ABSTRACT

Title of Dissertation: $\Sigma\Delta$ MODULATION AND CORRELATION CRITERIA FOR THE CONSTRUCTION OF FINITE FRAMES ARISING IN COMMUNICATION THEORY

Joseph Dennis Kolesar, Doctor of Philosophy, 2004

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

In this dissertation we first consider a problem in analog to digital (A/D) conversion. We compute the power spectra of the error arising from an A/D conversion. We then design various higher dimensional analogs of A/D schemes, and compare these schemes to a standard error diffusion scheme in digital halftoning.

Secondly, we study finite frames. We classify certain finite frames that are constructed as orbits of a group. These frames are seen to have subtle symmetry properties. We also study Grassmannian frames which are frames with minimal correlation. Grassmannian frames have an important intersection with spherical codes, erasure channel models, and communication theory. This is the main part of the dissertation, and we introduce new theory and algorithms to

decrease the maximum frame correlation and hence construct specific examples of Grassmannian frames.

A connection has been drawn between the two parts of this thesis, namely A/D conversion and finite frames. In particular, finite frames are used to expand vectors in $\mathbb{R}^d$, and then different quantization schemes are applied to the coefficients of these expansions. The advantage is that all possible outcomes of quantization can be considered because of the finite dimensionality.

$\Sigma\Delta$ MODULATION AND CORRELATION CRITERIA FOR THE CONSTRUCTION OF FINITE FRAMES ARISING IN COMMUNICATION THEORY

by

Joseph Dennis Kolesar

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
2004

Advisory Committee:

Professor John J. Benedetto, Chairman/Advisor
Professor William W. Adams
Professor Raymond L. Johnson
Professor Prakash Narayan
Professor Robert C. Warner

DEDICATION

To my wife, Carrie, our children, Eleanor, Margaret, William, and Samuel/Elizabeth, my parents, Dennis and Janet, and in memory of my grandmother, Olga.

ACKNOWLEDGEMENTS

I would very much like to thank my advisor, John Benedetto, for his generosity, encouragement, insight and patience during my time in graduate school. John has created a stimulating and exciting research community and it is a privilege to be one of his graduate students.

John has a fine collection of students working with him and the cross-fertilization of ideas that occurs with such a group has been instructive and helpful. Of these many students, I especially wish to thank Andy Kebo, Alex Powell, Özgür Yılmaz, Songkiat Sumetkijakan, Juan Romero and Alfredo Nava-Tudela for their helpful advice when listening to various drafts of this research.

I also acknowledge the support and guidance of the many people at the University of Maryland, the University of Vermont, and Duquesne University who have taught me lessons in mathematics and life. I especially acknowledge Professors Mazur, Larget, Keagy, Bradley,

Gaur, Taylor, Reiter, D'Amico, Foote, Dummit, Sands, Golden, Wilson, Adams, Schafer, Hamilton, Grove, Warner, Johnson, Healy, Ramachandran, and Czaja.

My parents and family have been a guide in all my stages of life and I thank them for their constant love, encouragement, and support.

Finally, I express my deep appreciation to my wife, Carrie, for her many sacrifices and constant love and support during my graduate career. I am truly blessed.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

In this thesis we consider two problems. First we study the frequency content of the round off error generated by the quantization step in an analog to digital (A/D) conversion. Not only do we consider the case of a uniform quantizer, that is a quantizer with equal spaced thresholds, but we also consider the case of a more sophisticated and popular quantizer called $\Sigma\Delta$ modulation. Further, we extend $\Sigma\Delta$ modulation from a quantization scheme for one dimensional signals to a scheme for two or d dimensional signals.

Next we study algebraic and analytic properties of finite frames. Algebraically we consider the orbit of a point under the action of a group on a finite dimensional vector space. In communication theory, orbits have been called geometrically uniform sets and are seen to have useful properties. Analytically we consider frames which are minimally correlated, called Grassmannian frames.

These two seemingly different themes of quantization in (A/D) and finite frames are in fact related. The relation is studied in the paper [BPY]. For perspective, we give a rough idea of this work in Section 1.4.

1.1 Analog to digital (A/D) conversion

There are many advantages to digital signal processing, yet many signals are inherently analog. Therefore it is often necessary to perform an A/D conversion. Mathematically, an analog signal can be thought of as a function $f : \mathbb{R} \to \mathbb{R}$, whereas a quantized digital signal is a function $q : \mathbb{Z} \to \{a_0, \ldots, a_M\}$. An A/D conversion can be roughly modeled by replacing $f : \mathbb{R} \to \mathbb{R}$ with $q : \mathbb{Z} \to \{a_0, \ldots, a_M\}$ so that $\|f - q\|$ is small for a given norm $\|\cdot\|$. It is natural to break an A/D conversion into two separate steps. First, discretize the domain of f, that is, replace the continuum of values $\{f(t) : t \in \mathbb{R}\}$ with the samples $\{f(nT) : n \in \mathbb{Z}\}$. This is called the *sampling step.* Second, discretize the range of f, that is, replace each $f(nT)$ with one of a finite number of predetermined quantization values, say $Q(f(nT)) := a_k \in \{a_0, \ldots, a_N\}$. This second step is called the *quantization step.* We shall consider the case of uniform quantization, that is, the case in which the quantization values, a_k, are equally spaced. See Figure 1.1 for an example where we use two different predetermined sets, one with only two values, the other with 6 values. It appears that more information is retained when we use more quantization values.

The sampling step is well understood. We have the following standard result [BF01, Ben97, DD03, OS99].

Classical Sampling Theorem. *Let $T, \Omega > 0$ and assume $0 < 2\Omega T \leq 1$. Let $g \in PW_{1/2T}$ satisfy $\hat{g} = 1$ on $[-\Omega, \Omega]$ and $\hat{g} \in L^\infty(\widehat{\mathbb{R}})$. Then*

$$\forall f \in PW_\Omega, \quad f(t) = T \sum_n f(nT) g(t - nT), \tag{1.1}$$

where convergence is in $L^2(\mathbb{R})$ and uniformly in $\mathbb{R}$.

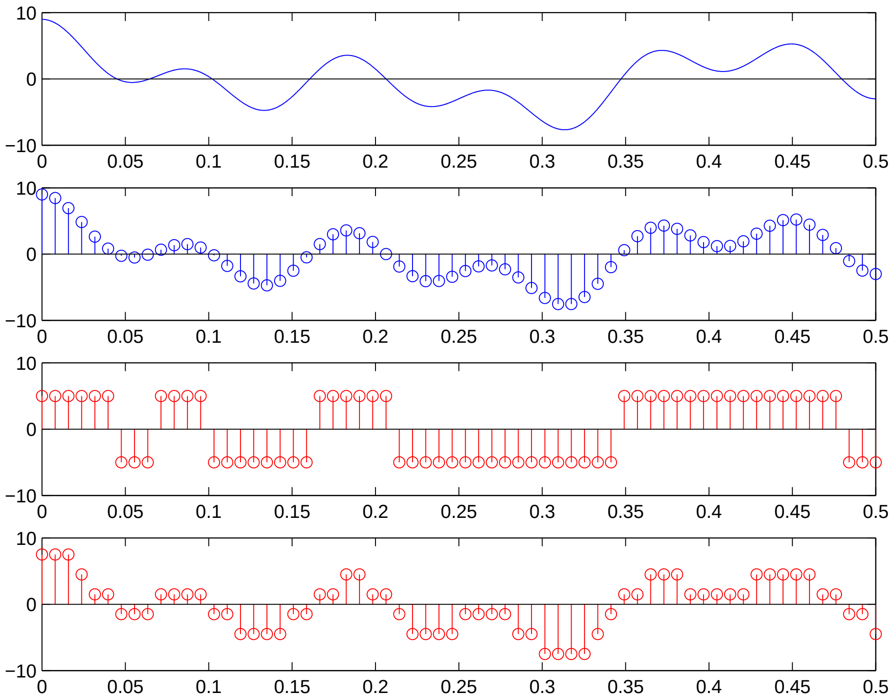


Figure 1.1: The two steps in an A/D conversion of the continuous time signal (A). First sample the time domain (B); second quantize the range values with only two values $\{\pm 5\}$ (C), and with six values $\{\pm 1.5, \pm 4.5, \pm 7.5\}$. It appears that less information is lost when we use more quantization values.

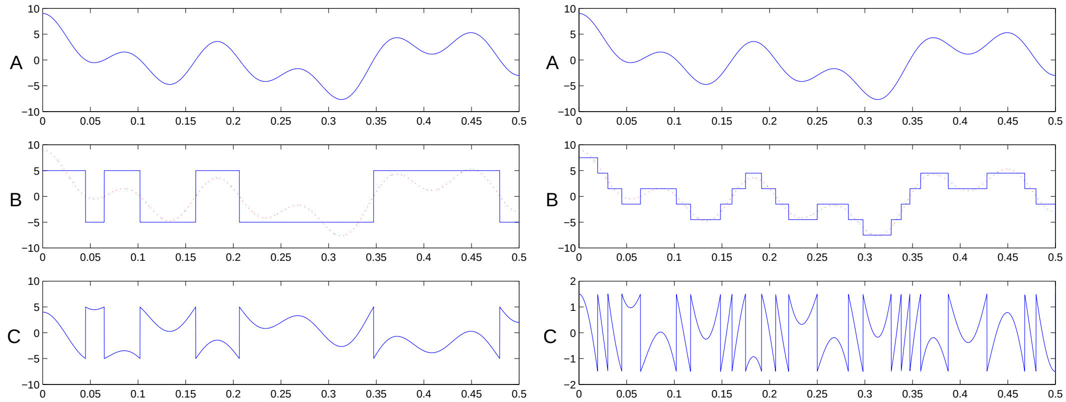


Figure 1.2: Two examples of quantizing before the sampling step. Note, in both figures, (C) is the quantization error which is highly correlated with the signal when we use 2 levels (left) but less correlated when we use 6 levels (right) .

T is the *sampling period*, Ω is the *sampling frequency*, and functions in the Paley-Wiener class, PW_Ω, are called Ω-*bandlimited*, where

$$PW_\Omega = \left\{ g \in L^2(\mathbb{R}) : \operatorname{supp}\{\hat{g}\} \subseteq [-\Omega, \Omega] \right\}.$$

From the Classical Sampling Theorem we know that if we sample with a small enough period, then, for Ω-bandlimited functions, no information is lost by the sampling step when measuring with either the L^2 norm or the L^∞ norm. This is not the case with the quantization step.

In Chapter 2, we shall study the frequency content of the error introduced during this quantization step. This error, $f(nT) - Q(f(nT))$, is called the *quantization error*, where Q is a *quantization rule*, see Figure 1.2 for an example of the quantization error when we quantize without sampling. It is common in the engineering community to assume the quantization error is not correlated with the signal, [EFKM03, OS99, ASVDS96, Gra90]. This assumption simplifies the analysis of the quantization effects. In Chapter 2, we describe a more accurate model of the quantization error.

1.2 Oversampled quantizers

In Figures 1.1 and 1.2 we see that increasing the number of quantization levels results in the quantized output retaining more of the information from the original signal, which therefore results in a quantization error retaining less of the information in the signal. Thus increasing the number of levels is one method of controlling the quantization error. Specifically, we consider the binary expansion of the signal samples

$$f(nT) = \sum_{k=1}^{\infty} b_k 2^{-k},$$

where $b_k \in \{0,1\}$ and where we have scaled $f(nT)$ to lie in the interval $[0,1]$. Now if we fix $K \in \mathbb{N}$, then the quantization rule

$$Q(f(nT)) = Q\left(\sum_{k=1}^{\infty} b_k 2^{-k}\right) = \sum_{k=1}^{K} b_k 2^{-k}, \tag{1.2}$$

corresponds to quantization with 2^K equal spaced levels. So increasing the number of levels, and hence the accuracy of the quantization, corresponds to increasing K and hence retaining more terms of the binary expansion.

Interestingly, increasing K is not the solution of choice in some practical situations [DD03, ASVDS96]. Instead we fix $K = 1$, so that we keep only one term of the expansion (1.2). Specifically, if $\{f(nT)\}$ is now scaled to lie in $[-1,1]$, then the quantization rule becomes

$$Q(f(nT)) = \operatorname{sign}\left(f(nT)\right).$$

Thus to control the information lost in quantization, we increase the number of samples we take for a given interval, i.e., we decrease the sampling period T. We then use the redundancy built into the samples to control the quantization error. A practical method of exploiting this redundancy is called $\Sigma\Delta$ *modulation*, [DD03, ASVDS96, OS99, Gra90], $\Sigma\Delta$ *quantization*, or *error diffusion* [EFKM03]. Specifically, to quantize $f(nT)$ we construct bits, q_n^T, and an auxiliary sequence, u_n, which satisfy the recursion

$$\begin{cases} u_n = u_{n-1} + f(nT) - q_n^T \\ q_n^T = Q(u_{n-1} + f(nT)) = \operatorname{sign}\left(u_{n-1} + f(nT)\right), \end{cases}$$

where we set $u_0 = c \in (-1, 1)$. Note, u_n is sometimes called the *internal state* of the $\Sigma\Delta$ quantizer, [DD03]; and u_n can also be referred to as the *quantization error* since $u_n = \big(u_{n-1} + f(nT)\big) - Q\big(u_{n-1} + f(nT)\big)$.

Now to quantize $f(nT)$ we replace each $f(nT)$ with the corresponding q_n^T. Thus a second type of error is introduced, namely $f(nT) - q_n^T$, the difference between the input to and the output from the entire $\Sigma\Delta$ scheme, not just the quantizer Q. The recursive nature of the $\Sigma\Delta$ scheme results in

$$\sum_{n=1}^{N} \left(f(nT) - q_n^T\right) = u_N - u_0 \leq \|u\|_\infty . \tag{1.3}$$

If $\|u\|_\infty < \infty$, then dividing (1.3) by N shows that the average of the samples $\frac{1}{N}\sum_{n=1}^{N} f(nT)$ approaches the average of the bits $\frac{1}{N}\sum_{n=1}^{N} q_n^T$. Furthermore, if we reconstruct $f(t)$ using $f_q(t) = T\sum_n q_n^T g(t - nT)$, then (1.1) and the first line of the $\Sigma\Delta$ recursion imply,

$$\begin{aligned} |f(t)-f_q(t)| &= T\left|\sum_n (f(nT)-q_n^T)g(t-nT)\right| \\ &= T\left|\sum_n (u_n-u_{n-1})g(t-nT)\right| \\ &= T\left|\sum_n u_n\left(g(t-nT)-g(t-(n+1)T)\right)\right| \qquad (1.4) \\ &\leq T\left\|u\right\|_\infty \sum_n |g(t-nT)-g(t-(n+1)T)| \\ &= T\left\|u\right\|_\infty \sum_n \left|\int_{t-(n+1)T}^{t-nT} g'(y)dy\right| \leq T\left\|u\right\|_\infty \left\|g'\right\|_{L^1}. \end{aligned}$$

Again we see that it is crucial to have $\left\|u\right\|_\infty < \infty$, for in this case, (1.4) implies that the sampling period controls the size of the pointwise reconstruction error.

In Chapter 2, we shall show how an analysis of the quantization error in a uniform quantizer also applies to a $\Sigma\Delta$ quantizer. Finally in Chapter 3, we extend the $\Sigma\Delta$ scheme to higher dimensions.

1.3 Frames for Hilbert space

Consider the expansion in the conclusion of the Classical Sampling Theorem, namely

$$\forall f \in PW_\Omega, \quad f(t) = T\sum_n f(nT)g(t-nT),$$

in the case $2T\Omega = 1$ and

$$g(t) = d_{2\pi\Omega(t)} = \frac{\sin(2\pi\Omega t)}{\pi t} = (\mathbf{1})^\wedge_{[-\Omega,\Omega]}(\gamma).$$

Then $\{g(t-nT)\}_{n\in\mathbb{Z}}$ is a basis for PW_Ω, whereas if $2T\Omega < 1$, then $\{g(t-nT)\}$ is an over complete spanning set in PW_Ω. In either case, we still have a decomposition for f. In the over complete case, $\{g(t-nT)\}_{n\in\mathbb{Z}}$ is called a frame. So we see that frames are a natural language in which to study the sampling step in an A/D conversion.

We now introduce the basic definitions of frame theory [DS52, BF03, BF94, Dau92, Chr02]. Let $\mathcal{H}$ be a separable Hilbert space, and let $X = \{x_n : n \in \mathcal{I}\} \subset \mathcal{H}$ where $\mathcal{I}$ is a countable indexing set. Consider the following map associated with the set X:

$$\begin{aligned} L : \mathcal{H} &\to \ell^2(\mathcal{I}) \\ y &\mapsto \{\langle y, x_n \rangle\}_{n\in\mathcal{I}} \,. \end{aligned}$$

If L is a well-defined linear map, i.e., if $\sum_{n\in\mathcal{I}} |\langle y, x_n\rangle|^2 < \infty$ for any $y \in \mathcal{H}$, then we call L a *Bessel map* and X a *Bessel sequence*. The *adjoint* of L is the map

$$\begin{aligned} L^* : \ell^2(\mathcal{I}) &\to \mathcal{H} \\ \{c[n]\}_{n\in\mathcal{I}} &\mapsto \sum_{n\in\mathcal{I}} c[n] x_n. \end{aligned}$$

Intuitively, L can be considered an analysis operator, and L^* a synthesis operator. The *frame operator* is the map $S : \mathcal{H} \to \mathcal{H}$ defined as L^*L. So, for any $y \in \mathcal{H}$,

$$S(y) = L^*\left(L(y)\right) = L^*\left(\{\langle y, x_n\rangle\}_{n\in\mathcal{I}}\right) = \sum_{n\in\mathcal{I}} \langle y, x_n\rangle\, x_n.$$

Finally, the *Grammian operator* is the map $G : \ell^2(\mathcal{I}) \to \ell^2(\mathcal{I})$ defined by $G = LL^*$. Note that both S and G are self adjoint.

