbb{C}$	.	.
14	..	..	.	.	.

	d				
N	2	3	4	5	6
15	..	..	.	.	.
16	..	..	$\mathbb{C}$	.	$\mathbb{R}$
17	..	..	..	.	.
18	..	..	..	.	.
19	..	..	..	.	.
20	..	..	..	.	.
21	..	..	..	$\mathbb{C}$	.
22	..	..	..	.	.
23	..	..	..	.	.
24	..	..	..	.	.
25	..	..	..	$\mathbb{C}$	.
26	..	..	..	..	.

	d				
N	2	3	4	5	6
27	..	..	..	..	.
28	..	..	..	..	.
29	..	..	..	..	.
30	..	..	..	..	.
31	..	..	..	..	$\mathbb{C}$
32	..	..	..	..	.
33	..	..	..	..	.
34	..	..	..	..	.
35	..	..	..	..	.
36	..	..	..	..	$\mathbb{C}$

Table 4.1: $\mathbb{R}$ and $\mathbb{C}$ indicate that the alternating projection method was able to compute an equiangular tight frame for $\mathbb{R}^d$ and $\mathbb{C}^d$, respectively. Every equiangular tight frame for $\mathbb{R}^d$ is automatically an equiangular tight frame for $\mathbb{C}^d$, so $\mathbb{C}$ in turn indicates that there is no equiangular tight frame for $\mathbb{R}^d$. One period (.) means that no equiangular tight frame for $\mathbb{R}^d$ exists, and two periods (..) mean that no equiangular tight frame exists at all. [73]

where σ is the surface measure on S^{d-1}, which is invariant to unitary operations.

There are several characterizations of spherical designs. We give one of the caracterizations of spherical 2-designs in Theorem 42, and provide an original proof.

Theorem 42. A set $\{x_i\}_{i=1}^N$ of unit norm vectors in $\mathbb{F}^d$ is a spherical 2-design if and only if $\{x_i\}_{i=1}^N$ is a zero sum FUNTF.

Proof. Suppose $\{x_i\}_{i=1}^N$ is a spherical 2-design, and let

$$Tx = \frac{1}{|S^{d-1}|}\int_{S^{d-1}} \langle x, y\rangle y d\sigma(y).$$

It is not hard to check that T is linear, and commute with every unitary matrix. As a result, $T = \lambda I$, for some constant λ.

Since $\{x_i\}_{i=1}^N$ is a spherical 2-design, for every $x \in \mathbb{R}^d$

$$\lambda||x||^2 = \langle Tx, x\rangle = \frac{1}{|S^{d-1}|}\int_{S^{d-1}} \langle |\langle x, y\rangle|^2\rangle d\sigma(y) = \frac{1}{N}\sum_{i=1}^N |\langle x, x_i\rangle|^2.$$

Therefore, $\{x_i\}_{i=1}^N$ is a FUNTF, and also $\lambda = 1/d$. Second, by symmetry,

$$\int_{S^{d-1}} y d\sigma(y) = 0.$$

Therefore, for every $x \in \mathbb{R}^d$,

$$\frac{1}{N}\langle \sum_{i=1}^N x_i, x\rangle = \frac{1}{|S^{d-1}|}\int_{S^{d-1}} \langle x, y\rangle d\sigma(y) = 0.$$

Thus, $\sum_{i=1}^N x_i = 0$.

Conversely, suppose $\{x_i\}_{i=1}^N$ is a zero sum FUNTF for $\mathbb{F}^d$. Every degree ≤ 2 polynomial is of the form

$$f(x) = a + \langle x, b\rangle + \langle Ax, x\rangle.$$

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N} a &= a = \frac{1}{|S^{d-1}|}\int_{S^{d-1}} a d\sigma(y), \\
\frac{1}{N}\sum_{i=1}^{N}\langle x_i, b\rangle &= 0 = \frac{1}{|S^{d-1}|}\int_{S^{d-1}} \langle y, b\rangle d\sigma(y),
\end{aligned}$$

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N}\langle Ax_i, x_i\rangle &= trace(A\frac{1}{N}S) \\
&= trace(A)/d \\
&= trace(AT) \\
&= \frac{1}{|S^{d-1}|}\int_{S^{d-1}} \langle Ay, y\rangle d\sigma(y).
\end{aligned}$$

Therefore, $\{x_i\}_{i=1}^{N}$ is a spherical 2-design. □

Blume-Kohout, Scott, Caves and Renes [65] generalized the concept of spherical designs to $\mathbb{C}^d$. Analogous to the sphere S^{d-1} in $\mathbb{R}^d$, they define

$$\mathbb{S}^{d-1} = \{x \in \mathbb{C}^d : \|x\| = 1\}.$$

They also define a measure σ on $\mathbb{S}^{d-1}$, which is invariant under unitary transformations. They proved Theorem 43 that links the (d, d^2) equiangular tight frames and the spherical 4-designs. The original theorem and its proof can be found in [65].

Theorem 43. A set $\{x_i\}_{i=1}^{d^2}$ of unit norm vectors is a spherical 4-design if and only if

$$\sum_{i,j=1}^{d^2} |\langle x_i, x_j\rangle|^4 = \frac{2d^3}{d+1}.$$

This value is the global minimum of $\sum_{i,j=1}^{d^2} |\langle x_i, x_j\rangle|^4$. Moreover, when this 4-design exists, we obtain an equiangular tight frame.

4.1.4 Optimal Frames for Erasures

The simplified scenario in communications is that a given data/signal is first encoded, then divided into packets, and then sent to another location through a channel. Simply, a channel is a medium between the two locations, the transmitter and the receiver.

Usually, the channel alters the data. A *noisy channel* adds a noise on the original data and a *lossy channel* erases some of the packets of the data in transit.

In our simplified scenario, $x \in \mathbb{C}^d$ represents a data vector. Given a finite unit-norm tight frame $\{x_i\}_{i=1}^N$ for $\mathbb{C}^d$, the frame coefficients $(\langle x, x_i \rangle)$ are the packets to be transmitted.

Tight frames are advantageous in this setting for many reasons including

i. Linear encoding/decoding,

ii. Noise reduction,

iii. Robustness to erasures.