A Bessel sequence X is a *frame* for $\mathcal{H}$ if there exist constants A, B with $0 < A \le B < \infty$ such that for any $y \in \mathcal{H}$

$$A \left\|y\right\|^2 \le \sum_{n\in\mathcal{I}} |\langle y, x_n\rangle|^2 \le B \left\|y\right\|^2.$$

Thus, given any frame, we have four natural maps: L, L^*, S, and G. If the indexing set $\mathcal{I}$ is finite then X is called a *finite frame*. Also, if $A = B$ then X is called a *tight frame*, or, if we wish to emphasize the bound, an A-*tight frame*.

The lower frame bound implies that S is invertible. Thus, we have the two frame reconstruction formulas,

$$y = SS^{-1}y = \sum_{n\in\mathcal{I}} \left\langle S^{-1}(y), x_n \right\rangle x_n = \sum_{n\in\mathcal{I}} \left\langle y, S^{-1}(x_n) \right\rangle x_n$$

and

$$y = S^{-1}Sy = S^{-1}\left(\sum_{n\in\mathcal{I}} \langle y, x_n\rangle x_n\right) = \sum_{n\in\mathcal{I}} \langle y, x_n\rangle S^{-1}(x_n).$$

The set $\{S^{-1}(x_n)\}$ is also a frame, and it called the *dual frame*. In general, it is difficult to invert the frame operator and compute the dual frame.

1.4 A/D conversion and finite frames

At the beginning of Section 1.1 we saw that frames are linked with the sampling step of an A/D conversion. Also,in the third paragraph of this thesis, we mentioned that finite frames have a nontrivial intersection with A/D conversion.

We now give a brief description of the main idea in the paper [BPY] which studies this intersection. View each vector in $\mathbb{R}^d$ as a distinct signal and consider only signals in a bounded region, say $R = \left\{v \in \mathbb{R}^d : \|v\| \leq 2\right\}$. Then, given a finite frame $\{x_k\}_{k=1}^N$ for $\mathbb{R}^d$, we can expand each vector in R in terms of the frame, i.e., $v = \sum_{k=1}^N \langle v, x_k\rangle S^{-1}x_k$. The coefficients, $\langle v, x_k\rangle$, of this frame expansion correspond to the sampling step in an A/D conversion. Notice that since v is bounded by two, the coefficients are also bounded by two. Next we consider the discrete set $D = \left\{\sum_{k=1}^N \varepsilon_k x_k : \varepsilon_k \in \{\pm 1\}\right\}$. D consists of all possible 2-bit quantizations of vectors in R, see Figure 1.3 for the complexity of patterns for different choices of D. Now given a $v \in R$, we want to study methods for choosing a $q \in D$ which is close to v. We can translate the quantization schemes used in

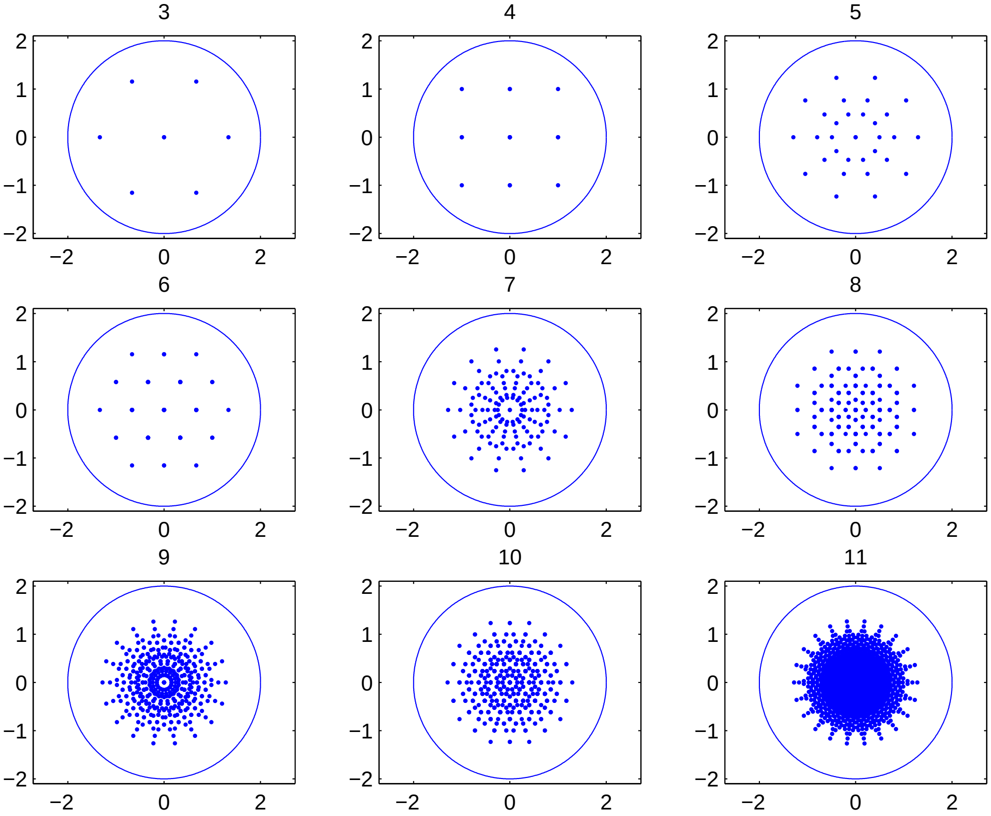


Figure 1.3: All possible quantizations in $\mathbb{R}^2$ for $N = 3, \ldots, 11$, using the quantization levels $\{\pm 1\}$, and the harmonic frames $x_k = \frac{2}{N} e^{2\pi i k/N}$ where $k = 1, \ldots, N$.

A/D conversion to this setting and study which points of D are chosen by the schemes. This is an active field of research.

1.5 Geometrically uniform frames

Let X_N be the Nth roots of unity considered in Section 1.4. We note that X_N has a high degree of symmetry, i.e., if we rotate the frame by the angle $\frac{2\pi}{N}$, we obtain the same frame again. In fact, if we let $x = (1,0)^T$ and consider rotating x by $0, \frac{2\pi}{N}, \ldots, \frac{2\pi(N-1)}{N}$ we obtain the frame X_N. Note, these N rotations form a group isomorphic to $\mathbb{Z}/N\mathbb{Z}$.

In attempting to construct finite frames with a high degree of symmetry, we can generalize the example of the Nth roots of unity to an arbitrary finite subgroup of $O_d(\mathbb{R})$, the $d \times d$ orthogonal matrices. Let us introduce some definitions from group theory. Let G be a group and X be a set. *G acts* on X if there is a function $G \times X \to X$, denoted by $(g,x) \mapsto gx$, such that

(i) $(gh)x = g(hx)$, for all $g,h \in G$ and $x \in X$,

(ii) $1x = x$, for all $x \in X$, where 1 is the identity of G.

For any $x \in X$, the *orbit* of x by G is the set

$$\mathrm{Orb}_G(x) = \{gx \in X : g \in G\}.$$

Let G be a finite subgroup of $O_d(\mathbb{R})$. Then G acts on the set $\mathbb{R}^d$. For any $x \in \mathbb{R}^d$, the set, $\mathrm{Orb}_G(x)$, is called a *geometrically uniform* (GU) set. These sets arise in coding theory and have a high degree of symmetry. In Chapter 3, we construct examples of GU sets. It would seem that such sets could play a role in the quantization sets in Figure 1.3.

1.6 Grassmannian frames

As suggested in [SH03], one way to construct frames which are similar to orthonormal bases is to consider the properties that define orthonormal bases and relax them slightly. For example, assume that for $n = 1, \ldots, d$, $\|x_n\|_{\mathbb{R}^d} = 1$ and that $\operatorname{span}\{x_n : n = 1, \ldots, d\} = \mathbb{R}^d$. Consider the following properties:

$$\forall y \in \mathbb{R}^d, \quad y = \sum_{n=1}^{d} \langle y, x_n \rangle \, x_n, \tag{1.5}$$

$$\forall m \neq n, \quad \langle x_n, x_m \rangle = 0. \tag{1.6}$$

If we assume that $\{x_n\}$ satisfies either (1.5) or (1.6), we can conclude that $\{x_n\}$ is an orthonormal basis. Now, if we relax (1.5) so that

$$\forall y \in \mathbb{R}^d, \quad y = \frac{d}{N} \sum_{n=1}^{N} \langle y, x_n \rangle \, x_n,$$

where $N > d$, then $\{x_n\}$ is no longer an orthonormal basis, but

$$\|y\|^2 = \left\langle y, \frac{d}{N} \sum_{n=1}^{N} \langle y, x_n \rangle \, x_n \right\rangle = \frac{d}{N} \sum_{n=1}^{N} |\langle y, x_n \rangle|^2 ,$$

i.e., $\{x_n\}$ is an $\frac{N}{d}$-tight frame.

Relaxing condition (1.6) gives a different type of frame, called a Grassmannian frame, which is also a generalization of an orthonormal basis. In Chapter 5, we study these frames in detail.

1.7 Results

We now list the results in this thesis.

In Chapter 5, we prove that Grassmannian frames exist for every $N \geq d$. We completely characterize all two dimensional Grassmannian frames up to rotations and sign changes. We also expand a theorem in [SH03], which provides a lower bound for the maximum correlation depending only on the number of frame elements N and the dimension of the space d. Furthermore we give a complete and detailed proof of this theorem which is not present in [SH03]. Next we take on the task of constructing 3 dimensional Grassmannian frames. We develop new theory for explicitly reducing the maximum correlation of a four element frame in $\mathbb{R}^3$. We then use this theory to provide another proof that the $(4, 3)$ -Grassmannian bound is $1/3$.

Since the results used in the $(4, 3)$ case do not immediately apply to $N > 4$, we need to use some notions from convex analysis when $N > 4$. Using these notions, we extend the algorithm of explicitly reducing the maximum correlation of a frame, to any N and d. We then use these ideas to give an explicit proof of the $(5, 3)$-Grassmannian bound. Finally we prove the $(6, 3)$ bound.

Also, in Chapter 5, we apply the method of reducing the maximum correlation in a given frame to the case $N > 3$ and $d \leq 2$. We observe that cyclically applying the algorithm to N elements results in an arrangement which is a subset of a Grassmannian frame with greater than N elements.

The apparently ad hoc methods and proofs in Chapter 5 are state of the art in the subject. In fact, we must first answer these basic combinatorial questions before proceeding to more advanced analytic questions. The results in Chapter 5 are in some sense parallel to the research program begun by J. Conway, R. Hardin, and N. Sloane, see [CHS96].

In Chapter 4, we construct specific examples of GU frames. We consider both Abelian and non-Abelian groups. In $\mathbb{R}^3$, we use the classification of all subgroups of SO_3 and O_3 to construct three dimensional GU frames. We also show that any finite frame can be orthogonally transformed into a frame with a diagonal frame operator. This generalizes the situation for A-tight frames where $S = AI_d$.

In Chapter 3, we provide an alternate proof of a two dimensional generalization of the one dimensional $\Sigma\Delta$ modulator. A similar generalization has been also studied independently in [Yıl02]. We then use the ideas in this proof to construct different quantization schemes. Also, we prove that a specific class of signals will make the original generalization unstable.

We begin the thesis in Chapter 2 by giving detailed calculations for power spectra of quantization error in the case of sinusoidal inputs. We then show how this is applied to $\Sigma\Delta$ modulation.

1.8 Notation

In this section we list the notation and standard theorems used through this thesis.

The *Fourier transform* of f on $\mathbb{R}$ is

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

with corresponding inversion formula

$$f(t) = \int_{-\infty}^{\infty} \hat{f}(\gamma)e^{2\pi it\gamma}d\gamma.$$

The torus is $\mathbb{T}_{2\Omega} = \widehat{\mathbb{R}}/(2\Omega\mathbb{Z})$. We take any fixed interval of length 2Ω to be the representatives of $\mathbb{T}_{2\Omega}$. Functions on $\mathbb{T}_{2\Omega}$ are 2Ω-periodic functions on $\mathbb{R}$. The *Fourier transform* of f on $\mathbb{Z}$ is

$$F(\gamma) = \sum_{n=-\infty}^{\infty} f[n]e^{-2\pi in\gamma/(2\Omega)},$$

with corresponding inversion formula

$$f[n] = \frac{1}{2\Omega}\int_{-\Omega}^{\Omega} F(\gamma)e^{2\pi in\gamma/(2\Omega)}d\gamma.$$

The representation, F, is called a Fourier series with Fourier coefficients, $\{f[n]\}_{n\in\mathbb{Z}}$.

The *deterministic autocorrelation* of a function $f : \mathbb{Z} \to \mathbb{R}$ is

$$r_f[k] = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} f[n+k]\overline{f[n]},$$

If f is p-periodic on $\mathbb{Z}$, then the autocorrelation is defined as

$$r_f[k] = \frac{1}{p}\sum_{n=1}^{p} f[n+k]\overline{f[n]},$$

for $k = 1, 2, \ldots, p$.

The *power spectrum* of f is the Fourier transform of the autocorrelation function,

$$S_f(\gamma) = \sum_{k=-\infty}^{\infty} r_f[k]e^{-2\pi in\gamma/(2\Omega)}$$

In some physical situations, the function f cannot be measured, yet the autocorrelation r_f can be measured. In these cases, the power spectrum of f contains information about the magnitude, but not phase, of the frequency components found in f. A common interpretation of the integral $\int_a^b S_f(\gamma)d\gamma$ is the average power contained in the frequency band $[a, b]$.

For this thesis, a function f is said to be uniformly distributed white noise if the power spectrum of f is a Dirac delta measure at 0.

The *z-transform* of a function $x : \mathbb{Z} \to \mathbb{C}$ is $X : \mathbb{C} \to \mathbb{C}$, defined by

$$X(z) = \sum_{n=-\infty}^{\infty} x[n]z^{-n}.$$

Note that $X(e^{2\pi i\gamma})$ is the Fourier transform of x, i.e., $X(e^{2\pi i\gamma}) = \hat{x}(\gamma)$.

A *Bessel function of order* m is denoted $J_m(z)$, and is defined as the coefficients of the Fourier series of the 2π-periodic function $e^{iz\sin(x)}$, i.e.,

$$e^{iz\sin(x)} = \sum_{m=-\infty}^{\infty} J_m(z)e^{imx}.$$

A function $f(m) = O(g(m))$ as $m \to \infty$ means, there exist a constant $B > 0$ such that $\lim_{m\to\infty} \left|\frac{f(m)}{g(m)}\right| = B$.

The *floor* of a number x, is denoted $\lfloor x \rfloor$, and is defined as the greatest integer less than or equal to x. The *fraction part* of x is $\langle x \rangle = x - \lfloor x \rfloor$.

The *unit sphere* in $\mathbb{R}^d$ is $S^{d-1} = \left\{x \in \mathbb{R}^d : \|x\| = 1\right\}$.

The *Dirac vector basis*, or *canonical basis* for $\mathbb{R}^d$ is $\mathcal{D} = \{\delta_k\}_{k=1}^{d}$, where $\delta_k[n]$ equals one if $k = n$, and 0 otherwise.

A $d \times d$ matrix U is *orthogonal* if the columns of U are orthonormal, i.e., $U^TU = I_d$, where U^T is the transpose of U and I_d is the $d \times d$ identity matrix. If U is orthogonal, then for any $x, y \in \mathbb{R}^d$, $\|Ux\| = \|x\|$ and $\langle Ux, Uy \rangle = \langle x, y \rangle$.

The set of all $d \times d$ orthogonal matrices forms the *orthogonal group*,

$$O_d = \left\{U \in GL(d, \mathbb{R}) : U^TU = 1\right\},$$

where $GL(d, \mathbb{R})$ is the group of $d \times d$ invertible matrices. Because

$$1 = \det(I) = \det(U^TU) = \det(U)^2,$$

we note $\det : O_d \to \{+1, -1\}$, and is a group homomorphism with kernel equal to the *special orthogonal group* $SO_d = \{U \in O_d : \det(U) = +1\}$, which is therefore a subgroup of index 2 in O_d. We think of SO_d as *rotations* and $O_d \setminus SO_d$ as *reflections*.

A $d \times d$ matrix A is *symmetric* if $A^T = A$. The spectral theorem for symmetric matrices is

Theorem 1.1 (Spectral Theorem). *A $d \times d$ symmetric matrix A has the following properties:*

$(a.)$ *A has n real eigenvalues counting multiplicities.*

$(b.)$ *The dimension of the eigenspace for each eigenvalue λ equals the multiplicity of λ as a root of the characteristic equation* $\det(A - \lambda I) = 0$.

$(c.)$ *The eigenspaces are mutually orthogonal, in the sense that eigenvectors corresponding to different eigenvalues are orthogonal.*

$(d.)$ *A is orthogonally diagonalizable, i.e., there is an orthonormal basis of eigenvectors for A.*

We also use the following classical result from ergodic theory,

Theorem 1.2 (Weyl). *If γ is irrational, and $h : \mathbb{R} \to \mathbb{C}$ is 1-periodic and Riemann integrable, then*

$$\lim_{N\to\infty} \frac{1}{2N+1} \sum_{k=-N}^{N} h(\langle k\gamma \rangle) = \int_0^1 h(x)dx.$$

Intuitively the sequence $\{\langle k\gamma \rangle\}$ fills the unit interval so the sums converge to the integral.

Finally, we briefly define the groups that appear in this thesis. A group is *Abelian* if for every $x, y \in G$, $xy = yx$.

$\mathbb{Z}/n\mathbb{Z}$ denotes the additive group of *integers modulo n* and is defined as the set $\{0, 1, \ldots, n-1\}$ with the group law being defined as addition mod n, i.e., $a + b = r$ where r is the remainder after dividing $a + b$ by n.

D_{2n} denotes the *dihedral group of order* $2n$ and is defined as the set of symmetries of a regular n-gon. More precisely, if R is rotation by $\frac{2\pi}{n}$, S is reflection through the x-axis, and the group law is symmetry composition, then R^n is the identity, $SR = R^{n-1}S$, and $D_{2n} = \{1, R, R^2, \ldots, R^{n-1}, S, SR, SR^2, \ldots, SR^{n-1}\}$.

S_n denotes the *symmetric group of degree* n and is defined as the set of all bijections or permutations of $\{1, \ldots, n\}$ with the group law being defined as function composition.

A_n denotes the *alternating group of degree* n. A_n is a subgroup of S_n and is defined as the set of all even permutations in S_n. A permutation is *even* if it can be written as a composition of an even number of 2-cycles. A permutation σ is a *2-cycle* if σ fixes all but two of the elements in its domain $\{1, \ldots, n\}$.

$\mathbb{Z}/n\mathbb{Z}$ is Abelian, D_{2n} and S_n are non-Abelian for $n \geq 3$, and A_n is non-Abelian for $n \geq 4$. Finally A_4, S_4, and A_5 are isomorphic to the rotational symmetries of a tetrahedron, cube/octahedron, and icosahedron/dodecahedron, respectively.

Chapter 2

Quantization and Power Spectra

It is important to understand, in detail, the information loss at the quantization step of an A/D conversion. We begin by recalling the first order ΣΔ quantization,

$$\begin{cases} e_n = e_{n-1} + f(nT) - q_n^T \\ q_n^T = Q(e_{n-1} + f(nT)), \end{cases} \tag{2.1}$$

where Q, called the *quantizer*, is some thresholding function such as $\operatorname{sign}(\cdot)$, $f(nT)$ is a sample from a bandlimited function, q_n^T is the associated output bit from the quantizer, and e_n is the *quantization error* which is also called the *internal state* in [DD03] and is labeled u_n there. To motivate this change in notation (from u_n to e_n), introduce the sequence $w_n = f(nT) + e_{n-1}$, called the *modified input*. Then we can rewrite (2.1) as

$$\begin{cases} w_n = f(nT) + e_{n-1} \\ q_n^T = Q(w_n) \\ e_n = w_n - q_n^T. \end{cases} \tag{2.2}$$

Studying the recursion (2.2) we see how the ΣΔ scheme can be split into the following three steps:

1. modify the input by adding previous quantization errors,
2. apply the quantization rule to the modified input,
3. compute the current quantization error.

The recursion (2.2) is the one dimensional analog of a common way to write error diffusion schemes which are used to halftone images. Also, e_n in (2.2) is consistent with the notation in [Gra90].

The first line of (2.1),

$$e_n = e_{n-1} + f(nT) - q_n^T, \tag{2.3}$$

displays a relationship between the two types of error found in $\Sigma\Delta$ schemes. The first kind of error, $f(nT) - q_n^T$, is the difference between the input and output of the entire scheme. The second type of error, e_n, is the error which we have referred to as the quantization error. It is the difference between the input and output of Q. In order to understand the relationship between these two types of errors, following the development in [EFKM03], we take the z-transform of (2.3). Thus we obtain,

$$E(z) = E(z)z^1 + X(z) - B(z), \tag{2.4}$$

where

$$E(z) = \sum_{n=-\infty}^{\infty} e_n z^{-n},$$

$$X(z) = \sum_{n=-\infty}^{\infty} f(nT) z^{-n},$$

and

$$B(z) = \sum_{n=-\infty}^{\infty} q_n^T z^{-n}.$$

We can rewrite (2.4) as

$$X(z) - B(z) = H(z)E(z), \quad \text{where} \quad H(z) = 1 - z. \tag{2.5}$$

Thus, if we know $E(z)$, the frequency content of e_n, then we can derive $X(z) - B(z)$, the frequency content of $f(nT) - q_n^T$. Since $H(z) = 0$ at $z = 1$, this frequency domain representation of (2.3) shows that the $\Sigma\Delta$ scheme is trying to minimize $X(1) - B(1) = \sum_{n=-\infty}^{\infty} f(nT) - \sum_{n=-\infty}^{\infty} q_n^T$.

Also notice that if we solve for $E(z)$ in (2.5) we obtain

$$E(z) = K(z)(X(z) - B(z)), \quad \text{where} \quad K(z) = \frac{1}{1-z};$$

and we see that $K(z)$ has a pole at $z = 1$, which suggests that $\Sigma\Delta$ can be roughly viewed as an error minimization that gives higher priority to the DC component.

2.1 A quantization error calculation

For $\Sigma\Delta$ modulation, we seek to understand the frequency components found in e_n. We shall do this by first isolating the quantizer, that is, considering the second line of (2.2) separate from the rest of the recursion. This is the approach taken in [Gra90]. To this end, let M be a positive even integer and let Δ be a positive real number. Consider the function $Q : \mathbb{R} \to \{a_1, a_2, \ldots, a_M\}$ given by

$$Q(w) = \sum_{j=1}^{M} a_j \mathbf{1}_{A_j}(w),$$

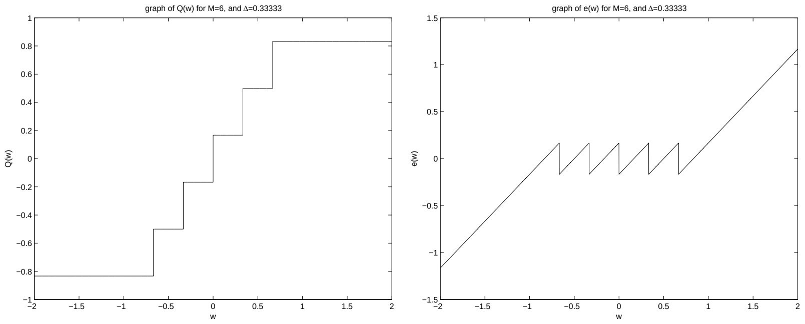


Figure 2.1: Graph of Q (left) and e (right), with $M = 6$ and $\Delta = 1/3$.

where

$$\begin{aligned}
a_j &= (-M + 2j - 1)\,\frac{\Delta}{2} \text{ if } j = 1, \ldots, M \\
A_j &= \left[a_j - \frac{\Delta}{2}, a_j + \frac{\Delta}{2}\right) \text{ if } j = 2, \ldots, M-1, \\
A_1 &= \left(-\infty, a_1 + \frac{\Delta}{2}\right), \\
A_M &= \left[a_M - \frac{\Delta}{2}, \infty\right).
\end{aligned}$$

We call Q a *uniform quantizer* with M levels and Δ spacing. If $M = 2$ and $\Delta = 1$, then $Q = \operatorname{sign}(\cdot)$. For other choices of M and Δ, the graph of Q is a staircase with M levels, rising at an angle of $\frac{\pi}{4}$, see Figure 2.1. We can also write Q in terms of the floor function $\lfloor \cdot \rfloor$,

$$Q(w) = \begin{cases} \Delta\left(\frac{1}{2} + \left\lfloor \frac{w}{\Delta} \right\rfloor\right), & \text{if } w \in \big[-M\Delta/2, M\Delta/2\big) \\ \operatorname{sign}(w)\frac{(M-1)\Delta}{2}, & \text{otherwise.} \end{cases} \tag{2.6}$$

Now, consider the quantization error function $e(w) = w - Q(w)$. By inspecting the graph of e in Figure 2.1, we notice that e is a Δ-periodic function in the interval $\left[-\frac{M}{2}\Delta, \frac{M}{2}\Delta\right)$. This interval is called the *no-overload* interval for Q and e. We have

Proposition 2.1. *Let $Q(w) = \sum_{j=1}^{M} a_j \mathbf{1}_{A_j}(w)$ be a uniform quantizer with M levels and Δ spacing, and let $e(w) = w - Q(w)$. For any w in the no-overload interval, i.e., $w \in [-M\Delta/2, M\Delta/2)$, we have*

$$e(w) = -\Delta\left(\frac{1}{2} - \left\langle\frac{w}{\Delta}\right\rangle\right), \tag{2.7}$$

and

$$e(w) = \sum_{l \neq 0} \frac{-\Delta}{2\pi i l} e^{2\pi i l w/\Delta}. \tag{2.8}$$

Proof. To prove (2.7), use (2.6). Hence for w in the no-overload interval,

$$\begin{aligned} e(w) = w - Q(w) &= \Delta\left(\frac{w}{\Delta}\right) - \left(\frac{\Delta}{2} + \Delta\left\lfloor\frac{w}{\Delta}\right\rfloor\right) \\ &= -\Delta\left(\frac{1}{2} - \frac{w}{\Delta} + \left\lfloor\frac{w}{\Delta}\right\rfloor\right) = -\Delta\left(\frac{1}{2} - \left\langle\frac{w}{\Delta}\right\rangle\right) \end{aligned}$$

To prove (2.8), since w is assumed in the no-overload interval, e is a Δ-periodic function of w, hence e has the Fourier series representation

$$e(w) = \sum_{l \neq 0} \frac{-\Delta}{2\pi i l} e^{2\pi i l w/\Delta}, \quad \text{for } w \in \left[-\frac{M}{2}\Delta, \frac{M}{2}\Delta\right).$$

where w is the value of the input. □

In view of this proposition, our goal is to choose the input sequence, w_n, so that equation (2.8) simplifies and makes computation of the power spectrum of $e(w_n)$ relatively easy.

2.2 Constant input

For the remainder of Chapter 2, let M and Δ be fixed and let Q be an M level Δ spaced uniform quantizer. Then, given an input w_n, let $e_n := e(w_n) = w_n - Q(w_n)$

be the *quantization error sequence* associated with w_n. With the goal that was stated at the end of the previous section in mind, first consider $w_n = c$, where c is a constant.

Proposition 2.2. *Let $w_n = c$ for all $n \in \mathbb{Z}$, and let e_n be the associated quantizer error sequence. Then the autocorrelation of e_n is*

$$r_e(k) = |c - Q(c)|^2 ,$$

and the power spectrum of e_n is

$$S_e(\gamma) = |c - Q(c)|^2 \delta_0(\gamma)$$

Proof. Since $w_n = c$, where c is a constant, we have that

$$e_n := e(w_n) = e(c) = c - Q(c).$$

Hence the autocorrelation is

$$r_e(k) = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} e_{n+k}\overline{e_n} = \lim_{N\to\infty} |c - Q(c)|^2 .$$

Thus, the power spectrum is

$$S_e = (r_e)^{\wedge} = |c - Q(c)|^2 (e^{2\pi i 0 n})^{\wedge} = |c - Q(c)|^2 \delta_0.$$

□

Therefore, as we would expect, all the power in the quantization error is concentrated in the DC component, i.e., 0 frequency.

2.3 Sinusoidal input

Next, consider the input $w_n = A\sin(n\omega_0 + \theta)$ where $\omega_0 = 2\pi\gamma_0$, $\gamma_0 > 0$, $\theta \in \mathbb{R}$, and $0 < A < \frac{M}{2}\Delta$. For this choice of A, w_n lies in the no-overload interval, hence the Fourier representation (2.8) is valid for each term in the sequence $\{w_n\}$. Using this representation, we show

Proposition 2.3. *Let $\gamma_0 > 0$, $\omega_0 = 2\pi\gamma_0$, $\theta \in \mathbb{R}$, and $0 < A < \frac{M}{2}\Delta$. Let $w_n = A\sin(n\omega_0 + \theta)$ and let e_n be the associated quantization error sequence. Then*

$$e_n = \sum_{m=-\infty}^{\infty} b_m e^{i(\omega_0 m)n}$$

where

$$b_{2m+1} = \frac{-\Delta e^{i(2m+1)\theta}}{\pi i(2m+1)} \left(1 + 2\sum_{k=1}^{\lfloor A/\Delta \rfloor} \cos\left((2m+1)\sin^{-1}(\Delta k/A)\right)\right) \tag{2.9}$$

and $b_{2m} = 0$.

Proof. Substituting this sinusoidal input into the Fourier expansion of the quantization error (2.8), we have

$$\forall n, \quad e_n := e(w_n) = \sum_{l\neq 0} \frac{-\Delta}{2\pi i l} e^{2\pi i l A \sin(n\omega_0+\theta)/\Delta}. \tag{2.10}$$

Now, we can generate the Bessel functions of order m by considering the Fourier transform of the 2π-periodic function $e^{iz\sin x}$. Then the synthesis equation gives

$$e^{iz\sin x} = \sum_{m=-\infty}^{\infty} J_m(z)e^{imx}. \tag{2.11}$$

Next, letting $z = 2\pi l A/\Delta$ and $x = n\omega_0 + \theta$ in (2.11), we can simplify (2.10) as follows:

$$\begin{aligned} e_n &= \sum_{l\neq 0} \frac{-\Delta}{2\pi i l}\left[\sum_{m=-\infty}^{\infty} J_m(2\pi l A/\Delta) e^{im(n\omega_0+\theta)}\right] \\ &= \sum_{m=-\infty}^{\infty}\left[\sum_{l\neq 0}\frac{-\Delta}{2\pi i l} J_m(2\pi l A/\Delta)e^{im\theta}\right] e^{imn\omega_0} \\ &= \sum_{m=-\infty}^{\infty} b_m e^{i(\omega_0 m)n}, \end{aligned} \tag{2.12}$$

where $b_m = \sum_{l\neq 0} \frac{-\Delta}{2\pi i l} J_m(2\pi l A/\Delta) e^{im\theta}$. Now,

$$\begin{aligned}
b_m &= \frac{-\Delta e^{im\theta}}{2\pi i}\left(\sum_{l=1}^{\infty}\frac{J_m(2\pi l A/\Delta)}{l} + \sum_{l=1}^{\infty}\frac{J_m(2\pi(-l)A/\Delta)}{-l}\right) \\
&= \frac{-\Delta e^{im\theta}}{2\pi i}\left(\sum_{l=1}^{\infty}\frac{J_m(2\pi l A/\Delta)}{l} + (-1)^{m+1}\sum_{l=1}^{\infty}\frac{J_m(2\pi(+l)A/\Delta)}{+l}\right) \qquad (2.13) \\
&= \begin{cases} 0, & \text{if } m \text{ is even} \\ \frac{-\Delta e^{im\theta}}{\pi i}\left(\sum_{l=1}^{\infty}\frac{J_m(2\pi l A/\Delta)}{l}\right), & \text{if } m \text{ is odd,} \end{cases} \qquad (2.14)
\end{aligned}$$

where (2.13) follows since Bessel functions satisfy the symmetry

$$J_m(-z) = (-1)^m J_m(z) = J_{-m}(z).$$

We have to verify (2.14). The first claim for m even, is clear. For odd indices, we shall show in Lemma 2.4 that

$$b_{2m+1} = \frac{-\Delta e^{i(2m+1)\theta}}{\pi i(2m+1)}\left(1 + 2\sum_{k=1}^{\lfloor A/\Delta\rfloor}\cos\left((2m+1)\sin^{-1}(\Delta k/A)\right)\right);$$

and this formula shows that $b_m \sim O(1/m)$, as $m \to \infty$. □

Lemma 2.4. *Let $J_m(z)$ be the Bessel function of order m defined in* (2.11)*. For $m = -1, 0, 1\ldots$, let $S(m) = \sum_{l=1}^{\infty}\frac{J_{2m+1}(2\pi l A/\Delta)}{l}$. Then*

$$\begin{aligned}
S(m) &= \frac{A\pi}{2\Delta}(\delta_{-1}[m] - \delta_0[m]) \\
&\qquad + \frac{1}{2m+1}\left(1 + 2\sum_{k=1}^{\lfloor A/\Delta\rfloor}\cos\left((2m+1)\sin^{-1}(k\Delta/A)\right)\right)
\end{aligned}$$