Noise reduction is achieved by projecting the noise onto the range of the analysis operator. If we transmit the frame coefficients through a noisy channel, the channel will alter each frame coefficient $\langle x, x_i \rangle$ by adding a certain amount of noise η_i. We can represent the noise as $\eta = \{\eta_i\}_{i=1}^N$ in the vector form. Therefore, what we obtain at the receiver after reconstruction is

$$\frac{d}{N} \sum_{i=1}^{N} (\langle x, x_i \rangle + \eta_i) x_i = x + \frac{d}{N} \sum_{i=1}^{N} \eta_i x_i.$$

Therefore, the effect of the noise realized at the receiver is

$$\eta' := \frac{d}{N}\sum_{i=1}^{N}\eta_i x_i.$$

The noise vector η does not necessarily lie in the range of the analysis operator L of the frame. Therefore, we can think of η as a sum of an in-space component, which lies in the range $\mathcal{R}(L)$ of L, and an out-of-space component, which lies in the orthogonal complement of $\mathcal{R}(L)$. Orthogonal complement of $\mathcal{R}(L)$ is the Kernel (or Null space) of the synthesis operator L^*. Therefore, the out-of-space component of η vanishes during the reconstruction.

Theorem 44 shows how much noise reduction is possible when the channel noise is a zero mean uncorrelated random noise.

Theorem 44. Suppose $x \in \mathbb{F}^d$ ($\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$) is a data vector, $\{x_i\}_{i=1}^N$ is a FUNTF for $\mathbb{F}^d$, and suppose that $(\langle x, x_i\rangle)_{i=1}^N$ is sent from a transmitter to a receiver through a noisy channel. Suppose also that the channel noise is zero mean uncorrelated random noise with variance σ^2. Then, the sum of the variances of the noise realized at the receiver is $d^2\sigma^2/N$, i.e.,

$$E(||\eta'||^2) = \frac{d^2}{N^2}E(||\eta||^2) = \frac{d^2}{N}\sigma^2,$$

where $E(X)$ denotes the expected value of a random variable X.

Proof. The realized noise η' is given by

$$\eta' = \frac{d}{N}\sum_{i=1}^{N}\eta_i x_i.$$

Then,

$$||\eta'||^2 = \frac{d^2}{N^2}\sum_{i,j=1}^{N}\eta_i\eta_j\langle x_i, x_j\rangle.$$

Since η_i and η_j are uncorrelated for $i \neq j$, $E(\eta_i \eta_j) = 0$, and so

$$E(||\eta'||^2) = \frac{d^2}{N^2} \sum_{i,j=1}^{N} E(\eta_i \eta_j) \langle x_i, x_j \rangle = \frac{d^2}{N^2} \sum_{i=1}^{N} E(|\eta_i|^2) = \frac{d^2}{N} \sigma^2.$$

□

In the erasure problem, some of the packets, i.e., the frame coefficients, might be delayed for too long, or get lost inside of the channel. These lost or delayed packets/coefficients are *erasures*.

Tight frames are robust to erasures in the sense of Definition 17. One can perfectly reconstruct the data x from a subset of frame coefficients provided that the number of erasures is not too big. For instance, if J is the set of indices corresponding to the frame coefficients that are received, and if $\{x_i\}_{i \in J}$ still constitutes a frame, then, we can compute the dual of this frame, apply this dual to the data received, and reconstruct x accurately.

Definition 17. A frame $\{x_i\}_{i=1}^N$ is *robust* to m erasures if $\{x_i\}_{i \in J}$ is still a frame for any index set J with $|J| = N - m$.

Theorem 45. *Let $\{x_i\}_{i=1}^N$ be a FUNTF for $\mathbb{F}^d$, and $N > dm$. Then, $\{x_i\}_{i=1}^N$ is robust to m erasures.*

Proof. Let $|J| = N - m$. Then, for each $x \in \mathbb{F}^d$,

$$\frac{N}{d} ||x||^2 - \sum_{i \in J} |\langle x, x_i \rangle|^2 = \sum_{i \notin J} |\langle x, x_i \rangle|^2 \leq m ||x||^2.$$

Therefore,

$$(\frac{N}{d} - m) ||x||^2 \leq \sum_{i \in J} |\langle x, x_i \rangle|^2 \leq \frac{N}{d} ||x||^2.$$

Hence, $\{x_i\}_{i \in J}$ is still a frame with frame bounds $\frac{N}{d} - m$ and $\frac{N}{d}$ □

By Theorem 45, the perfect reconstruction is possible in the presence of up to $m < N/d$ erasures if we use a FUNTF. However, even though the perfect reconstruction is possible, computing the duals of frames $\{x_i\}_{i\in J}$ might not be preferred for every application. We might have time constraints or scarce resources, or we might be willing to trade precision for speeding up the reconstruction process. In either case, using the synthesis operator of the original FUNTF is more attractive rather than using the duals of frames $\{x_i\}_{i\in J}$ that were subject to erasures. However, we must make sure that the error due to erasures is below a reasonable level.

If we know in advance that the number of erasures is limited by a certain number, then dividing data evenly into equal sized packets minimizes the maximum loss due to erasures. In fact, Holmes and Paulsen [49] define the optimal frames for erasures in the following way

Definition 18. A FUNTF $\{x_i\}_{i=1}^N$ is *optimal for m erasures* if it is a minimizer of the function

$$F_m(\{x_i\}_{i=1}^N) = \max_{||x||\leq 1} \quad \max_{|J|=N-m} \left\| x - \frac{d}{N} \sum_{i\in J} \langle x, x_i\rangle x_i \right\|$$

among all FUNTFs.

Theorem 46 is from [49]. We provide a slightly different proof here.

Theorem 46. A FUNTF $\{x_i\}_{i=1}^N$ is optimal for 2 erasures in the sense of Definition 18 if and only if it minimizes the quantity

$$\max_{i\neq j} |\langle x_i, x_j\rangle|$$

among all FUNTFs.

Proof. $\{x_i\}_{i=1}^N$ is optimal for 2-erasures in the sense of Definition 18 if it attains the minimum of

$$F_2(\{x_i\}_{i=1}^N) = \max_{||x||\leq 1} \max_{i\neq j} \left\| \frac{d}{N}(\langle x, x_i\rangle x_i + \langle x, x_j\rangle x_j) \right\|$$

among all FUNTFs.