Proof. We first consider the analysis equation for the Fourier representation (2.11) and obtain the integral formula

$$\begin{aligned}
J_m(z) &= \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{iz\sin x} e^{-imx} dx \\
&= \frac{1}{2\pi}\int_{-\pi}^{\pi}\cos(z\sin x - mx)dx + \frac{i}{2\pi}\int_{-\pi}^{\pi}\sin(z\sin x - mx)dx \qquad (2.15) \\
&= \frac{1}{\pi}\int_{0}^{\pi}\cos(z\sin x - mx)dx \qquad (2.16)
\end{aligned}$$

where (2.16) follows since if $\Theta(x) = z\sin x - mx$, then

$$\Theta(-x) = z\sin(-x) + mx = -(z\sin x - mx) = -\Theta(x)$$

and therefore $\cos(\Theta(x))$ is even and $\sin(\Theta(x))$ is odd. So, if we split $\int_{-\pi}^{\pi}$ into $\int_{-\pi}^{0} + \int_{0}^{\pi}$ in (2.15), then the cos integrals combine and the sin integrals cancel. Next, we can further simplify (2.16) using the sum formulas for cos and sin. We obtain

$$\begin{aligned}
J_m(z) &= \frac{1}{\pi}\int_0^{\pi} \cos(z\sin x)\cos(mx) + \sin(z\sin x)\sin(mx)dx \\
&= \frac{1}{\pi}\int_0^{\pi/2} \cos(z\sin x)\cos(mx) + \sin(z\sin x)\sin(mx)dx \\
&\quad + \frac{1}{\pi}\int_{\pi/2}^{\pi} \cos(z\sin x)\cos(mx) + \sin(z\sin x)\sin(mx)dx \\
&= \frac{1}{\pi}\int_0^{\pi/2} \cos(z\sin x)\cos(mx) + \sin(z\sin x)\sin(mx)dx \\
&\quad + \frac{1}{\pi}\int_0^{\pi/2} (-1)^m\cos(z\sin x)\cos(mx) + (-1)^{m+1}\sin(z\sin x)\sin(mx)dx \\
&= \begin{cases} \frac{2}{\pi}\int_0^{\pi/2}\cos(z\sin x)\cos(mx)dx, & \text{if } m \text{ is even} \\ \frac{2}{\pi}\int_0^{\pi/2}\sin(z\sin x)\sin(mx)dx, & \text{if } m \text{ is odd.} \end{cases}
\end{aligned}$$

Using this to simplify $S(m)$, which was defined as the infinite sum in the formula for b_{2m+1} (equation (2.12)), we have

$$\begin{aligned}
S(m) &= \sum_{l=1}^{\infty}\frac{1}{l}\frac{2}{\pi}\int_0^{\pi/2}\sin\left(\frac{2\pi A}{\Delta}l\sin x\right)\sin((2m+1)x)dx \\
&= \frac{2}{\pi}\int_0^{\pi/2}\sin((2m+1)x)\left(\sum_{l=1}^{\infty}\frac{\sin\left(2\pi\left\langle\frac{A}{\Delta}\sin x\right\rangle l\right)}{l}\right)dx \qquad (2.17)
\end{aligned}$$

From a standard calculation in Fourier analysis, the sum on the right side of equation (2.17) has the closed formula,

$$\sum_{l=1}^{\infty}\frac{\sin(l\theta)}{l} = \frac{\pi}{2} - \frac{\theta}{2}, \quad \text{for } 0 < \theta < 2\pi;$$

and so (2.17) becomes

$$S(m) = \frac{2}{\pi}\int_0^{\pi/2} \sin((2m+1)x)\left(\frac{\pi}{2} - \frac{2\pi\left\langle \frac{A}{\Delta}\sin x\right\rangle}{2}\right)dx$$
$$= \underbrace{\int_0^{\pi/2} \sin((2m+1)x)dx}_{I_1(m)} - 2\underbrace{\int_0^{\pi/2} \sin((2m+1)x)\left\langle \frac{A}{\Delta}\sin x\right\rangle dx.}_{I_2(m)}$$

Now, $I_1(m) = \frac{1}{2m+1}$. In order to compute $I_2(m)$ we first need a formula for $\left\langle \frac{A}{\Delta}\sin x\right\rangle$ for $x \in [0, \pi/2]$. Observe, for $x \in [0, \pi/2]$, that

$$\left\langle \frac{A}{\Delta}\sin x\right\rangle = \frac{A}{\Delta}\sin x - k, \quad \text{when } k \le \frac{A}{\Delta}\sin x < k+1,$$

i.e., when $\sin^{-1}\left(\frac{k\Delta}{A}\right) \le x < \sin^{-1}\left(\frac{(k+1)\Delta}{A}\right)$. Therefore, set $K = \lfloor A/\Delta \rfloor$, assume $\lfloor A/\Delta \rfloor \ne A/\Delta$, and set

$$\alpha_0 = 0,$$
$$\alpha_k = \sin^{-1}\left(\frac{\Delta}{A}k\right), \quad \text{for } k = 1, 2, \ldots, K,$$
$$\alpha_{K+1} = \pi/2,$$

see Figure 2.2. Then we have

$$\left\langle \frac{A}{\Delta}\sin x\right\rangle = \sum_{k=0}^{K}\left(\frac{A}{\Delta}\sin x - k\right)\mathbf{1}_{[\alpha_k,\alpha_{k+1})}(x). \tag{2.18}$$

Using (2.18), we can compute $I_2(m)$,

$$I_2(m) = \int_0^{\pi/2} \sin((2m+1)x)\sum_{k=0}^{K}\left(\frac{A}{\Delta}\sin x - k\right)\mathbf{1}_{[\alpha_k,\alpha_{k+1})}(x)dx$$
$$= \sum_{k=0}^{K}\underbrace{\int_{\alpha_k}^{\alpha_{k+1}} \sin((2m+1)x)\frac{A}{\Delta}\sin x dx}_{I_3(m)} - \underbrace{\sum_{k=0}^{K}\int_{\alpha_k}^{\alpha_{k+1}} k\sin((2m+1)x)dx,}_{S_4(m)}$$

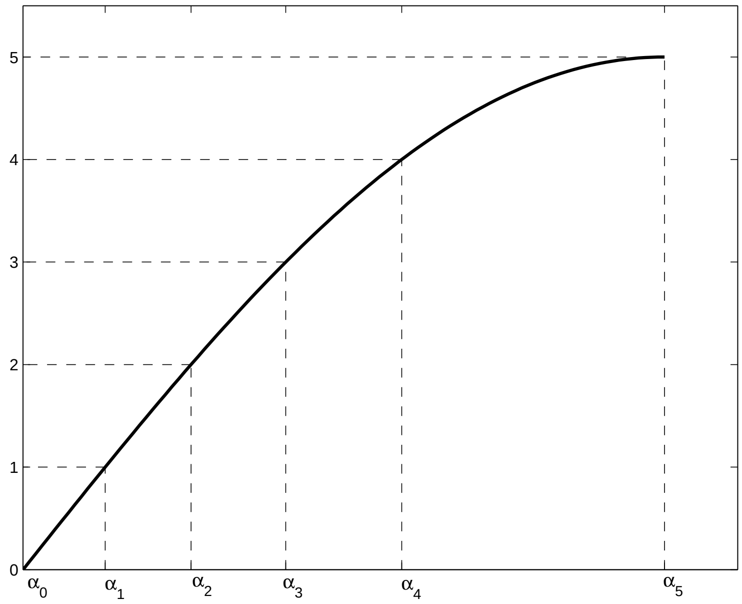


Figure 2.2: Computing $\left\langle \frac{A}{\Delta}\sin(x)\right\rangle$ with $A = 5$ and $\Delta = 1$. The curve is $y = 5\sin(x)$, $\alpha_0 = 0$, and $\alpha_5 = \pi/2$.

and we can compute $I_3(m)$ using standard trigonometric formulas. In fact,

$$I_3(m) = \begin{cases} \frac{-A}{2\Delta}\frac{\pi}{2}, & \text{if } m = -1 \\ \frac{A}{2\Delta}\frac{\pi}{2}, & \text{if } m = 0 \\ 0, & \text{otherwise} \end{cases} .$$

Furthermore, $S_4(m)$ has a telescoping property,

$$\begin{aligned}
S_4(m) &= \sum_{k=0}^{K} \frac{-k\cos((2m+1)x)}{2m+1}\bigg|_{\alpha_k}^{\alpha_{k+1}} \\
&= \sum_{k=0}^{K} \frac{-k}{2m+1}\cos((2m+1)\alpha_{k+1}) + \sum_{k=0}^{K} \frac{k}{2m+1}\cos((2m+1)\alpha_k) \\
&= \frac{-K}{2m+1}\cos\left((2m+1)\frac{\pi}{2}\right) + \frac{0}{2m+1}\cos((2m+1)0) \\
&\quad + \sum_{k=1}^{K} \frac{k-(k-1)}{2m+1}\cos((2m+1)\alpha_k) \\
&= \frac{1}{2m+1}\sum_{k=1}^{K}\cos((2m+1)\alpha_k).
\end{aligned}$$

Thus,

$$\begin{aligned}
S(m) &= I_1(m) + I_2(m) \\
&= I_1(m) + I_3(m) + S_4(m) \\
&= \frac{1}{2m+1} + \frac{A\pi}{2\Delta}[\delta_{-1} - \delta_0] + \frac{2}{2m+1}\sum_{k=1}^{\lfloor A/\Delta \rfloor} \cos\left((2m+1)\sin^{-1}(k\Delta/A)\right)
\end{aligned}$$

□

Now using this result, since $b_{2m+1} = \frac{-\Delta e^{i(2m+1)x}}{\pi i(2m+1)} S(m)$, we obtain the formula (2.9) in Proposition 2.3.

2.4 Growth of b_m

With formula (2.9), we see that $\{b_m\} \in \ell^2(\mathbb{Z})$,

$$\begin{aligned}
&\sum_{m=-\infty}^{\infty} |b_m|^2 \\
&= 2\left(\frac{\pi A}{2\Delta}\right)^2 + \frac{1}{\pi^2}\sum_{m=-\infty}^{\infty}\left(\frac{1}{2m+1}\right)^2 \left(1 + 2\sum_{k=1}^{\lfloor A/\Delta \rfloor} \cos\left((2m+1)\sin^{-1}(k\Delta/A)\right)\right) \\
&\leq 2\left(\frac{\pi A}{2\Delta}\right)^2 + \frac{1}{\pi^2}\sum_{m=-\infty}^{\infty}\left(\frac{1}{2m+1}\right)^2 (1 + 2\lfloor A/\Delta \rfloor) \\
&< 2\left(\frac{\pi A}{2\Delta}\right)^2 + \frac{1}{\pi^2}2\left(\frac{\pi^2}{6}\right)(1 + 2\lfloor A/\Delta \rfloor) \\
&= \frac{\pi^2 A^2}{2\Delta^2} + \frac{1}{3}(1 + 2\lfloor A/\Delta \rfloor) \\
&< \infty
\end{aligned}$$

We can also see that (for $m \neq 0, 1$) the size of $|b_{2m+1}|^2$ is ultimately controlled by the size of $\left(\frac{1}{2m+1}\right)^2$:

$$|b_{2m+1}|^2 \leq \frac{1}{\pi^2}\left|\frac{1}{2m+1}\right|^2 |1 + 2\lfloor A/\Delta \rfloor|^2 = O\left(\left(\frac{1}{2m+1}\right)^2\right), \quad m \to \infty.$$

2.5 Power spectrum of e_n

Next, we compute the power spectrum of the quantization error e_n associated with a sinusoidal input $w_n = A\sin(n\omega_0 + \theta)$. We have two cases, $\omega_0 \in 2\pi\mathbb{Q}$ and $\omega_0 \notin 2\pi\mathbb{Q}$.

Proposition 2.5. *Let $e_n = \sum_{m=-\infty}^{\infty} b_m e^{i(\omega_0 m)n}$ be the quantization error associated with a sinusoidal input $w_n = A\sin(n\omega_0 + \theta)$ computed in Proposition 2.3, and let $\omega_0 = 2\pi\alpha/\beta$, where $\alpha < \beta$. Assume for every $p \in \{0, 1, \ldots, \beta - 1\}$, we* *have*

$$\left|\sum_{m'\in\mathbb{Z}} b_{m'\beta+p}\right| < \infty.$$

Then the power spectrum of e_n *is*

$$S_e = (r_e)^\wedge = \sum_{p=0}^{\beta-1} |c_p|^2 \delta_{\left\langle \frac{\alpha p}{\beta} \right\rangle},$$

where $c_p = \sum_{m'\in\mathbb{Z}} b_{m'\beta+p}$.

Proof. Let $\gamma_0 = \alpha/\beta$, so $\omega_0 = 2\pi\alpha/\beta$. Let $m = m'\beta + p$ where $m' \in \mathbb{Z}$ and $p = 0, 1, \ldots, \beta - 1$. Then (2.12) becomes

$$e_n = \sum_{p=0}^{\beta-1} \sum_{m'\in\mathbb{Z}} b_{m'\beta+p} e^{2\pi i \frac{\alpha}{\beta}(m'\beta+p)n}. \tag{2.19}$$

Clearly, $e^{2\pi i \frac{\alpha}{\beta}(m'\beta+p)n} = e^{2\pi i \frac{\alpha}{\beta} pn}$; so that if we let $c_p = \sum_{m'\in\mathbb{Z}} b_{m'\beta+p}$, then (2.19) becomes

$$e_n = \sum_{p=0}^{\beta-1} c_p e^{2\pi i \frac{\alpha}{\beta} pn} = \sum_{p=0}^{\beta-1} c_p e^{i\omega_0 pn}. \tag{2.20}$$

Since by assumption $|c_p| < \infty$ for all p, we have that e_n is well defined and therefore β-periodic in n.

Now, for a β-periodic function, the deterministic autocorrelation is

$$r_e(k) = \frac{1}{\beta} \sum_{n=0}^{\beta-1} e_{k+n} \overline{e_n}. \tag{2.21}$$

Thus, substituting (2.20) into (2.21), we obtain

$$
\begin{aligned}
r_e(k) &= \frac{1}{\beta}\sum_{n=0}^{\beta-1}\left(\sum_{p=0}^{\beta-1} c_p e^{i\omega_0 p(k+n)}\right)\left(\sum_{q=0}^{\beta-1}\overline{c_q}e^{-i\omega_0 qn}\right)\\
&= \frac{1}{\beta}\sum_{p=0}^{\beta-1}\sum_{q=0}^{\beta-1} c_p\overline{c_q}\left(\sum_{n=0}^{\beta-1} e^{i\omega_0[p(k+n)-qn]}\right)\\
&= \frac{1}{\beta}\sum_{p=q}|c_p|^2 e^{i\omega_0 pk}\left(\sum_{n=0}^{\beta-1}1\right)\\
&\qquad + \frac{1}{\beta}\sum_{p\neq q} c_p\overline{c_q}e^{i\omega_0 pk}\left(\sum_{n=0}^{\beta-1}\left(e^{i\omega_0(p-q)}\right)^n\right) \qquad (2.22)\\
&= \sum_{p=q}|c_p|^2 e^{i\omega_0 pk} + 0, \qquad (2.23)
\end{aligned}
$$

where (2.23) follows since the sum in parentheses in (2.22) is zero by the geometric series formula, i.e.,

$$
\sum_{n=0}^{\beta-1}\left(e^{i\omega_0(p-q)}\right)^n = \frac{1-\left(e^{2\pi i\frac{\alpha}{\beta}(p-q)}\right)^n}{1-e^{2\pi i\frac{\alpha}{\beta}(p-q)}} = \frac{1-1}{1-e^{2\pi i\frac{\alpha}{\beta}(p-q)}} = 0.
$$

From equation (2.23) we can compute the power spectrum of e_n. First $(\delta_{\lambda_0})^\vee(k) = e^{+2\pi ik\lambda_0}$, so setting $\lambda_0 = \frac{\alpha p}{\beta}$, we have

$$
r_e(k) = \sum_{p=0}^{\beta-1}|c_p|^2\left(\delta_{\frac{\alpha p}{\beta}}\right)^\vee(k) = \left(\sum_{p=0}^{\beta-1}|c_p|^2\delta_{\frac{\alpha p}{\beta}}\right)^\vee(k).
$$

Hence if $\omega_0 = 2\pi\alpha/\beta$, then the power spectrum of e_n is

$$
S_e = (r_e)^\wedge = \sum_{p=0}^{\beta-1}|c_p|^2\delta_{\frac{\alpha p}{\beta}},
$$

□

Thus, if $\omega_0 = 2\pi\frac{\alpha}{\beta}$, where $\alpha < \beta$ and the greatest common divisor of α and β is 1, then the quantization error e_n associated with the input $w_n = A\sin(n\omega_0 + \theta)$ has power concentrated at β evenly spaced deltas of height $|c_p|^2$ where $p =$

$0, 1, \ldots, \beta - 1$. The assumption $|c_p| < \infty$ is necessary since (2.9) shows that $b_m \sim O(1/m)$, as $m \to \infty$, and therefore, $c_p \sim \sum_{m' \neq 0} 1/m'$.

If $\omega_0 \notin 2\pi\mathbb{Q}$, then a different type of simplification occurs.

Proposition 2.6. *Let $e_n = \sum_{m=-\infty}^{\infty} b_m e^{i(\omega_0 m)n}$ be the quantization error associated with a sinusoidal input $w_n = A\sin(n\omega_0 + \theta)$ computed in Proposition 2.3, and let $\omega_0 \notin 2\pi\mathbb{Q}$. Then the power spectrum of e_n is*

$$= \sum_{p=-\infty}^{\infty} |b_p|^2 \delta_{\langle(2p+1)\gamma_0\rangle}$$

where $\omega_0 = 2\pi\gamma_0$.

Proof. The deterministic autocorrelation of e_n is

$$r_e(k) = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} e_{n+k}\overline{e_n}.$$

Hence, we compute

$$\begin{aligned} r_e(k) &= \lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} e_{n+k}\overline{e_n} \\ &= \lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} \left(\sum_{p=-\infty}^{\infty} b_p e^{i\lambda_p(n+k)} \right) \left(\sum_{q=-\infty}^{\infty} \overline{b_q} e^{-i\lambda_p n} \right) && (2.24) \\ &= \sum_{p=-\infty}^{\infty} \sum_{q=-\infty}^{\infty} b_p \overline{b_q} e^{i\lambda_p k} \left(\lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} e^{in(\lambda_p - \lambda_q)} \right) \\ &= \sum_{p=q} |b_p|^2 e^{i\lambda_p k} \left(\lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} 1 \right) && (2.25) \\ &\quad + \sum_{p\neq q} b_p \overline{b_q} e^{i\lambda_p k} \left(\lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} e^{i\langle n(\lambda_p - \lambda_q)\rangle} \right) && (2.26) \end{aligned}$$

where the switching of lim and $\sum_{-\infty}^{\infty} \sum_{-\infty}^{\infty}$ in (2.24) is formal. Now, if we set $h(x) = e^{2\pi i x}$ and

$$\gamma_{p,q} = \frac{\lambda_p - \lambda_q}{2\pi} = \langle(2p+1)\gamma_0\rangle - \langle(2q+1)\gamma_0\rangle,$$

then since $\lambda_p - \lambda_q \notin 2\pi\mathbb{Q}$, we have that $\gamma_{p,q} \notin \mathbb{Q}$ and therefore Weyl's uniform distribution theorem applies, see Section 1.8. Hence the limit in parentheses in (2.26) becomes

$$\begin{aligned}\lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} e^{2\pi i \langle n\gamma_{p,q}\rangle} &= \lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} h\big(\, \langle n\gamma_{p,q}\rangle \,\big) \\ &= \int_0^1 h(x) d(x) = 0.\end{aligned}$$

Therefore, we have

$$r_e(k) = \sum_{p=-\infty}^{\infty} |b_p|^2 e^{i\lambda_p k}, \tag{2.27}$$

where $b_p = \sum_{l=1}^{\infty} \frac{-\Delta}{\pi i l} J_{2p+1}(2\pi l A/\Delta) e^{i(2p+1)\theta}$ and $\lambda_p = 2\pi \left\langle (2p+1)\gamma_0 \right\rangle$.

We can now compute the power spectrum of e. Since $(\delta_{\frac{\lambda_p}{2\pi}})^\vee = e^{i\lambda_p(\cdot)}$, we have

$$\begin{aligned} S_e = (r_e)^\wedge = \left(\sum_{p=-\infty}^{\infty} |b_p|^2 e^{i\lambda_p(\cdot)} \right)^\wedge &= \left(\sum_{p=-\infty}^{\infty} |b_p|^2 \left(\delta_{\frac{\lambda_p}{2\pi}}\right)^\vee \right)^\wedge \\ = \left(\sum_{p=-\infty}^{\infty} |b_p|^2 \delta_{\frac{\lambda_p}{2\pi}} \right)^{\vee\wedge} &= \sum_{p=-\infty}^{\infty} |b_p|^2 \delta_{\frac{\lambda_p}{2\pi}}, \end{aligned}$$

and $\frac{\lambda_p}{2\pi} = \frac{2\pi\langle(2p+1)\gamma_0\rangle}{2\pi} = 2\pi \left\langle (2p+1)\gamma_0 \right\rangle$. □

Thus, for $\omega_0 \notin \mathbb{Q}$, the quantization error associated with the input $w_n = A\sin(n\omega_0 + \theta)$ has frequency components whose magnitude squared is $|b_p|^2$ at the points of the uniformly distributed sequence $\langle (2p+1)\gamma_0 \rangle$, where $\omega_0 = 2\pi\gamma_0$.

2.6 Power spectra and $\Sigma\Delta$ modulation error

Next, we demonstrate how the above analysis can be applied to the $\Sigma\Delta$ modulator (2.1). Rewrite the $\Sigma\Delta$ modulator (2.1) and (2.2) as

$$\begin{cases} w_n = x_n + e_{n-1} \\ e_n = w_n - Q(w_n). \end{cases} \tag{2.28}$$

where we have replaced $f(nT)$ by x_n, and where Q is an M level, Δ space uniform quantizer. We first seek to show that for bounded inputs x_n, the modified input w_n remains in the no-overload interval for Q.

Proposition 2.7. *Let $B > 0$, and let $\{x_n\}$ be a sequence bounded by B, i.e., $|x_n| \leq B$. Let $M \in \mathbb{N}$, let $\Delta \leq \frac{2B}{M-1}$, and let Q be the associated M level, Δ spaced uniform quantizer. Let $|e_0| \leq \frac{\Delta}{2}$, and for $n \geq 1$, let w_n and e_n be defined by the $\Sigma\Delta$ recursion* (2.28)*. Then $|e_n| \leq \frac{\Delta}{2}$, hence we can deduce w_n lies in the no-overload region for Q, i.e., $|w_n| \leq \frac{M\Delta}{2}$.*

Proof. First note that $\Delta \leq \frac{2B}{M-1}$ implies that $B \leq \frac{(M-1)\Delta}{2}$, hence $|e_{n-1}| \leq \frac{\Delta}{2}$ implies

$$|w_n| = |x_n + e_{n-1}| \leq |x_n| + |e_{n-1}| \leq B + \frac{\Delta}{2} \leq \frac{(M-1)\Delta}{2} + \frac{\Delta}{2} = \frac{M\Delta}{2}.$$

Thus, if the quantization error is bounded by $\frac{\Delta}{2}$, then the modified input w_n is in the no-overload interval. Proceeding by induction, we first note that the base case, $n = 0$, is satisfied by assumption, i.e., $|e_0| \leq \frac{\Delta}{2}$.

For the induction step, let $n > 0$ and assume $|e_{n-1}| \leq \frac{\Delta}{2}$. By the above observation, w_n lies in the no-overload interval, hence (2.7) is valid, i.e.,

$$e(w_n) = -\Delta\left(\frac{1}{2} - \left\langle\frac{w_n}{\Delta}\right\rangle\right).$$

Hence,

$$|e_n| = |e(w_n)| = \Delta \Big| \underbrace{\underbrace{\left\langle \frac{w_n}{\Delta} \right\rangle}_{\in[0,1]} - \frac{1}{2}}_{\in[-1/2,1/2]} \Big|,$$

so by induction, $|e_n| \leq \frac{\Delta}{2}$ for all $n \in \mathbb{N}$. □

By virtue of Proposition 2.7, we can use Equations (2.7) and (2.8) to derive a formula for the quantizer error in a $\Sigma\Delta$ for a given bounded input x_n.

Proposition 2.8. *Under the same assumptions as Proposition 2.7, if* $e_0 = \frac{\Delta}{2}$*, then*

$$e_n = -\Delta \left(\frac{1}{2} - \left\langle \frac{\frac{-n}{2} + \sum_{k=0}^{n} x_k}{\Delta} \right\rangle \right).$$

Proof. By Proposition 2.7, since w_n is in the no-overload interval, we have that

$$e_n = -\Delta \left(\frac{1}{2} - \left\langle \frac{w_n}{\Delta} \right\rangle \right) = \frac{-\Delta}{2} + \Delta \left\langle \frac{x_n + e_{n-1}}{\Delta} \right\rangle, \qquad (2.29)$$

where the second equality follows from (2.28). Now, let $y_n = e_n + \frac{\Delta}{2}$. Then (2.29) implies

$$y_n = \Delta \left\langle \frac{x_n}{\Delta} + \frac{y_{n-1}}{\Delta} - \frac{1}{2} \right\rangle.$$

Now by induction, we have $y_n = \Delta \left\langle \frac{1}{\Delta} \sum_{k=1}^{n} x_k - \frac{n}{2} \right\rangle$, since for $n = 0$, $y_0 = e_0 - \Delta/2 = 0$ and for $n > 0$, if $y_{n-1} = \Delta \left\langle \frac{1}{\Delta} \sum_{k=1}^{n-1} x_k - \frac{n-1}{2} \right\rangle$, then

$$\begin{aligned} y_n = \Delta \left\langle \frac{x_n}{\Delta} + \frac{y_{n-1}}{\Delta} - \frac{1}{2} \right\rangle &= \Delta \left\langle \frac{x_n}{\Delta} + \langle \frac{1}{\Delta} \sum_{k=1}^{n-1} x_k - \frac{n-1}{2} \rangle - \frac{1}{2} \right\rangle \\ &= \Delta \left\langle \frac{x_n}{\Delta} + \frac{1}{\Delta} \sum_{k=1}^{n-1} x_k - \frac{n-1}{2} - \frac{1}{2} \right\rangle = \Delta \left\langle \frac{1}{\Delta} \sum_{k=1}^{n} x_k - \frac{n}{2} \right\rangle. \end{aligned}$$

Substituting this into $e_n = y_n - \frac{\Delta}{2}$ we have,

$$e_n = -\Delta \left(\frac{1}{2} - \left\langle \frac{\frac{-n}{2} + \sum_{k=0}^{n} x_k}{\Delta} \right\rangle \right).$$

□

Now, since (2.7) is equivalent to (2.8), we have

$$\begin{aligned} e_n &= \sum_{l \neq 0} \frac{-\Delta}{2\pi i l} e^{2\pi i l \left(\frac{-n}{2} + \sum_{k=1}^{n} x_k \right) / \Delta} \\ &= \sum_{l \neq 0} \frac{-\Delta}{2\pi i l} e^{-\pi i n l/\Delta} e^{2\pi i l x_0/\Delta} \cdots e^{2\pi i l x_n/\Delta}. \end{aligned}$$

So, to compute the power spectrum of the quantization error arising from a $\Sigma\Delta$ modulator with a bounded input, we must simplfy the above expression. This is a possible direction of future research.

Chapter 3

Quantization in Higher Dimensions

Next we consider the problem of quantizing two dimensional functions. Thus we could consider either $x : \mathbb{Z}^2 \to \mathbb{R}$, $x : \mathbb{Z} \to \mathbb{R}^2$, or $x : \mathbb{Z}^2 \to \mathbb{R}^2$. We shall only study the first case here. We would like to develop schemes which behave similar to the $\Sigma\Delta$ scheme in one dimension. For example, given a two dimensional sequence of samples $x_{m,n}$ on $\{0, 1, \ldots, M\} \times \{0, 1, \ldots, N\}$, we wish to construct a binary sequence, $q_{m,n}$, satisfying the stability condition that if the input $x_{m,n}$ is bounded, then the sum of the differences $x_{m,n} - q_{m,n}$, is bounded. That is, for any B, there is a C_B such that $|x_{m,n}| \leq B$ implies $\sum_m \sum_n (x_{m,n} - q_{m,n}) \leq C_B$. See (1.4) for a brief explanation as to why this bound is important.

3.1 Standard scheme

We first generalize the one dimensional $\Sigma\Delta$ recursion which quantizes a function $x : \mathbb{Z} \to [-1, 1]$ by producing a function $q : \mathbb{Z} \to \{\pm 1\}$ such that $\sum_n (x_n - q_n) \leq 2 = m[-1, 1]$. Thus, we first consider a function $x : \{0, 1, \ldots, M\} \times$

$\{0, 1, \ldots, N\} \to [-1, 1]$ as an $(M+1) \times (N+1)$ matrix

$$\begin{pmatrix} x_{0,0} & \ldots & x_{0,N} \\ \vdots & \ddots & \vdots \\ x_{M,0} & \ldots & x_{M,N} \end{pmatrix}.$$

We construct $(M+1) \times (N+1)$ matrices u, q as follows. Given $u_{0,0} = c$, where c is a constant, construct the 0th column and 0th row of u and q using a one dimensional $\Sigma\Delta$ scheme. That is, for column 0 and for $m > 0$, define $u_{m,0}$ and $q_{m,0}$ recursively by,

$$\begin{cases} u_{m,0} = u_{m-1,0} + x_{m,0} - q_{m,0} \\ q_{m,0} = Q(u_{m-1,0} + x_{m,0}) \end{cases} \tag{3.1}$$

where $Q(y) = \text{sign}(y)$. Likewise, for row 0 and for $n > 0$, define $u_{0,n}$ and $q_{0,n}$ satisfying

$$\begin{cases} u_{0,n} = u_{0,n-1} + x_{0,n} - q_{0,n} \\ q_{0,n} = Q(u_{0,n-1} + x_{0,n}). \end{cases} \tag{3.2}$$

Thus we have the leftmost row and topmost column of both u and q defined, i.e.,

$$\begin{pmatrix} u_{0,0} & \ldots & u_{0,N} \\ \vdots & ? & ? \\ u_{M,0} & ? & ? \end{pmatrix}, \quad \begin{pmatrix} q_{0,0} & \ldots & q_{0,N} \\ \vdots & ? & ? \\ q_{M,0} & ? & ? \end{pmatrix}.$$

Next, to define the inside of u and q, for $n, m \geq 1$, use the initial data on the edges of u and q and recursively construct $u_{m,n}$ and $q_{m,n}$ to satisfy

$$\begin{cases} u_{m,n} &= u_{m-1,n} - u_{m-1,n-1} + u_{m,n-1} + x_{m,n} - q_{m,n} \\ q_{m,n} &= Q(u_{m-1,n} - u_{m-1,n-1} + u_{m,n-1} + x_{m,n}), \end{cases} \tag{3.3}$$

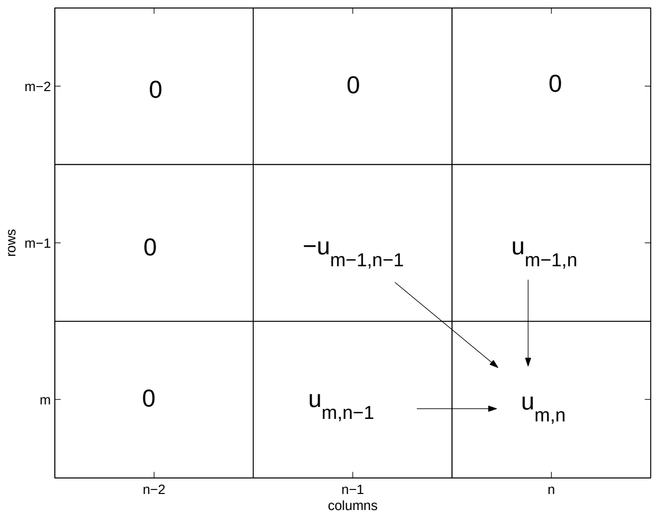


Figure 3.1: Graphical representation of the 2 dimensional $\Sigma\Delta$ recursion. Different schemes can be constructed by considering more nonzero entries when recursively defining $u_{m,n}$.

see Figure 3.1. Note, we have some freedom in what order we will construct the $u_{m,n}$ and $q_{m,n}$. We can work down rows, across columns or along consecutive reverse subdiagonals (similar to a common rule used to enumerate the rational numbers). Also note that because we are using (3.3), we have the following constraint. When we are constructing the (m, n)-entry, we must have already constructed the $(m-1, n)$, $(m, n-1)$ and $(m-1, n-1)$-entries. (Note that the diagonal scheme does generalize to an infinite dimensional input function $x_{m,n}$ easily.)

Using the above recursions, we have

Proposition 3.1. *Let* $x : \{0, \ldots, M\} \times \{0, \ldots, N\} \to [-1, 1]$*, and let* $u_{m,n}$ *and* $q_{m,n}$ *be defined by recursions* (3.1)*,* (3.2)*, and* (3.3)*. For any* $1 \leq m \leq M$ *and*

$1 \leq n \leq N$,

$$\sum_{j=1}^{m}\sum_{k=1}^{n}(x_{j,k} - q_{j,k}) = u_{0,0} - u_{m,0} - u_{0,n} + u_{m,n}. \tag{3.4}$$

Furthermore,

$$\sum_{\substack{j=1 \\ (j,k)\neq(0,0)}}^{m}\sum_{k=1}^{n}(x_{j,k} - q_{j,k}) = u_{m,n} - u_{0,0}. \tag{3.5}$$

Proof. Now at any point $(m, n) \in \{0, 1, \ldots, M\} \times \{0, 1, \ldots, N\}$, (3.3) implies

$$\begin{aligned}
&\sum_{j=1}^{m}\left(\sum_{k=1}^{n} x_{j,k} - q_{j,k}\right) \\
&= \sum_{j=1}^{m}\sum_{k=1}^{n}[u_{j,k} - u_{j-1,k} + u_{j-1,k-1} - u_{j,k-1}] \\
&= \sum_{j=1}^{m}\sum_{k=1}^{n}[u_{j,k} + (-u_{j,k} + u_{j,k}) - u_{j-1,k} + u_{j-1,k-1} - u_{j,k-1}] \\
&= \underbrace{\sum_{j=1}^{m}\sum_{k=1}^{n}(u_{j,k} - u_{j-1,k})}_{d_1} + \underbrace{\sum_{j=1}^{m}\sum_{k=1}^{n}(u_{j-1,k-1} - u_{j,k})}_{d_2} + \underbrace{\sum_{j=1}^{m}\sum_{k=1}^{n}(u_{j,k} - u_{j,k-1})}_{d_3},
\end{aligned}$$

and if we closely inspect these sums, we see cancellation due to telescoping,

$$d_1 = \sum_{k=1}^{n}\sum_{j=1}^{m}(u_{j,k} - u_{j-1,k}) = \sum_{k=1}^{n}(u_{m,k} - u_{0,k}),$$

$$d_3 = \sum_{j=1}^{m}\sum_{k=1}^{n}(u_{j,k} - u_{j,k-1}) = \sum_{j=1}^{m}(u_{j,n} - u_{j,0}),$$

and

$$\begin{aligned} d_2 &= \sum_{j=1}^{m}\sum_{k=1}^{n}(u_{j-1,k-1} - u_{j,k}) \\ &= \sum_{j=1}^{m}\sum_{k=1}^{n}(u_{j-1,k-1} - u_{j-1,k} + u_{j-1,k} - u_{j,k}) \\ &= \sum_{j=1}^{m}\sum_{k=1}^{n}(u_{j-1,k-1} - u_{j-1,k}) + \sum_{k=1}^{n}\sum_{j=1}^{m}(u_{j-1,k} - u_{j,k}) \\ &= \sum_{j=1}^{m}(u_{j-1,0} - u_{j-1,n}) + \sum_{k=1}^{n}(u_{0,k} - u_{m,k}). \end{aligned}$$