Let $S_{ij}x = \langle x, x_i\rangle x_i + \langle x, x_j\rangle x_j$. It is not hard to show that S_{ij} has precisely two nonzero eigenvalues $1 \pm |\langle x_i, x_j\rangle|$, and the corresponding eigenvectors are

$$x_i \mp \frac{\langle x_i, x_j\rangle}{|\langle x_i, x_j\rangle|} x_j.$$

Therefore, the operator norm of S_{ij},

$$\|S_{ij}\| = 1 + |\langle x_i, x_j\rangle|.$$

Then,

$$\frac{N}{d} F_2(\{x_i\}_{i=1}^N) = \max_{i\neq j} ||S_{ij}|| = 1 + \max_{i\neq j} |\langle x_i, x_j\rangle|.$$

Therefore, $\{x_i\}_{i=1}^N$ is optimal for 2-erasures if and only if it minimizes the quantity $\max_{i\neq j} |\langle x_i, x_j\rangle|$ among all FUNTFs. □

By Theorem 36, Equiangular tight frames are minimizers of the quantity $\max_{i\neq j} |\langle x_i, x_j\rangle|$ among all FUNTFs. Therefore, equiangular tight frames are optimal for 2 erasures by Theorem 46.

4.1.5 Graph Theory Connection

Paulsen and Bodmann [13] characterized the equiangular tight frames for $\mathbb{R}^d$ in terms of certain graphs, called regular 2-graphs. We shall provide the characterization in Theorem 47, and refer the reader to [13] for the proof.

Definition 19. Given a graph with N vertices, *Seidel adjacency matrix* is the $N \times N$ matrix $Q = (q_{ij})$ where $q_{ii} = 0$, $q_{ij} = 1$ if ith and jth vertices are adjacent, and $q_{ij} = -1$ if not.

A *two-graph* (Ω, Δ) consists of a vertex set Ω and a set Δ of three element subsets of Ω such that every four element subset contains an even number of sets from Δ. A two-graph is *regular* if every two element subset of Ω is contained in same number of sets in Δ.

Theorem 47. Let $\{x_i\}_{i=1}^N$ be a FUNTF for $\mathbb{R}^d$ and L be its Bessel map. Then, $\{x_i\}_{i=1}^N$ is an equiangular tight frame for $\mathbb{R}^d$ if and only if

$$Q = \sqrt{\frac{d(N-1)}{N-d}}(LL^* - I)$$

is the Seidel adjacency matrix of a regular two-graph.

4.1.6 Grassmanian Packing Problem

The Grassmannian $G_k(V)$ is the set of all k-dimensional subspaces of a d-dimensional vector space V. Thus, the Grassmannian $G_1(V)$ is the space of lines through the origin in V, i.e., it is the projective space $P(V)$.

$G_k(V)$ has a topology induced by the metric

$$d(\ell_1, \ell_2) = \|P_1 - P_2\|, \qquad \ell_1, \ell_2 \in G_k(V)$$

where P_i is the orthogonal projection onto ℓ_i, and $\|.\|$ is the operator norm.

Grassmanian packing problem is the problem of locating N k-dimensional subspaces of V so that the minimum distance with respect to this metric between any

two subspaces are maximized. Such those packings are called the optimal packings. In other words, Grassmanian packing problem is the min-max problem

$$\max_{(\ell_i)_{i=1}^N} \min_{i \neq j} d(\ell_i, \ell_j). \tag{4.1}$$

The Grassmanians are compact metric spaces [18, 19], therefore, a solution to the problem (4.1) always exists.

Theorem 48. Let $\{\ell_i\}_{i=1}^N \subseteq G_1(V)$, and $x_i \in \ell_i$ be a unit norm vector. Then, $\{\ell_i\}_{i=1}^N$ is an optimal packing of lines if and only if $\{x_i\}_{i=1}^N$ attains the minimum of

$$\max_{i \neq j} |\langle x_i, x_j \rangle|$$

among all sets $\{x_i\}_{i=1}^N$ of unit norm vectors.

Proof. Let $S_{ij} = P_i - P_j$. Then, S_{ij} is Hermitian, and it has precisely two nonzero eigenvalues. Let

$$b_\pm = \frac{-1 \pm \sqrt{1 - |\langle x_i, x_j \rangle|^2}}{\langle x_j, x_i \rangle}, \quad \lambda_\pm = \pm\sqrt{1 - |\langle x_i, x_j \rangle|^2}, \quad v_\pm = x_i + b_\pm x_j.$$

It is not hard to show that $\lambda_\pm$ are eigenvalues of S_{ij} with corresponding eigenvectors $v_\pm$. Therefore,

$$d(\ell_i, \ell_j) = \|P_i - P_j\| = \|S_{ij}\| = \sqrt{1 - |\langle x_i, x_j \rangle|^2}.$$

Hence, $\{x_i\}_{i=1}^N$ is the solution of the Grassmanian line packing problem (4.1) if and only if it minimizes the quantity $\max_{i \neq j} |\langle x_i, x_j \rangle|$ among all sets $\{x_i\}_{i=1}^N$ of unit norm vectors. □

Definition 20. An equinorm set of vectors $\{x_i\}_{i=1}^N$ in $\mathbb{F}^d$ is a *Grassmanian frame* if it attains the minimum of

$$\max_{i \neq j} |\langle x_i, x_j \rangle|$$

among all sets $\{x_i\}_{i=1}^N$ of unit norm vectors.

Unlike equiangular frames, Grassmanian frames always exist for any pair (d, N) due to compactness. More precisely, for a fixed N, the set of all unit norm frames

$$\{\{x_i\}_{i=1}^N : x_i \in \mathbb{F}^d,\ \|x_i\| = 1\}$$

is a compact metric space, endowed with the metric

$$d(\{x_i\}_{i=1}^N, \{y_i\}_{i=1}^N) = \left(\sum_{i=1}^N \|x_i - y_i\|^2 \right)^{1/2},$$

and $f(\{x_i\}_{i=1}^N) = \max_{i \neq j} |\langle x_i, x_j \rangle|$ is continuous in this metric. Every continuous function attains its minimum value on a compact set.

Let $\{x_i\}_{i=1}^N$ be a frame for $\mathbb{F}^d$. In Theorem 36, we proved that

$$\max_{i \neq j} |\langle x_i, x_j \rangle| \geq \sqrt{\frac{N-d}{d(N-1)}},$$

and that this bound is attained if and only if $\{x_i\}_{i=1}^N$ is an equiangular tight frame. Therefore, equiangular tight frames are Grassmanian frames.

4.2 New Results

4.2.1 p-th Frame Potential

Benedetto and Fickus [5] proved that finite unit-norm tight frames are minimizers of the *frame potential function*

$$FP(\{x_i\}_{i=1}^N) = \sum_{i \neq j} |\langle x_i, x_j \rangle|^2.$$

In analogy to the frame potential, Blume-Kohout, Scott, Caves and Renes [65] defined the *p-th frame potential* (Definition 21). They proved that (d, d^2) complex equiangular tight frames are the minimizers of the second frame potential function, whenever they exist.