So when we add d_1, d_2, d_3, we have

$$\begin{aligned} &\sum_{j=1}^{m}\sum_{k=1}^{n}(x_{j,k} - q_{j.k}) \\ &\quad = d_1 + d_2 + d_3 \\ &\quad = \sum_{j=1}^{m}(u_{j,n} - u_{j,0} + u_{j-1,0} - u_{j-1,n}) + \sum_{k=1}^{n}(u_{m,k} - u_{0,k} + u_{0,k} - u_{m,k}) \\ &\quad = \sum_{j=1}^{m}(u_{j-1,0} - u_{j,0}) + \sum_{j=1}^{m}(u_{j,n} - u_{j-1,n}) + 0 \\ &\quad = u_{0,0} - u_{m,0} - u_{0,n} + u_{m,n}, \end{aligned}$$

and we have shown (3.4). Furthermore, since the 0th row and column satisfy the one dimensional $\Sigma\Delta$ recursion, by induction,

$$\begin{aligned} u_{0,0} - u_{m,0} - u_{0,n} + u_{m,n} &= (u_{0,0} - u_{m,0}) + (u_{0,0} - u_{0,n}) + (u_{m,n} - u_{0,0}) \\ &= \left(\sum_{j=1}^{m}(x_{j,0} - q_{j,0})\right) + \left(\sum_{k=1}^{n}(x_{0,k} - q_{0,k})\right) + (u_{m,n} - u_{0,0}). \end{aligned}$$

Thus, if we bring the sums to the other side of the equation in (3.4), we have proven (3.5). □

If we let $\Delta_j(u_{j,k}) = u_{j,k} - u_{j-1,k}$, and $\Delta_k(u_{j,k}) = u_{j,k} - u_{j,k-1}$, then we see that the first line of recursion (3.3) becomes $x_{j,k} - q_{j,k} = \Delta_j\Delta_k(u_{j,k})$, [Yıl02]. Since

$\sum_{j=1}^{m}\sum_{k=1}^{n}$ and $\Delta_j\Delta_k$ are inverses of each other, we see a heuristic reason why recursion(3.3) implies Proposition 3.1. By this reasoning, it is intuitively clear how we would generalize to quantizing functions whose domain is a subset of $\mathbb{Z}^d$, i.e., d-dimensional $\Sigma\Delta$.

Since we now have the sum of the quantization errors bounded by the internal state sequence, $u_{m,n}$, we seek to have control over the size of $u_{m,n}$, given that our input $x(m,n)$ is bounded by 1, i.e., $x_{m,n} < 1$. Interestingly, there is no such control as Proposition 3.2 shows.

Proposition 3.2. *For any $B > 0$, there exists $K = K_B \in \mathbb{N}$ and a bounded signal $x : \{0,\ldots,K\} \times \{0,\ldots,K\} \to [-1,1]$, such that if $u_{m,n}$ and $q_{m,n}$ are defined by the two dimensional $\Sigma\Delta$ recursion* (3.1)*,* (3.2)*,* (3.3)*, then*

$$\max_{m,n\in\{0,\ldots,K\}} |u_{m,n}| > B.$$

Proof. Let $B > 0$ be given, choose $K \in \mathbb{N}$ such that $K > 1 + \frac{B}{2}$. For $(m,n) \in \{0,1,\ldots,K\} \times \{0,1,\ldots,K\}$, set

$$x_{m,n} = \begin{cases} 1 - \frac{1}{K}, & \text{if } n = 0 \text{ and } m = 0,\ldots,K, \\ 1 - \frac{1}{K}, & \text{if } m = 0 \text{ and } n = 0,\ldots,K, \\ -1, & \text{if } m,n \geq 1 \text{ and } m+n \leq K, \\ 1, & \text{if } m,n \geq 1 \text{ and } m+n \geq K, \end{cases} \tag{3.6}$$

see figure Figure 3.2 with $K = 10$. Let $u_{0,0} = 0$, and $m,n = 0,\ldots,K$, let $u_{m,n}$ and $q_{m,n}$ be defined by the two dimensional $\Sigma\Delta$ recursion (3.1), (3.2), (3.3). Then by induction we prove that

$$u_{m,n} = \begin{cases} -\frac{m+n}{K}, & \text{if } m+n \leq K, \\ \left\{2 - \frac{1}{K}\right\}(m+n) - 2K, & \text{if } m+n > K. \end{cases} \tag{3.7}$$

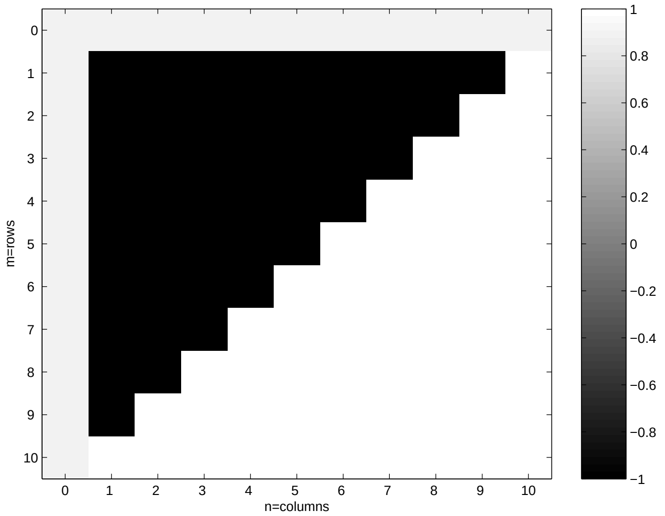


Figure 3.2: An example of the two dimensional signal $x_{m,n}$ defined in (3.6). Note $K = 10$ so $x_{m,n} = 0.9$ along the 0th row and column, $x_{m,n} = -1$ in the shaded upper left, and $x_{m,n} = 1$ in the unshaded lower right.

For the base case of our induction we notice that if $m = 0 = n$ then $u_{0,0} = 0 = \frac{0+0}{K}$.

Next, we induct along the 0th column. Let $n = 0$ and for $m = 1, 2, \ldots, K$, assume

$$u_{m-1,0} = -\frac{m-1+0}{K}. \tag{3.8}$$

then by (3.1), (3.6), and the induction hypothesis (3.8),

$$u_{m,0} = u_{m-1,0} + x_{m,0} - q_{m,0} = \left(1 - \frac{m}{K}\right) - \text{sign}\,(1 - \frac{m}{K}) = -\frac{m}{K}$$

and induction give the result along the 0th column. Since the 0th row is defined by (3.2) which is the same recursion as (3.1) except it runs along the row instead of the column, we have

$$u_{0,n} = -\frac{0+n}{K}, \quad \text{for } n = 0, \ldots, K \tag{3.9}$$

$$u_{m,0} = -\frac{m+0}{K}, \quad \text{for } m = 0, \ldots, K. \tag{3.10}$$

Now, let $P(n)$ be the statement

$$u_{m,n} = -\frac{m+n}{K}, \quad \text{for } m = 1, \ldots, K-n,$$

and let $Q(m, n)$ be the statement

$$u_{m,n} = -\frac{m+n}{K}.$$

Equation (3.10) shows $P(n)$ for $n = 0$. Now, for any $n = 1, 2, \ldots, K$ assume $P(n-1)$. For this fixed n, we next induct on m. Equation (3.9) shows $Q(0, n)$ and the base case $m = 0$ holds. Now, for $m = 1, \ldots, K-n$, assume the induction hypothesis $Q(m-1, n)$. Then $P(n)$, $Q(m-1, n)$, and recursion (3.3) imply

$$\begin{aligned} u_{m,n} &= u_{m-1,n} - u_{m-1,n-1} + u_{m,n-1} + x_{m,n} - q_{m,n} \\ &= -\frac{m-1+n}{K} - \frac{m+n-1}{K} + \frac{m-1+n-1}{K} + 1 - q_{m,n} \\ &= \left(1 - \frac{m+n}{K}\right) - \text{sign}\left(1 - \frac{m+n}{K}\right) \\ &= -\frac{m+n}{K}, \end{aligned}$$

where $1 - \frac{m+n}{K} > 0$, since $m \leq K - n$. So, by induction on m, we see that for this fixed n, $Q(m, n)$ for $m = 1, \ldots, K-n$, i.e., for this fixed n, $P(n)$. Thus by induction on n we have proven

$$u_{m,n} = -\frac{m+n}{K}, \quad \text{for } m + n \leq K, \tag{3.11}$$

which is the first line of (3.7). Thus in the region above the reverse diagonal, the internal state $|u_{m,n}|$ is bounded by 1, see Figure 3.3.

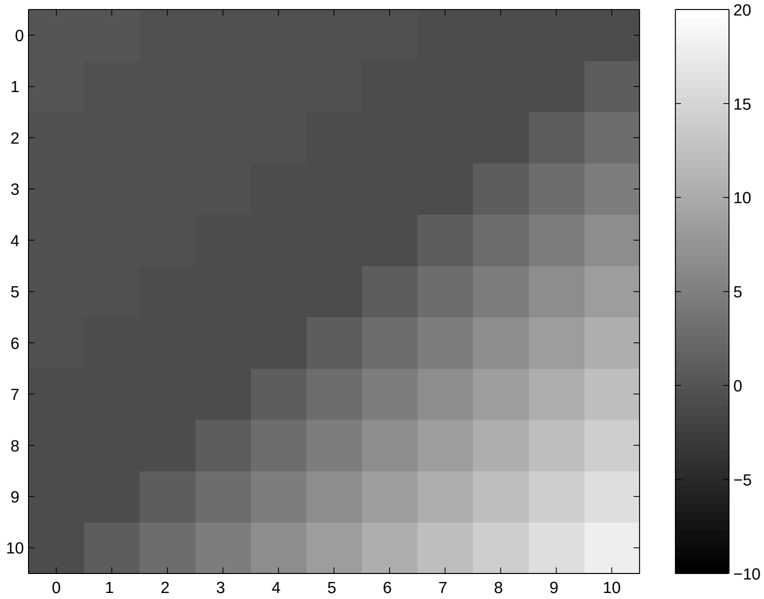


Figure 3.3: The internal state $u_{m,n}$ which satisfies (3.1), (3.2), (3.3) where $x_{m,n}$ is defined in (3.6) with $K = 10$, and so $B \leq 17$. Note $-1 \leq u_{m,n} \leq 18$.

Next we show that as we cross the reverse diagonal the internal state $u_{m,n}$ grows larger than the given B.

For $j = 0, \ldots, K$, let $m + n = K + j$. Proceeding by induction on j, we show that the second line of (3.7) notice for $j = 0$, (3.11) implies

$$u_{m,n} = -\frac{m+n}{K} = -1 = \left(2 - \frac{1}{K}\right)(m-n) - 2K.$$

For any $j = 1, \ldots, K$, assume

$$\left(2 - \frac{1}{K}\right)(m+n) - 2K, \quad \text{for } m+n \leq K - j + 1 \tag{3.12}$$

then for $m + n = j$, (3.3) and the induction hypothesis (3.12) imply

$$\begin{aligned} u_{m,n} &= u_{m-1,n} - u_{m-1,n-1} + u_{m,n-1} + x_{m,n} - q_{m,n} \\ &= 2(m+n-K) - \frac{m+n}{K} + 1 - \operatorname{sign}\left\{2(m+n-K) - \frac{m+n}{K} + 1\right\} \\ &= 2(m+n-K) - \frac{m+n}{K} \end{aligned} \tag{3.13}$$

where (3.13) follows since

$$\begin{aligned} 2(m+n-K) - \frac{m+n}{K} + 1 &= 2(K+j-K) - \frac{K+j}{K} + 1 \\ &= \left\{2 - \frac{1}{K}\right\} j \\ &> 0. \end{aligned}$$

Therefore we have shown (3.7). Finally, (3.7) implies $u_{K,K} = 2(K-1) > B$, see Figure 3.3. □

Reflecting on the proof of this proposition, we note that the constructed signal $x_{m,n}$ equals 1 when $u_{m,n}$ grows large, and 1 is the absolute value of the bit $q_{m,n}$. Thus, we next check that requiring $x_{m,n}$ to be bounded away from $|q_{m,n}|$ will reduce the size of the bound on the internal state $u_{m,n}$ in this specific example. That is, let

$$x_{m,n}(\alpha,\beta,\gamma) = \begin{cases} \alpha, & \text{if } n = 0 \text{ and } m = 0, \dots, K, \\ \alpha, & \text{if } m = 0 \text{ and } n = 0, \dots, K, \\ \beta, & \text{if } m, n \geq 1 \text{ and } m + n \leq K, \\ \gamma, & \text{if } m, n \geq 1 \text{ and } m + n \geq K, \end{cases} \tag{3.14}$$

and compare the result in Proposition 3.2 with the size of the bound on u when we input the sequence $x_{m,n}(\alpha,\beta,\gamma)$, with $\alpha = 1 - 1/K$, $\beta = .9$, and $\gamma = .9$, into the two dimensional $\Sigma\Delta$ scheme (3.1), (3.2), (3.3). For this example, when $K = 10$, the largest value of u decreases from 18 to 3.

3.2 Constant input

We observed in the last section that a signal x which is constant above and below the reverse diagonal repectively, and has a large jump across the reverse diagonal results in large bound for the internal state u. We also observed that if the input signal x is bounded away from $|q|$, then, in a specific example, the bound on u is significantly smaller. Hence, it is plausible that if $|x_{m,n}| \leq a < 1$ and $u_{m,n}$ is defined by the two dimensional $\Sigma\Delta$ recursion (3.3), then $|u_{m,n}| < B_a$.

In order to obtain some inutition on the relationship between a and B_a, we now reduce to the special case of a constant input signal, $x_{m,n} = a$. As Figure 3.4 shows, the bound on u decreases as the distance from a to $1 = |q_{m,n}|$ increases. Figure 3.4 also shows that the relationship between a and $\max(u_{m,n})$ is quite intricate.

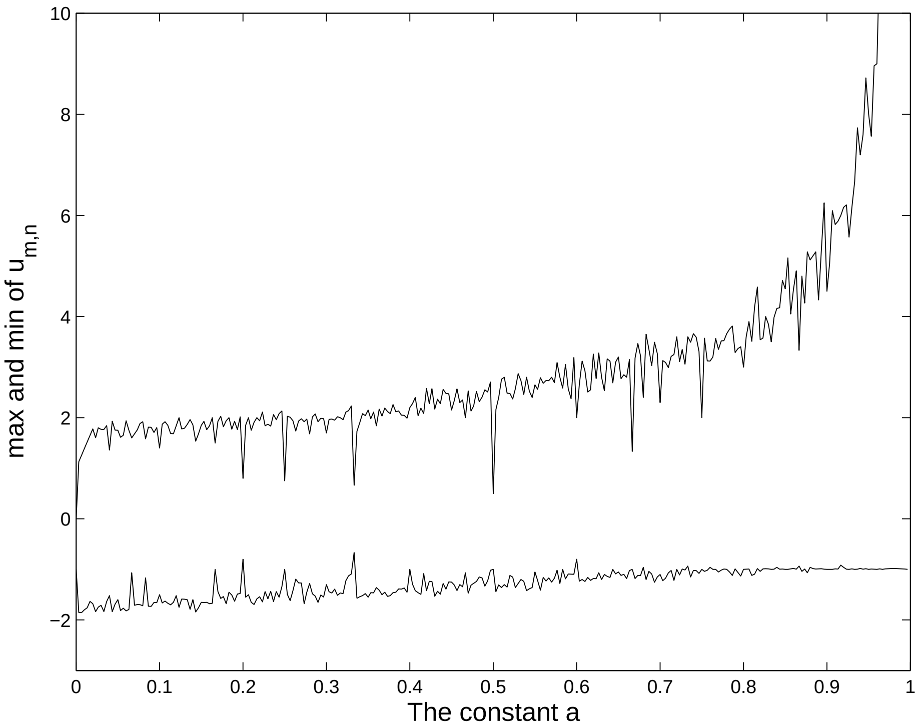


Figure 3.4: We let $x_{m,n} = a$ for $m, n = 0, \ldots 50$ and construct $u_{m,n}$ using (3.3). The top curve is a versus the maximum of u while the bottom is a versus the minimum of u.

3.3 A different quantization rule

The two dimensional $\Sigma\Delta$ recursion in (3.3) is not the only possible generalization of a one dimensional $\Sigma\Delta$ modulator.