Definition 21. Let $p > 1$ and let N be a positive integer. Let $\{x_i\}_{i=1}^N$ be a set of unit norm vectors in $\mathbb{F}^d$ ($\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$). The *p-th frame potential function* FP_p is defined by

$$FP_p(\{x_i\}_{i=1}^N) = \sum_{i \neq j} |\langle x_i, x_j \rangle|^{2p}.$$

We generalize the result of Blume-Kohout, Scott, Caves and Renes to an arbitrary (d, N) in Theorem 50. In order to prove Theorem 50, we need Theorem 49. Theorem 49 is taken from [5], and a proof can be found in [5].

Theorem 49. Let $d < N$, and let $\{x_i\}_{i=1}^N$ be a set of unit norm vectors in $\mathbb{F}^d$ ($\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$). Then,

$$FP(\{x_i\}_{i=1}^N) = \sum_{i \neq j} |\langle x_i, x_j \rangle|^2 \geq \frac{N(N-d)}{d}.$$

Furthermore, the lower bound is achieved if and only if $\{x_i\}_{i=1}^N$ is tight.

Theorem 50. Let $d < N$, $1 < p < \infty$, and let $\{x_i\}_{i=1}^N$ be a set of unit norm vectors in $\mathbb{F}^d$ ($\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$). Then,

$$FP_p(\{x_i\}_{i=1}^N) = \sum_{i \neq j}^{N} |\langle x_i, x_j \rangle|^{2p} \geq N(N-1) \left(\frac{N-d}{d(N-1)} \right)^p. \tag{4.2}$$

Furthermore, the lower bound is achieved if and only if $\{x_i\}_{i=1}^N$ is an equiangular tight frame.

Proof. Let $\mathcal{A}_N$ be the image of set of all unit-norm frames for $\mathbb{F}^d$ consisting of N elements under the map

$$\{x_i\}_{i=1}^N \to (\langle x_i, x_j \rangle)_{i \neq j} \in \mathbb{F}^{N(N-1)}.$$

For any $p > 1$, and $\psi \in \mathcal{A}_N$, we have

$$||\psi||_{2p} \geq (N(N-1))^{(\frac{1}{2p} - \frac{1}{2})} ||\psi||_2$$

by Hölder's inequality. Moreover, the equality holds if and only if

$$\exists c > 0 \quad \forall k = 1, \ldots, N(N-1), \qquad |\psi[k]| = c. \tag{4.3}$$

Therefore,

$$\begin{aligned} \sum_{i \neq j}^{N} |\langle x_i, x_j \rangle|^{2p} &\geq (N(N-1))^{2p(\frac{1}{2p} - \frac{1}{2})} \left(\sum_{i \neq j}^{N} |\langle x_i, x_j \rangle|^2 \right)^p \\ &\geq N(N-1) \left(\frac{N-d}{d(N-1)} \right)^p. \end{aligned} \tag{4.4}$$

We used Theorem 49 for the second inequality in (4.4).

If FP_p attains the lower bound in (4.2), then, $\{x_i\}_{i=1}^N$ must be equiangular by (4.3). Moreover, by (4.4) we must have $\sum_{i \neq j}^{N} |\langle x_i, x_j \rangle|^2 = N(N-d)/d$, and so,

$\{x_i\}_{i=1}^N$ must be tight by Theorem 49. Therefore, $\{x_i\}_{i=1}^N$ must be an equiangular tight frame.

Conversely, if $\{x_i\}_{i=1}^N$ is an equiangular tight frame, then

$$\forall i \neq j \quad |\langle x_i, x_j \rangle| = \sqrt{\frac{N-d}{d(N-1)}}$$

by Theorem 36. Therefore, FP_p attains the lower bound at $\{x_i\}_{i=1}^N$. □

4.2.2 Equiangular Tight Frames for $\mathbb{C}^d$ with Maximum Redundancy

Notation 2. We use the notation xx^* to denote the linear map $y \to \langle y, x \rangle x$, i.e.,

$$(xx^*)y = \langle y, x \rangle x.$$

Lemma 9. Let $\{x_i\}_{i=1}^N$ be a frame for $\mathbb{C}^d$, let L be its Bessel map, and let L^* be the adjoint of L. Then, span $\{x_i x_i^* : i = 1, \ldots, N\} = \{L^* DL : D \text{ diagonal}\}$, and

$$dim(span\{x_i x_i^*\}) = N - dim(W)$$

where $W = \{D \text{ diagonal} : L^* DL = 0\}$.

Proof.

$$\begin{aligned} A \in \text{span}\{x_i x_i^*\} \quad &\Leftrightarrow \quad A = \sum_{i=1}^N \lambda_i x_i x_i^* \quad \text{for some} \quad \lambda_i \in \mathbb{C} \\ &\Leftrightarrow \quad \forall y \in \mathbb{F}^d, \quad Ay = \sum_{i=1}^N \lambda_i \langle y, x_i \rangle x_i = L^* DLy \end{aligned}$$

where $D = \text{diag}(\lambda_1, \ldots, \lambda_N)$.

Second, the map $D \to L^* DL$ is linear, and W is its Null space. As a result, $dim(span\{x_i x_i^*\}) = N - dim(W)$. □

Lemma 10. If $\{x_i\}_{i=1}^N$ is an equiangular tight frame, then $dim(span\{x_i x_i^*\}) = N$

Proof. If $L^*DL = 0$ for some $D = \text{diag}(\lambda_1, \ldots, \lambda_N)$

$$\begin{aligned} &\Rightarrow \quad \sum_{i=1}^{N} \lambda_i |\langle x_j, x_i \rangle|^2 = 0, \ \forall j = 1, \ldots, N \\ &\Rightarrow \quad \frac{N-d}{d(N-1)} \left(\sum_{i=1}^{N} \lambda_i \right) + \left(1 - \frac{N-d}{d(N-1)} \right) \lambda_j, \ \forall j = 1, \ldots, N \end{aligned}$$

but then there is a λ such that $\lambda_j = \lambda$ for every j. Then, $\sum_{i=1}^N \lambda |\langle x_1, x_i \rangle|^2 = 0$, so $\lambda = 0$. Therefore, $W = \{0\}$. □

Theorem 51. If $\{x_i\}_{i=1}^N$ is an equiangular tight frame for $\mathbb{F}^d$, then

$$N \leq d(d+1)/2 \quad \text{if} \quad \mathbb{F} = \mathbb{R}$$

$$N \leq d^2 \qquad\quad \text{if} \quad \mathbb{F} = \mathbb{C}$$

Proof. $M(\mathbb{F}^d)$, the set of all $d \times d$ matrices over $\mathbb{F}$, is a d^2 dimensional vector space. Then, $N = dim(span\{x_i x_i^*\}) \leq d^2$.

When $\mathbb{F} = \mathbb{R}$, $\text{span}\{x_i x_i^*\}$ is a subspace of the space of all real symmetric matrices $SM(\mathbb{R}^d)$, which is $d(d+1)/2$ dimensional, so

$$N = dim(span\{x_i x_i^*\}) \leq d(d+1)/2.$$

□

Theorem 52. let $\{x_i\}_{i=1}^{d^2}$ be a set of unit norm vectors in $\mathbb{C}^d$. Then, the following assertions are equivalent.

i. $\{x_i\}_{i=1}^{d^2}$ is an equiangular tight frame for $\mathbb{C}^d$,

ii. For any $d \times d$ complex matrix B

$$\sum_{i=1}^{d^2} \langle Bx_i, x_i \rangle x_i x_i^* = \frac{d}{d+1}(B + trace(B)I),$$

iii. For every $k, l, k', l' = 1, \ldots, d$

$$\sum_{i=1}^{d^2} x_i(k)\ \overline{x_i(l)}\ \overline{x_i(k')}\ x_i(l') = \frac{d}{d+1}[\delta(k-k')\delta(l-l') + \delta(k-l)\delta(k'-l')],$$

iv. For every $y \in \mathbb{C}^d$

$$\sum_{i=1}^{d^2} |\langle y, x_i\rangle|^4 = \frac{2d}{d+1}||y||^4.$$

Proof. $(i \Rightarrow ii)$ If $\{x_i\}_{i=1}^{d^2}$ is an equiangular tight frame, then $\{x_i x_i^*\}$ spans $M(\mathbb{C}^d)$ by Lemma 10. Then, for any $B \in GL(\mathbb{C}^d)$, there are λ_i s such that $B = \sum \lambda_i x_i x_i^*$.

Next, define

$$\mathcal{S}(B) = \sum_{i=1}^{d^2} \langle Bx_i, x_i\rangle x_i x_i^*.$$

Then, $\mathcal{S}$ is linear, and $\mathcal{S}(x_j x_j^*) = \sum_{i=1}^{d^2} |\langle x_j, x_i\rangle|^2 x_i x_i^* = \frac{d}{d+1}(I + x_j x_j^*)$. By linearity,

$$\mathcal{S}(B) = \frac{d}{d+1}\sum_i \lambda_i (I + x_i x_i^*) = \frac{d}{d+1}\left(\sum_i \lambda_i\right) I + \frac{d}{d+1} B$$

and the result follows with $trace(B) = trace(\sum \lambda_i x_i x_i^*) = \sum \lambda_i$.

$(ii \Rightarrow iii)$ Assume (ii). Then, for any matrix $B = [b_{kl}]$, we have

$$\sum_{i=1}^{d^2} |\langle Bx_i, x_i\rangle|^2 = \frac{d}{d+1} trace(BB^*) + \frac{d}{d+1}|trace(B)|^2.$$

$$\Leftrightarrow \quad \sum_{i=1}^{d^2} \left|\sum_{k,l} b_{kl} x_i(l)\overline{x_i(l)}\right|^2 = \frac{d}{d+1}\left(\sum_{k,l} |b_{kl}|^2 + |\sum_k b_{kk}|^2\right)$$

$$\Leftrightarrow \quad \sum_{k,l}\sum_{k',l'} b_{kl}\overline{b_{k'l'}}\left(\sum_{i=1}^{d^2} x_i(k)\ \overline{x_i(l)}\ \overline{x_i(k')}\ x_i(l')\right) = \frac{d}{d+1}\left(\sum_{k,l} |b_{kl}|^2 + b_{kk}\overline{b_{ll}}\right)$$

$$\Leftrightarrow \quad \sum_{k,l}\sum_{k',l'} b_{kl}\overline{b_{k'l'}}\left(\sum_{i=1}^{d^2} x_i(k)\ \overline{x_i(l)}\ \overline{x_i(k')}\ x_i(l') - \frac{d}{d+1}[\delta(k-k')\delta(l-l') + \delta(k-l)\delta(k'-l')]\right) =$$

Since this is true for any matrix B, (iii) follows.

$(iii \Rightarrow iv)$ Assume (iii). Then, For every $y \in \mathbb{C}^d$

$$\begin{aligned}
\sum_{i=1}^{d^2} |\langle y, x_i\rangle|^4 &= \sum_{i=1}^{d^2} \left| \sum_{k=1}^{d} y(k)\ \overline{x_i(k)} \right|^4 \\
&= \sum_{k,l,k',l'=1}^{d} \sum_{i=1}^{d^2} y(k')\ \overline{y(k)}\ \overline{y(l')}\ y(l) x_i(k)\ \overline{x_i(l)}\ \overline{x_i(k')}\ x_i(l') \\
&= \frac{d}{d+1} \sum_{k,l,k',l'=1}^{d} y(k')\ \overline{y(k)}\ \overline{y(l')}\ y(l)\ [\delta(k-k')\delta(l-l') + \delta(k-l)\delta(k'-l')] \\
&= \frac{2d}{d+1} ||y||^4.
\end{aligned}$$

$(iv \Rightarrow i)$ Assume (iv). Then,

$$\sum_{i,j=1}^{d^2} |\langle x_i, x_j\rangle|^4 = \sum_{j=1}^{d^2} \frac{2d}{d+1} ||x_j||^4 = \frac{2d^3}{d+1}.$$

Then, by Theorem 43, $\{x_i\}_{i=1}^{d^2}$ is an equiangular tight frame for $\mathbb{C}^d$. □

The following theorem is analogous to Theorem 52 for the real equiangular frames with maximum redundancy. The proof is the same as the proof of Theorem 52, except we use Theorem 50 to prove $(iii \Rightarrow i)$ instead of Theorem 43. Therefore, we shall not provide a separate proof for Theorem 53.