The following is a scheme inspired by [DD03]. Let $|x_{m,n}| \leq a < 1$, $u_{0,0} = 0 = v_{0,0}$, and use (3.1) and (3.2) to construct the 0th row and 0th column of the matrices $u_{m,n}$ and $v_{m,n}$. For $m, n > 0$, construct $u_{m,n}$ and $v_{m,n}$ using the recursion

$$\begin{cases} v_{m,n} = v_{m-1,n} + x_{m,n} - q_{m,n} \\ u_{m,n} = u_{m,n-1} + v_{m,n} \\ q_{m,n} = \operatorname{sign}\left(v_{m-1,n} + x_{m,n} + C \operatorname{sign}(u_{m,n-1})\right) \end{cases} \tag{3.15}$$

where $C \geq 1 + 2a$. Note the more complicated quantization rule, and that $v_{m,n}$ is like a derivative with respect to m. Now

$$\begin{aligned} \sum_{j=1}^{m}\sum_{k=1}^{n}(x_{j,k} - q_{j,k}) &= \sum_{j=1}^{m}\sum_{k=1}^{n}(v_{m,n} - v_{m-1,n}) \\ &= \sum_{j=1}^{m}\sum_{k=1}^{n}\big((u_{m,n} - u_{m,n-1}) - (u_{m-1,n} - u_{m-1,n-1})\big) \\ &= u_{0,0} - u_{m,0} - u_{0,n} + u_{m,n}, \end{aligned}$$

where the last equality follws by (3.4). Hence, we want to show that $\|u_{m,n}\|_\infty < \infty$. Using the same ideas from [DD03], we show

Lemma 3.3. *For any $n > 0$, if $|v_{m-1,n}| \leq C + 1$ then $|v_{m,n}| \leq C + 1$.*

Proof. By the first line of (3.15),

$$\begin{aligned} v_{m,n} &= v_{m-1,n} + x_{m,n} - q_{m,n} \\ &= \underbrace{\underbrace{v_{m-1,n}}_{\in[-C-1,C+1]} + \underbrace{x_{m,n}}_{\in[-a,a]}}_{\in[-C-1-a,C+1+a]} - \operatorname{sign}\big(\underbrace{v_{m-1,n} + x_{m,n} + C \operatorname{sign}(u_{m,n-1})}_{w_{m,n}}\big) \end{aligned}$$

where we have set $w_{m,n} = v_{m-1,n} + x_{m,n} + C\,\mathrm{sign}(u_{m,n-1})$. We have three cases,

Case 1. $v_{m,n-1} + x_{m,n} \in (C, C+1+a]$

Then to compute $q_{m,n} = \mathrm{sign}(w_{m,n})$ we note that regardless of the sign of $u_{m,n-1}$, we have that $w_{m,n} \in (C, C+1+a] \pm C \subset (0, 2C+1+a]$. Hence $q_{m,n} = 1$ and

$$v_{m,n} = v_{m,n-1} + x_{m,n} - 1 \in (C, C+1+a] - 1 = (C-1, C+a] \subset [-C-1, C+1].$$

Case 2. $v_{m,n-1} + x_{m,n} \in [-C-1-a, -C)$

Again to compute $q_{m,n} = \mathrm{sign}(w_{m,n})$ we note that regardless of the sign of $u_{m,n-1}$, we have that $w_{m,n} \in [-C-1-a, -C) \pm C \subset [-2C-1-a, 0)$. Hence $q_{m,n} = -1$ and

$$v_{m,n} = v_{m,n-1} + x_{m,n} + 1 \in [-C-1-a, -C) + 1 = (-C-a, -C+1] \subset [-C-1, C+1].$$

Case 3. $v_{m,n-1} + x_{m,n} \in [-C, C]$ In this case we do not need to inspect $w_{m,n}$, we simply note that

$$v_{m,n} = v_{m,n-1} + x_{m,n} \pm 1 \in [-C, C] \pm 1 = [-C-1, C+1].$$

□

Unlike the one dimensional case in [DD03], Lemma 3.3 does not translate into a bound on $u_{m,n}$. This is because the bound in Lemma 3.3 is a bound for column n of $v_{m,n}$, but the second line of (3.15) shows that $v_{m,n} = u_{m,n} - u_{m,n-1}$ which is a row relationship. So, the rows of u are well behaved, but we have no control over the growth of the columns. in fact the signal in Proposition 3.2 grows large for (3.15) just as it did for (3.3).

3.4 Halftoning

The quantization rule considered in Section 3.3 is just one of many possible choices. In fact, there is a research field in image processing which studies the effect of different quantization rules, namely digital halftoning by error diffusion. Digital halftoning, refers to any process by which a continuous gray-scale image is converted to a binary image by the judicious arrangement of binary picture elements, see Figure 3.5, [Uli87, EFKM03, Kit98]. One method of digital halftoning is called error diffusion. Error difusion uses feedback to pick the binary picture elements. It is error diffusion which is seen to be a generalization of one dimensional $\Sigma\Delta$ modulation.

The first example of an error diffusion quantization rule is due to Floyd and Steinberg [FS76]. The rule is

$$\begin{cases} u_{m,n} = \frac{7}{16}u_{m-1,n} + \frac{1}{16}u_{m-1,n-1} + \frac{5}{16}u_{m,n-1} + \frac{3}{16}u_{m+1,n-1} - x_{m,n} + q_{m,n} \\ q_{m,n} = \operatorname{sign}\left(x_{m,n} - (\frac{7}{16}u_{m-1,n} + \frac{1}{16}u_{m-1,n-1} + \frac{5}{16}u_{m,n-1} + \frac{3}{16}u_{m+1,n-1})\right) \end{cases} \tag{3.16}$$

see Figure 3.6. We note that in (3.16), the sum of the weights is $\frac{7}{16}+\frac{1}{16}+\frac{5}{16}+\frac{3}{16} = 1$ and in (3.3) the sum of the wieghts is $1 - 1 + 1 = 1$. The fact that these weights sum to one means that in both schemes the error in the surronding pixels $u_{m-1,n}, u_{m-1,n-1}, \ldots$ is not amplified when it is used to construct $u_{m,n}$.

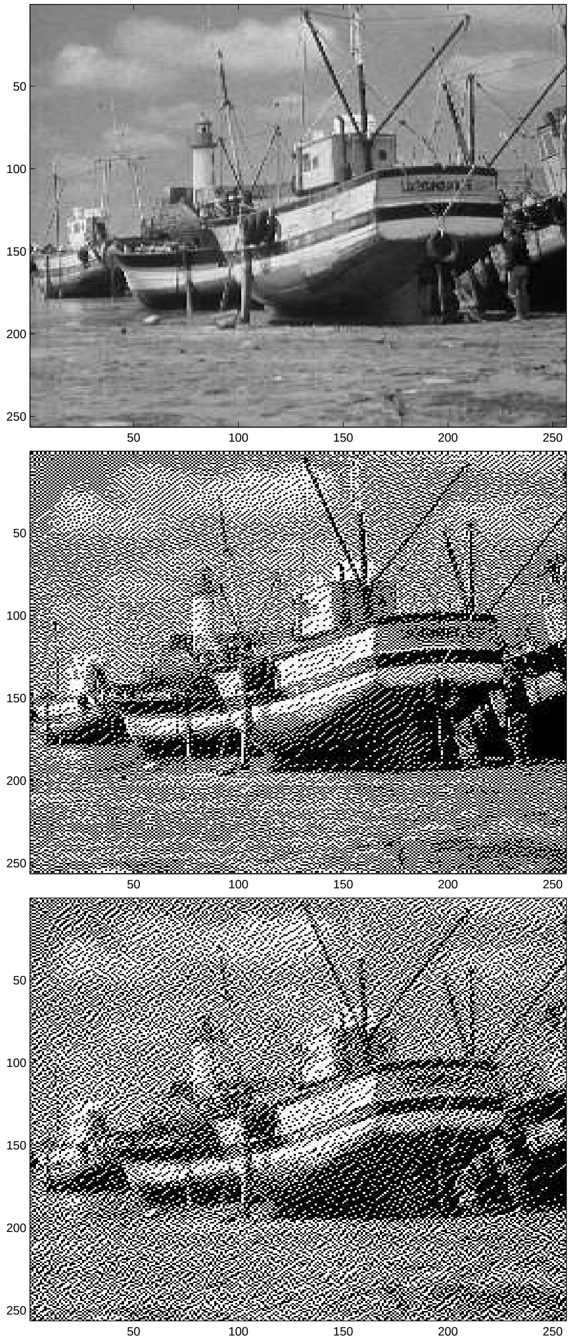


Figure 3.5: An example of halftoning a continuous gray-scale image. The top figure is a 256 × 256 continuous gray scale image, the middle figure is a halftone of the top using (3.16), the bottom figure is a halftone of the top using (3.3).

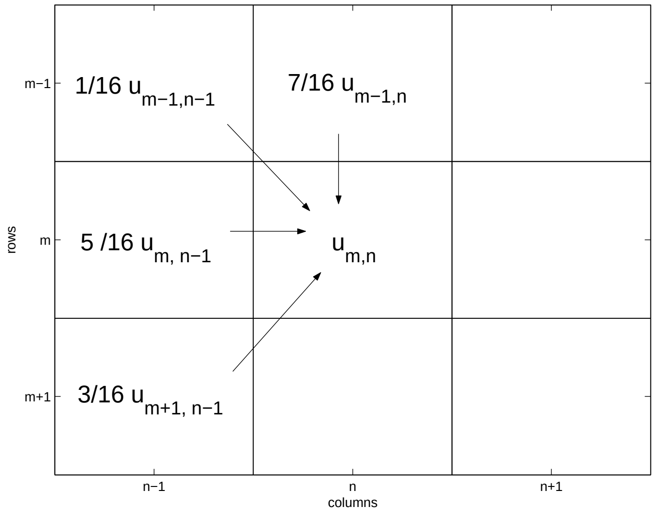


Figure 3.6: Graphical representation of the Floyd Steinberg error diffusion halftoning scheme, compare with Figure 3.1.

3.5 $\Sigma\Delta$ and space filling curves

Assume $|f(x_1, x_2)| \leq 1$ and that f is bandlimited to $(-\Omega_1, \Omega_1) \times (-\Omega_2, \Omega_2)$, i.e., $\hat{f}(\gamma_1, \gamma_2) = 0$ for $|\gamma_k| > \Omega_k$, $k = 1, 2$. Choose T_1, T_2 such that $2T_k\Omega_k \leq 1$ for $k = 1, 2$. Let H be the image of $\mathbb{Z}^2$ under the matrix $\begin{pmatrix} T_1 & 0 \\ 0 & T_2 \end{pmatrix}$, That is

$$H := \left\{ \begin{pmatrix} T_1 & 0 \\ 0 & T_2 \end{pmatrix} \begin{pmatrix} m \\ n \end{pmatrix} : m, n \in \mathbb{Z} \right\} \subseteq \mathbb{R}^2,$$

So a reciprocal lattice is

$$\Lambda := \left\{ \begin{pmatrix} \frac{1}{T_1} & 0 \\ 0 & \frac{1}{T_2} \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} : a, b \in \mathbb{Z} \right\} \subseteq \widehat{\mathbb{R}^2}.$$

Notice

$$\begin{pmatrix} T_1 & 0 \\ 0 & T_2 \end{pmatrix}^{-1} = \frac{1}{T_1 T_2} \begin{pmatrix} T_2 & 0 \\ 0 & T_1 \end{pmatrix} = \begin{pmatrix} \frac{1}{T_1} & 0 \\ 0 & \frac{1}{T_2} \end{pmatrix}.$$

$k =$	1	2	3	4	5	6	7	8	9
$\sigma(k) =$	(0,0)	(0,1)	(1,1)	(1,0)	(1,-1)	(0,-1)	(-1,-1)	(-1,0)	(-1,1)

Table 3.1: The first nine points of the bijection σ in Figure 3.7.

Consider the unit cell, $E \subseteq \widehat{\mathbb{R}^2}$ given by

$$E = \left(-\frac{1}{2T_1}, \frac{1}{2T_1}\right) \times \left(-\frac{1}{2T_2}, \frac{1}{2T_2}\right).$$

Now consider the sampling kernel associated with E, defined on $\mathbb{R}^2$ by

$$s_E(x_1, x_2) = \frac{1}{|E|}\int_E e^{2\pi i(x_1\gamma_1 + x_2\gamma_2)} d(\gamma_1 \times \gamma_2).$$

We could instead define $s_E \in C^\infty(\mathbb{R}^2)$ such that

$$\hat{s_E} = 1 \text{ on } (-\Omega_1, \Omega_1) \times (-\Omega_2, \Omega_2), \text{ and } \mathrm{supp}(\hat{s_E}) \subseteq E.$$

Then we have a two dimensional sampling theorem [BF01]. Namely, if f is continuous on $\mathbb{R}^2$, then

$$\lim_{r\to\infty}\left\| f(x_1, x_2) - \sum_{|m|,|n|\leq r}\sum f(mT_1, nT_2) s_E(x_1 - mT_1, x_2 - nT_2)\right\|$$

where the norm is either L^2 or L^∞.

A spiral $\Sigma\Delta$ Now, given the samples $f(mT_1, nT_2)$ construct $q_{m,n}^{T_T,T_2} = q_{m,n}$, using a spiral which fills out the lattice $\mathbb{Z}^2$. That is, let $\sigma = (\sigma_1, \sigma_2) : \mathbb{Z} \to \mathbb{Z}^2$ be an ordering of the integer points on the spiral in Figure 3.7, see Table 3.1. If $S_r = \{(m, n) \in \mathbb{Z}^2 : |m|, |n| \leq r\}$ and if $\partial S_r = \{(m, n) \in \mathbb{Z}^2 : |m| = r \text{ or } |n| = r\}$, then we see ∂S_r has $4 \cdot 2r$ points for $r \geq 1$ so S_r has $4(2 + 4 + \ldots + 2r) + 1$ points and $|S_r| = (2r-1)^2$. So for any $r \in \mathbb{N}$, σ is a bijection between $\{1, 2, \ldots, (2r-1)^2\} \subset \mathbb{Z}$ and the square $\{(m, n) \in \mathbb{Z}^2 : |m|, |n| \leq r\}$.

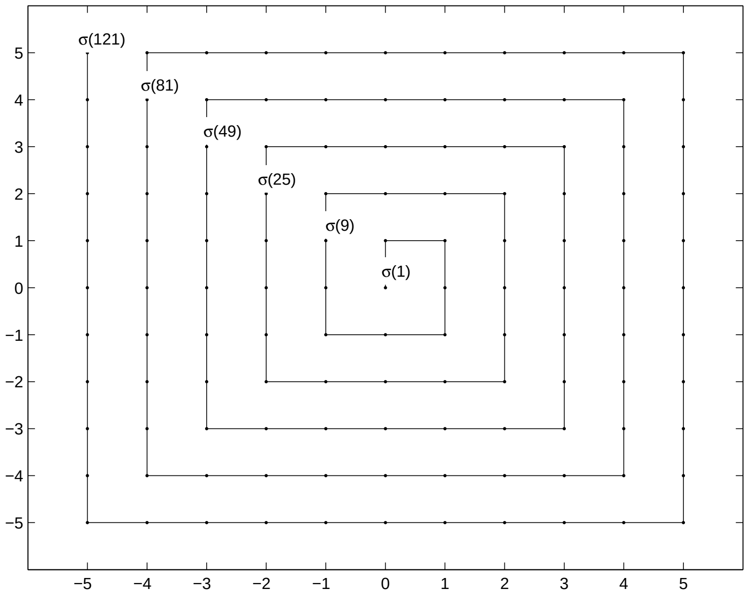


Figure 3.7: The bijection $\sigma : \mathbb{Z} \to \mathbb{Z}^2$ gives a natural ordering to the points on the spiral.

Now let $x_{m,n} = f(mT_1, nT_2)$ and construct $q_{m,n}$ using the initial value $u_{\sigma(0)} = c$ and the recursion

$$\begin{cases} u_{\sigma(k)} = u_{\sigma(k-1)} + x_{\sigma(k)} - q_{\sigma(k)} \\ q_{\sigma(k)} = \operatorname{sign}\left(u_{\sigma(k-1)} + x_{\sigma(k)}\right) \end{cases}$$

The using the one dimensional theory we have

$$\begin{aligned} \sum\sum_{|m|,|n|\le r} \left(f(mT_1, nT_2) - q_{m,n}^{T_T,T_2}\right) &= \sum_{k=1}^{(2r-1)^2} \left(x_{\sigma(k)} - q_{\sigma(k)}\right) \\ &= \sum_{k=1}^{(2r-1)^2} \left(u_{\sigma(k)} - u_{\sigma(k-1)}\right) \\ &= u_{\sigma((2r-1)^2)} - c. \end{aligned}$$

So, within the square with vertices $(\pm r, \pm r)$, the sum of the $q_{m,n}$s is close to the sum of the samples.

Next we reconstruct f using the $q_{m,n}$s, i.e.,

$$\sum_{|m|,|n|\leq r}\sum q_{m,n} s_E(x_1 - mT_1, x_2 - nT_2).$$

Let $s_k = s_E(x_1 - \sigma_1(k)T_1, x_2 - \sigma_2(k)T_2)$ and compute for any $r \geq 1$,

$$\begin{aligned}
&\sum_{|m|,|n|\leq r}\sum \left(f(mT_1, nT_2) - q_{m,n}\right) s_E(x_1 - mT_1, x_2 - nT_2) \\
&\quad = \sum_{k=1}^{(2r-1)^2} \left(f(mT_1, nT_2) - q_{m,n}\right) s_k \\
&\quad = \sum_{k=1}^{(2r-1)^2} \left(u_{\sigma(k)} - u_{\sigma(k-1)}\right) s_k \\
&\quad = \sum_{k=1}^{(2r-1)^2} u_{\sigma(k)} s_k + \sum_{k=1}^{(2r-1)^2} u_{\sigma(k-1)} s_k \\
&\quad = \sum_{k=1}^{(2r-1)^2} u_{\sigma(k)} s_k + \sum_{k=0}^{(2r-1)^2-1} u_{\sigma(k)} s_{k+1} \\
&\quad = \underbrace{\sum_{k=1}^{(2r-1)^2-1} u_{\sigma(k)} \left(s_k - s_{k+1}\right)}_{S} + \underbrace{u_{\sigma((2r-1)^2)} s_{(2r-1)^2} + u_{\sigma(0)} s_1}_{\partial S}.
\end{aligned}$$

It is not clear if S and ∂S are bounded, this is a possible direction for further research.