Theorem 53. let $N = d(d+1)/2$ and let $\{x_i\}_{i=1}^{N}$ be a set of unit norm vectors in $\mathbb{R}^d$. Then, the following assertions are equivalent.

i. $\{x_i\}_{i=1}^{N}$ is an equiangular tight frame for $\mathbb{R}^d$,

ii. For any $d \times d$ real symmetric matrix B

$$\sum_{i=1}^{N} \langle Bx_i, x_i \rangle x_i x_i^* = \frac{d}{d+1}(B + (1/2)trace(B)I),$$

iii. For every $y \in \mathbb{R}^d$

$$\sum_{i=1}^{N} |\langle y, x_i \rangle|^4 = \frac{3(d+1)}{2(d+2)}||y||^4.$$

It has been conjectured that for every d, there is a finite Heisenberg frame for $\mathbb{C}^{d^2}$, which is also an equiangular tight frame. In fact, there is a short list of such frames for $d = 2, 3$ and 4 in [65], for which there is a (d, d^2) equiangular tight frame. Renes, Blume-Kohout, Scott and Caves [65] also claim that they could numerically calculate an equiangular Heisenberg frame for $5 \leq d \leq 45$.

Definition 22. Let $\phi = (\phi(a))_{a \in \mathbb{Z}_d} \in \mathbb{C}^d$. The modulation and the translation operators are defined as follows:

$$\begin{aligned} M_b\phi(a) &= e^{ib.a}\phi(a), \\ \tau_b\phi(a) &= \phi(a-b). \end{aligned}$$

let $\omega = e^{2\pi i/d}$. The finite group

$$\mathbb{G} = \{T(n, a, b) = \omega^n M_b \tau_a : a, b, n \in \mathbb{Z}_d\}$$

is the *finite Heisenberg-Weyl group.*

The center of this group, the subset of elements that commute with every element in the group, is $Z(\mathbb{G}) = \{\omega^n I : n \in \mathbb{Z}_d\}$. Also, it has no nontrivial invariant subspace of $\mathbb{C}^d$, i.e., $\mathbb{G}$ is a d-dimensional irreducible unitary representation of itself [50].

Definition 23. Let $x \in \mathbb{C}^d$ be a unit-norm vector. The system $\{M_b \tau_a x : a, b \in \mathbb{Z}_d\}$ constitutes a frame for $\mathbb{C}^d$, is called a *Heisenberg frame*.

Heisenberg frames are the orbits of the factor group $\mathbb{G}/Z(\mathbb{G})$. Thus, irreducibility of $\mathbb{G}$ implies the irreducibility of $\mathbb{G}/Z(\mathbb{G})$. Then, Heisenberg frames are always FUNTFs by Theorem 55.

Definition 24. Let G be a group and let $GL(\mathbb{C}^d)$ be the group of all invertible $d \times d$ complex matrices. A group homomorphism $\rho : G \to \mathbb{M}(\mathbb{C}^d)$ is a *representation* of G in $GL(\mathbb{C}^d)$.

A subspace V of $GL(\mathbb{C}^d)$ is an *invariant subspace* of $\rho(G)$ if

$$\forall g \in G, \quad \rho(g)(V) \subseteq V.$$

ρ is an *irreducible representation* if $\rho(G)$ has no invariant proper subspace.

We need the following well-known result of Representation Theory known as "Schur's Lemma", which we use to prove Theorem 55.

Theorem 54. Let G be a group, let ρ be an irreducible representation of G in $GL(\mathbb{C}^d)$, and let $S \in GL(\mathbb{C}^d)$ be a matrix that commutes with every element in $\rho(G)$, i.e.,

$$\forall g \in G, \quad S\rho(g) = \rho(g)S.$$

Then, S is a constant multiple of the $d \times d$ identity matrix I.

Proof. Let λ be an eigenvalue of S, and E_λ be the corresponding eigenspace. Since S commutes with every $\rho(g) \in \rho(G)$, we have

$$\forall v \in E_\lambda, \quad (S - \lambda I)\rho(g)v = \rho(g)(S - \lambda I)v = 0.$$

Thus, $\rho(g)v \in E_\lambda$ for every g. Hence, E_λ must be an invariant subspace of $\rho(G)$. But, ρ is an irreducible representation, and so $\rho(G)$ does not have any proper invariant subspace. Hence, $E_\lambda = \mathbb{C}^d$, and so $S = \lambda I$. □

Theorem 55. Let G be a finite group of matrices over $\mathbb{C}^d$ that has no proper subspace, let $|G| \geq d$, and let $x \in \mathbb{C}^d$ be a unit norm vector. Then,

$$\{gx : g \in G\}$$

constitutes a FUNTF for $\mathbb{C}^d$.

Proof. The frame operator is defined by $Sy = \sum_{g\in G}\langle y, gx\rangle gx$ and it satisfies

$$\forall h \in G, \forall y \in \mathbb{C}^d \quad (hS)y = \sum_{g\in G}\langle y, gx\rangle hgx = (Sh)y,$$

i.e., S commutes with every $g \in G$. Then, by Theorem 54, $S = \lambda I$ for some constant λ. Moreover, since S is positive definite, $\lambda > 0$.

$\{gx : g \in G\}$ is a spanning set, for otherwise $span\{gx : g \in G\}$ would be an invariant proper subspace of G, which contradicts one of the hypotheses.

Hence, $\{gx : g \in G\}$ is a FUNTF for $\mathbb{C}^d$. □

The following theorem is a direct result of the Theorem 52, and an alternative proof is in [14].

Theorem 56. Let $x \in \mathbb{C}^d$ has unit norm, and

$$\begin{aligned} M_b x(n) &= e^{2\pi ibn/d}x(n), \\ T_a x(n) &= x(n+a). \end{aligned}$$

Then, $(M_b T_a x)_{a,b=0}^{d-1}$ is an equiangular tight frame for $\mathbb{C}^d$ if and only if

$$\sum_{a=0}^{d-1} x(a)\ \overline{x(a+k)}\ \overline{x(a+l)}\ x(a+k+l) = \frac{1}{d+1}[\delta(k)+\delta(l)].$$

Proof. By Theorem 52 $\{x_i\}_{i=1}^{d^2}$ is an equiangular tight frame for $\mathbb{C}^d$ if and only if for every $k, l, k', l' = 1, \ldots, d$

$$\sum_{i=1}^{d^2} x_i(k)\ \overline{x_i(l)}\ \overline{x_i(k')}\ x_i(l') = \frac{d}{d+1}[\delta(k-k')\delta(l-l') + \delta(k-l)\delta(k'-l')].$$

Substituting the expression $(M_b T_a x)_{a,b=0}^{d-1}$ for $\{x_i\}_{i=1}^{d^2}$ and simplifying, we obtain the result. □

Bibliography

[1] N. N. Andreev, *Disposition points on a sphere with minimum energy*, Proc. Steklov Inst. Math. **219** (1997), 20–24.

[2] D.M. Appleby, *Sic-povms and the extended clifford group*, (2004).

[3] P. M. Aziz, H. V. Sorensen, and J. Van Der Spiegel, *An overview of Sigma-Delta converters*, IEEE Signal Processing Magazine **13(1)** (1996), 6184.

[4] J. J. Benedetto, *Constructive approximation in waveform design*, Advances in Constructive Approximation Theory (2004), 89–108.

[5] J. J. Benedetto and M. Fickus, *Finite normalized tight frames*, Advances in Computational Mathematics **18:(2-4)** (February 2003), 357–385.

[6] J. J. Benedetto and M. W. Frazier (eds.), *Wavelets: mathematics and applications*, CRC Press, Boca Raton, FL, 1994.

[7] J. J. Benedetto and J. Kolesar, *Geometric properties of grassmanian frames*, EURASIP, Journal of Signal Processing.

[8] J. J. Benedetto, O. Oktay, and A. Tangboondouangjit, *Complex sigma-delta quantization algorithms for finite frames*, AMS Contemporary Mathematics, to Appear.

[9] J. J. Benedetto, A. Powell, and Ö. Yılmaz, *Second order Sigma-Delta ($\Sigma\Delta$) quantization of finite frame expansions*, Applied and Computational Harmonic Analysis **20(1)** (2006), 126–148.

[10] J. J. Benedetto, A. M. Powell, and Ö. Yılmaz, *Sigma-Delta quantization and finite frames*, IEEE Trans. Information Theory **52** (2006), 1990–2005.

[11] J. J. Benedetto and O. M. Treiber, *Wavelet frames: Multiresolution analysis and extension principles*, Wavelet Transforms and Time-Frequency Signal Analysis (L. Debnath, ed.), Birkhäuser, 2001.

[12] W. R. Bennett, *Spectra of quantized signals*, Bell Syst. Tech. J. **27** (1948), 446472.

[13] B. Bodmann and V. Paulsen, *Frames, graphs and erasures*, Preprint (September 2004).

[14] L. Bos and S. Waldron, *Some remarks on heisenberg frames and sets of equiangular lines*, Technical Report (2005).

[15] J. C. Candy and G. C. Temes (eds.), *Oversampling Delta-Sigma Data Converters*, IEEE Press, 1992.

[16] P. Casazza and J. Kovačević, *Equal-norm tight frames with erasures*, Advances in Computational Mathematics **18** (2003), 387–430.

[17] W. Chen and B. Han, *Improving the accuracy estimate for the first order sigma-delta modulator*, J. Amer. Math. Soc. (submitted in 2003).

[18] J.H. Conway, R. Hardin, and N. Sloane, *Packing lines, planes, etc.: Packings in grassmanian spaces*, Preprint.

[19] J.H. Conway and N. Sloane, *Sphere Packing Lattices and Groups*, Springer-Verlag, 1999.

[20] J. H. Cozzens and L. A. Finkelstein, *Computing the discrete fourier transform using residue number systems in a ring of algebraic integers*, IEEE Transactions on Information Theory **31** (1985), 580–588.

[21] J. Cui and W. Freeden, *Equidistribution on the sphere*, SIAM J. Sci. Comput. **18** (1997), no. 2, 595–609.

[22] Z. Cvetković, *Resilience properties of redundant expansions under additive noise and quantization*, IEEE Trans. Information Theory **49(3)** (2003), 644–656.

[23] Z. Cvetković and M. Vetterli, *Deterministic analysis of errors in oversampled A/D conversion and quantization of Weyl-Heisenberg frame expansions*, 1996, submitted to IEEE Trans. on Information Theory.

[24] ———, *Overcomplete expansions and robustness*, Proc. IEEE-SP Int.Symp. on Time-Frequency and Time-Scale Analysis (Paris, France), 1996, pp. 325–328.

[25] ———, *Overcomplete expansions and robustness*, Signal and image representation in combined spaces, Wavelet Anal. Appl., vol. 7, Academic Press, San Diego, CA, 1998, pp. 301–338. MR 99h:94006

[26] Z. Cvetković and M. Vetterli, *On simple oversampled A/D conversion in $L^2(\mathbb{R})$*, IEEE Trans. Information Theory **47** (2001), no. 1, 146–154.

[27] I. Daubechies, *Ten Lectures on Wavelets*, SIAM, Philadelphia, PA, 1992.

[28] I. Daubechies and R. DeVore, *Reconstructing a bandlimited function from very coarsely quantized data: A family of stable sigma-delta modulators of arbitrary order*, Annals of Mathematics **158 (2)** (2003), 679–710.

[29] D. Donoho, *For most large underdetermined systems of linear equations the minimal ℓ^1-norm solution is also the sparsest solution*, Manuscript (2004).

[30] D. Donoho and M. Elad, *Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ^1 minimization*, Proc. Natl. Acad. Sci. **100** (2002), 2197–2202.

[31] D. Donoho and X. Huo, *Uncertainty principles and ideal atomic decomposition*, IEEE Transactions on Information Theory **47** (2001), 2845–2862.

[32] D. Donoho, M. Vetterli, R. DeVore, and I. Daubechies, *Data compression and harmonic analysis*, IEEE Trans. Inform. Th. **44** (1998), no. 6, 2435–2476.

[33] R.J. Duffin and A.C. Schaeffer, *A class of nonharmonic fourier series*, Transactions of the American Matematical Society **72** (1952), no. 2, 341–366.

[34] M. Elad and A. M. Bruckstein, *A generalized uncertainty principle and sparse representation in pairs of* $\mathbb{R}^n$ *bases*, IEEE Transactions on Information Theory **48** (2002), 2558–2567.

[35] R. A. Games, *Complex approximations using algebraic integers*, IEEE Transactions On Information Theory **31** (1985), no. 5, 565–579.

[36] ———, *An algorithm for complex approximations in* $\mathbb{Z}[e^{2\pi i/8}]$, IEEE Transactions On Information Theory **32** (1986), no. 4, 603–607.

[37] R. A. Games, D. Moulin, S. O'Neil, and J. Rushanan, *Algebraic-integer quantization and residue number system processing*, Proc. ICASSP, 1989, pp. 948–951.

[38] A. Gersho and R.M. Gray, *Vector Quantization and Signal Compression*, Kluwer Academic Press/Springer, New York, NY, 1992.

[39] V. K. Goyal, J. Kovačević, and J. A. Kelner, *Quantized frame expansions with erasures*, Appl. Comput. Harmon. Anal. **10** (2001), 203233.

[40] V. K. Goyal, M. Vetterli, and N. T. Thao, *Quantized overcomplete expansions in* $\mathbb{R}^n$*: Analysis, synthesis, and algorithms*, IEEE Transactions on Information Theory **44(1)** (1998), 1631.

[41] R. M. Gray, *Quantization noise spectra*, IEEE Transactions on Information Theory **36(6)** (1990), 12201244.

[42] C. S. Güntürk, *Harmonic analysis of two problems in signal quantization and compression*, Ph.D. thesis, Princeton University, 2000.

[43] C. S. Güntürk, I. Daubechies, R. DeVore, and V. Vaishampayan, *A/D conversion with imperfect quantizers*, IEEE Transactions on Information Theory **52(3)** (2006), 874–885.

[44] C. S. Güntürk, J. C. Lagarias, and V. A. Vaishampayan, *On the robustness of single loop Sigma-Delta modulation*, IEEE Transactions on Information Theory **12(1)** (2001), 6379.

[45] C. S. Güntürk and T. Nguyen, *Ergodic dynamics in sigma delta quantization: tiling invariant sets and spectral analysis of error*, Advances in Applied Mathematics **34** (2005), 523–560.

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

[47] C.S. Güntürk, *Approximating a bandlimited function using very coarsely quantized data: Improved error estimates in sigma-delta modulation*, Journal of the American Mathematical Society **17** (2004), 229–242.

[48] B. Hochwald, T. Marzetta, T. Richardson, W. Sweldens, and R. Urbanke, *Systematic design of unitary space-time constellations*, IEEE Trans. Inform. Theory **46(6)** (2000), 19621973.

[49] R. Holmes and V. Paulsen, *Optimal frames for erasures*, Lin. Alg. Appl. **377** (January 2004), 31–51.

[50] S. Howard, R. Calderbank, W. Moran, H. Schmitt, and C. Savage, *Relationships between radar ambiguity and coding theory*, IEEE International Conference on Acoustoustics, Speech and Signal Processing (Philadelphia, PA, USA), vol. 5, 2005, pp. 897–900.

[51] D. Jimenez, L. Wang, and Y. Wang, *PCM quantization errors and the white noise hypothesis*, SIAM Journal on Mathematical Analysis **38** (2007), no. 6, 2042–2056.

[52] J. Kovačević, P. Dragotti, and V.K. Goyal, *Filter bank frame expansions with erasures*, IEEE Trans. Inform. Th. **48(6)** (2002), 1439–1450, Special Issue in honor of Aaron D. Wyner.

[53] J. Kovačević, M. Vetterli, and V.K. Goyal, *Multiple descripton transform coding: Robustness to erasures using tight frame expansions*, in Proc. International Symposium on Information Theory(ISIT) (1998), 326–335.

[54] ———, *Quantized frame expansions as source-channel codes for erasure channels*, in Proc. IEEE Data Compression Conference (1999), 326–335.

[55] A.B. Kuijlaars and E.B. Saff, *Asymptotics for minimal discrete energy on the sphere*, Trans. Amer. Math. Soc. **350(2)** (1998), 523538.

[56] L.C.Washington, *Introduction to Cyclotomic Fields*, Springer-Verlag, 1982.

[57] P. Lemmens and J. Seidel, *Equiangular lines*, Journal of Algebra **24** (1973), 494–512.

[58] S. Mallat, *A Wavelet Tour of Signal Processing*, 2 ed., Academic Press, 1999.

[59] J. Munch, *Noise reduction in tight weyl-heisenberg frames*, IEEE Transactions on Information Theory **38(2)** (1992), 608616.

[60] S.R. Norsworthy, R.Schreier, and G.C. Temes (eds.), *Delta-Sigma Data Converters*, IEEE Press, 1997.

[61] E. L. Pennec and S. Mallat, *Sparse geometric image representation with bandelets*, IEEE Trans. Image Proc. **14** (2005), 423–438.

[62] L. Perko, *Differential Equations and Dynamical Systems*, Springer-Verlag, New York, NY, 1991.

[63] G. Rath and C. Guillemot, *Syndrome decoding and performance analysis of DFT codes with bursty erasures*, In Proc. Data Compression Conference (2002), 282291.

[64] ———, *Recent advances in DFT codes based on quantized finite frames expansions for ereasure channels*, Preprint (2003).

[65] J. Renes, R. Blume-Kohout, J. Scott, and C. Caves, *Symmetric informationally complete quantum measurements*, Journal Of Mathematical Physics **45** (2004), no. 6, 2171.

[66] L. I. Rudin, S. Osher, and E. Fatemi, *Nonlinear total variation based noise removal algorithms*, Physica D **60** (1992), 259–268.

[67] W. Rudin, *Fourier Analysis on Groups*, Interscience Publishers - John Wiley and Sons, New York, NY, 1962.

[68] S. Sarvotham, D. Baron, and R. G. Baraniuk, *Measurements vs. bits: Compressed sensing meets information theory*, Proc. 44th Allerton Conference on Communication, Control, and Computing (Monticello, IL), 2006.

[69] T. Strohmer and R. Heath, *Grassmanian frames with applications to coding and communications*, Appl. Comput. Harmon. Anal. **14(3)** (2003), 257–275.

[70] M. Sustik, J. Tropp, I. Dhillon, and R. Heath, *On the existence of equiangular tight frames*, Preprint (2004).

[71] J. Tropp, I. Dhillon, and R. Heath, *Finite-step algorithms for constructing optimal cdma signature sequences*, IEEE Trans. Info. Theory **50** (2004), no. 11, 2916–2921.

[72] J. Tropp, I. Dhillon, R. Heath, and T. Strohmer, *Construction of equiangular signatures for synchronous cdma systems*, Proc. of IEEE Int. Sym. on Spread Spectrum Techniques and Applications (Sydney), 2004.

[73] ———, *Designing structured tight frames via an alternating projection method*, IEEE Transactions on Information Theory **51** (2005), no. 1, 188–209.

[74] S. Waldron, *Generalized Welch bound equality sequences are tight frames*, IEEE Trans. Info. Th. **49**, no. 9, 2307–2309.

[75] Y. Wang, *Sigma-delta quantization errors and the traveling salesman problem*, To Appear.

[76] Ö. Yılmaz, *Stability analysis for several sigma-delta methods of coarse quantization of bandlimited functions*, Constructive Approximation **18** (2002), 599623.

[77] ———, *Coarse quantization of highly redundant time-frequency representations of square-integrable functions*, Appl. Comput. Harmon. Anal. **14** (2003), 107132.

[78] G. Zimmermann, *Normalized tight frames in finite dimensions*, Recent Progress in Multivariate Approximation (K. Jetter, W. Haussmann, and M. Reimer, eds.), Birkhäuser, 2001.