Chapter 4

Finite Frames and Groups

As we stated in the introduction, frames for an infinite dimensional Hilbert space are a natural language to study the sampling step of an A/D conversion. Now, we study frames for a finite dimensional Hilbert space. We then specialize to two classes of finite frames, viz., geometrically uniform (GU) frames and Grassmannian frames.

4.1 Preliminaries for finite frames

Let $X = \{x_1, \ldots, x_N\}$ be a frame for $\mathbb{R}^d$ and let L, L^*, S, and G be the Bessel map, its adjoint, the frame operator, and the Grammian, repectively, see Section 1.3. Let $\mathcal{E} = \{e_1, \ldots e_d\}$ be an orthonormal basis for $\mathbb{R}^d$, and $\mathcal{D} = \{\delta_1, \ldots, \delta_N\}$ be the canonical basis of Dirac vectors, i.e., δ_m has a one in the mth position and zero elsewhere. We now derive the matrix representation of the maps L, L^*, S, and G with repect to the bases $\mathcal{D}$ and $\mathcal{E}$. First, consider the Bessel map

$$L = L_X : \mathbb{R}^d \to \mathbb{R}^N$$

given by

$$L(y) = \left(\langle y, x_k \rangle\right)_{k=1}^{N}.$$

The adjoint, $L^* : \mathbb{R}^N \to \mathbb{R}^d$, of L is defined by $\langle Ly, v \rangle = \langle y, L^* v \rangle$ for $y \in \mathbb{R}^d$ and $v \in \mathbb{R}^N$. Since we can expand any $v \in \mathbb{R}^N$ in the canonical basis $\{\delta_n\}_{n=1}^{N}$ as $v = \sum_{n=1}^{N} \langle v, \delta_n \rangle \delta_n$, we have, for any $y \in \mathbb{R}^d$,

$$\begin{aligned}
\langle y, L^* v \rangle &= \left\langle Ly, \sum_{n=1}^{N} \langle v, \delta_n \rangle \delta_n \right\rangle \\
&= \left\langle \sum_{k=1}^{N} \langle y, x_k \rangle \delta_k, \sum_{n=1}^{N} \langle v, \delta_n \rangle \delta_n \right\rangle \\
&= \sum_{k=1}^{N} \sum_{n=1}^{N} \langle y, x_k \rangle \langle v, \delta_n \rangle \langle \delta_k, \delta_n \rangle \\
&= \sum_{k=1}^{N} \langle y, x_k \rangle \langle v, \delta_k \rangle \\
&= \left\langle y, \sum_{k=1}^{N} \langle v, \delta_k \rangle x_k \right\rangle.
\end{aligned}$$

Thus, $L^* v = \sum_{k=1}^{N} \langle v, \delta_k \rangle x_k$.

Now, since

$$L^* \delta_n = \sum_{k=1}^{N} \langle \delta_n, \delta_k \rangle x_k = x_n = \sum_{j=1}^{d} \langle x_n, e_j \rangle e_j,$$

L^* can be represented as a matrix whose columns are the coordinates of the frame vectors with respect to the orthonormal basis $\mathcal{E}$, i.e.,

$$L^* = \begin{pmatrix} \langle x_1, e_1 \rangle & \dots & \langle x_N, e_1 \rangle \\ \vdots & \ddots & \vdots \\ \langle x_1, e_d \rangle & \dots & \langle x_N, e_d \rangle \end{pmatrix},$$

a $d \times N$ matrix. Therefore, the matrix representation of the Bessel map L is just the adjoint of L^*, i.e.,

$$L = \begin{pmatrix} \overline{\langle x_1, e_1\rangle} & \dots & \overline{\langle x_1, e_d\rangle} \\ \vdots & \ddots & \vdots \\ \overline{\langle x_N, e_1\rangle} & \dots & \overline{\langle x_N, e_d\rangle} \end{pmatrix} = \begin{pmatrix} \langle e_1, x_1\rangle & \dots & \langle e_d, x_1\rangle \\ \vdots & \ddots & \vdots \\ \langle e_1, x_N\rangle & \dots & \langle e_d, x_N\rangle \end{pmatrix},$$

an $N \times d$ matrix. The frame operator for the frame X is

$$S = S_X : \mathbb{R}^d \to \mathbb{R}^d$$

given by

$$S(y) = \sum_{k=1}^{N} \langle y, x_k\rangle\, x_k,$$

and it has the matrix representation $S = L^*L$. We note that

$$S^* = (L^*L)^* = L^*L^{**} = S,$$

and therefore S is symmetric.

As an aside, we mentioned in the introduction that one problem with frame reconstruction is the computation of S^{-1}. Now if S is a diagonal matrix, this inversion is easily accomplished. We have the following fact.

Proposition 4.1. *Any finite frame in $\mathbb{R}^d$ can be orthogonally transformed into a frame with a diagonal frame operator.*

Proof. Let $X = \{x_n : n = 1, \dots, N\}$ be a frame in $\mathbb{R}^d$ with frame operator S and Bessel map L. By the spectral theorem [Str88, Lay03, GVL83] there is an orthogonal matrix P and a diagonal matrix D such that $S = P^*DP$ Thus we have

$$D = PSP^* = (PL^*)(LP^*) = (LP^*)^*(LP^*),$$

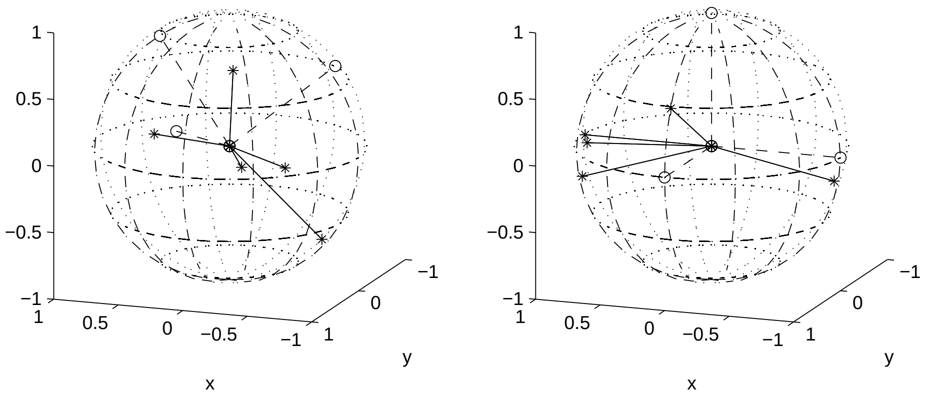


Figure 4.1: An example diagonalizing the frame operator. On the left is a unit norm frame with five elements; the eigenvectors of S are also plotted with $\circ$. On the right we apply the orthogonal matrix P to the frame X to get S_{PX} diagonal. Note the rotation/reflection of the eigenvectors.

where the kth column of $(LP^*)^*$ is Px_k. If we apply the orthogonal matrix P to the frame X we get the frame $PX = \{Px_1, \ldots, Px_N\}$, which has a frame operator $S_{PX} = D$ which is diagonal, see Figure 4.1. □

The fact that $S_{PX} = \text{diag}(\lambda_1, \ldots, \lambda_d)$ means that $S_{PX}^{-1} = \text{diag}(\lambda_1^{-1}, \ldots, \lambda_d^{-1})$, and hence the dual frame and the frame reconstruction formula can be computed efficiently.

We now prove a fact that will be use implicitly in the remainder of the thesis. Namely, any finite set of vectors forms a frame for its span with the frame bounds being the largest and smallest eigenvalues of the frame operator.

Proposition 4.2. *The following three statements are equivalent:*

$(a.)$ $\{x_n\}_{n=1}^{N}$ *is a frame for* $\mathbb{R}^d$

$(b.)$ $\exists A > 0$ *such that* $\forall y \in \mathbb{R}^d$, $A\,\|y\|^2 \leq \sum_{n=1}^{N} |\langle y, x_n\rangle|^2$

(*c*.) $\text{span}\{x_n\}_{n=1}^N = \mathbb{R}^d$.

Proof: $(a \Rightarrow b)$. This is the first inequality in the definition of a frame.

Proof: $(b \Rightarrow a)$. We use finiteness and the Cauchy-Schwarz inequality, $|\langle y, z\rangle| \leq \|y\| \cdot \|z\|$. So if we let $B = \sum_{k=1}^N \|x_k\|^2$, then

$$\sum_{k=1}^N |\langle y, x_n\rangle|^2 \leq \sum_{k=1}^N \left(\|y\|^2 \|x_k\|^2\right) = B\|y\|^2 .$$

Proof: $(b \Rightarrow c)$. Suppose $V := \text{span}\{x_n\}_{n=1}^N \neq \mathbb{R}^d$. Take $y \notin V$, and let $\{e_j\}_{j=1}^{j_0}$ be an orthonormal basis for V. Let $\text{proj}_V(y) = \sum_{j=1}^{j_0} \langle y, e_j\rangle e_j$ be the projection of y onto V, and set $\tilde{y} = y - \text{proj}_V$. Then $\tilde{y} \neq 0$ otherwise $y \in V$. By construction, $\langle \tilde{y}, x_k\rangle = 0$ for $k = 1, \ldots, N$, so

$$\forall A > 0, \quad A\|\tilde{y}\|^2 > 0 = \sum_{k=1}^N |\langle y, x_n\rangle|^2.$$

so by contraposition we have shown the forward direction of the second equivalence.

Proof: $(c \Rightarrow b)$. Consider the frame operator $S : \mathbb{R}^d \to \mathbb{R}^d$ given by $S(y) = \sum_{k=1}^N \langle y, x_k\rangle x_k$, and notice

$$\langle Sy, y\rangle = \sum_{k=1}^N \langle\langle y, x_k\rangle x_k, y\rangle = \sum_{k=1}^N |\langle y, x_n\rangle|^2.$$

Since $S = L^*L$, S is symmetric, and therefore has a full set of orthonormal eigenvectors say $\{v_k\}_{k=1}^d$. Given y, there are coefficients c_k such that $y = \sum_{k=1}^d c_k v_k$, thus

$$\langle Sy, y\rangle = y^T S y = \sum_{k=1}^d c_k^2 \lambda_k \geq (\min_{k=1,\ldots,d} \lambda_k) \sum_{k=1}^d c_k^2 = (\min_{k=1,\ldots,d} \lambda_k)\|y\|^2 .$$

So it is enough to show the eigenvalues of S are all positive. If S is positive definite, i.e., $\langle Sy, y\rangle > 0$, for $y \neq 0$, then letting y be a unit eigenvector, we see that $\lambda_y = \langle Sy, y\rangle > 0$, so it is enough to show that S is positive definite, i.e.,

$$\sum_{k=1}^{N} |\langle y, x_k\rangle|^2 = \langle Sy, y\rangle > 0, \quad \text{for } y \neq 0.$$

This is a sum of positive numbers, so it is enough to show at least one term is positive, i.e., show

$$\forall y \neq 0,\ \exists k \text{ such that } |\langle y, x_k\rangle|^2 > 0. \tag{4.1}$$

Now, to show $(c \Rightarrow b)$, we will show that if (4.1) is false we have a contradiction. Assume that span $\{x_k\}_{n=1}^{N} = \mathbb{R}^d$ and that

$$\forall k = 1, \ldots, N,\ \exists y \neq 0 \text{ such that } |\langle y, x_k\rangle|^2 = 0.$$

Let $\{e_k\}_{k=1}^{d}$ be an orthonormal basis for $\mathbb{R}^d$. Since the x_ks span, let $e_i = \sum_{k=1}^{N} c_k^{(i)} x_k$. Then

$$\forall i, \quad |\langle y, e_i\rangle| = \left|\sum_{k=1}^{N} \left\langle y, c_k^{(i)} x_k\right\rangle\right| = 0,$$

and so $y = 0$, a contradiction. □

The following alternate proof is from [Chr02].

Alternate Proof: $(c \Rightarrow b)$. Consider the map

$$\begin{aligned} \phi : \mathbb{R}^d &\to \mathbb{R} \\ y &\mapsto \sum_{n=1}^{N} |\langle y, x_n\rangle|^2 . \end{aligned}$$

It can be shown that ϕ is continuous. Then since the S^{d-1} is compact in $\mathbb{R}^d$, there is a $z \in S^{d-1}$ such that $\phi(z) = \inf\left\{\phi(y) : y \in S^{d-1}\right\}$, i.e.,

$$A := \sum_{n=1}^{N} |\langle z, x_n\rangle|^2 = \inf\left\{\sum_{n=1}^{N} |\langle y, x_n\rangle|^2 : y \in S^{d-1}\right\}.$$

Now $A > 0$ by (4.1). Thus for any $y \in \mathbb{R}^d$

$$\sum_{n=1}^{N} |\langle y, x_n\rangle|^2 = \sum_{n=1}^{N} \left|\left\langle \frac{y}{\|y\|}, x_n\right\rangle\right|^2 \|y\|^2 \geq \sum_{n=1}^{N} |\langle z, x_n\rangle|^2 \|y\|^2 = A\,\|y\|^2 .$$

□

4.2 Geometrically uniform (GU) frames

Next we consider a class of finite frames similar to Gabor and wavelet frames in that they are generated by a group acting on a vector. We shall see that these frames have subtle symmetry properties. [Sle68, For91, EB03, VW]

Definition 4.3. A set $X = \{x_1, \ldots, x_N\} \subset \mathbb{R}^d$ is a *geometrically uniform (GU)* set if there is group of orthogonal matrices $G = \{U_1, \ldots, U_N\}$ and a generating vector $x \in \mathbb{R}^d$ such that $x_n = U_n x$ for $n = 1, \ldots, N$. In the case that G is (non)abelian we call X a *GU (non)abelian* set.

As noted in the introduction, in the language of group actions, a GU set X is simply the orbit of and element $x \in X$ by the group G, i.e., $X = \mathrm{Orb}(x)$. As we shall see, a GU set is a special case of a Slepian-type group code, [Sle68, For91].

Definition 4.4. A set $X = \{x_1, \ldots, x_N\} \subset \mathbb{R}^d$ is a *Slepian-type group code* if given any two vectors $x_i, x_j \in X \subset \mathbb{R}^d$, there is an isometry $Z_{ij} : \mathbb{R}^d \to \mathbb{R}^d$ such that

$$Z_{ij}(x_i) = x_j, \quad \text{and} \quad Z_{ij}(X) = X.$$

We have,

Proposition 4.5. *If X is a GU set, then X is a Slepian-type group code.*

Proof. Let $x_i, x_j \in X \subset \mathbb{R}^d$. By assumption, $x_i = U_i x$, and $x_j = U_j x$. Thus consider the map

$$\begin{aligned} Z : \mathbb{R}^d &\to \mathbb{R}^d \\ v &\mapsto U_j U_i^T v. \end{aligned}$$

Then $U = U_j U_i^T$ is orthogonal, hence $\|Z(v)\| = \|Uv\| = \|v\|$, hence, Z is an isometry. Also, $Z(x_i) = Ux_i = U_j U_i^T x_i = U_j x = x_j$. Finally, since G is a group, $U \in G$ and for any $x_k \in X$, $U^T x_k = U^T U_k x$ and $U^T U_k \in G$, hence $U^T x_k \in X$ and $Z\left(U^T x_k\right) = UU^T x_k = x_k$.

First consider GU sets in $\mathbb{R}^2$. Now, any element of O_2 has the form

$$\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \in SO_2, \quad \text{or} \quad \begin{pmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{pmatrix} \notin SO_2. \tag{4.2}$$

Furthermore, if G is a subgroup of O_2, then G is isomorphic to $\mathbb{Z}/n\mathbb{Z}$ the cyclic group of order n, or D_{2n} the dihedral group of order $2n$ which contains $\mathbb{Z}/n\mathbb{Z}$ as a subgroup of index 2, [GB85]. Therefore we can classify all the GU sets in $\mathbb{R}^2$.

Theorem 4.6 (2d GU sets). *Let X be a GU set in $\mathbb{R}^2$, i.e., let $X = \mathrm{Orb}_G(x)$, where G is a subgroup of $O(2, \mathbb{R})$, and $x \in \mathbb{R}^2 \setminus \{0\}$. Then X is either*

(a.) the vertices of a regular n-gon,

(b.) the union of the vertices of two regular n-gons, where the angle between n-gons is twice the angle between x and the closest line of reflection.

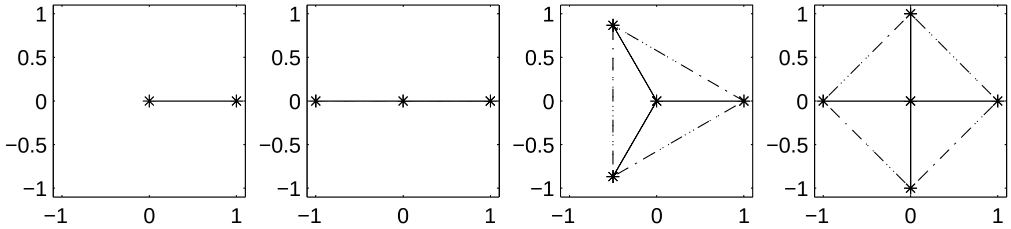


Figure 4.2: Four examples of GU frames Gx where, from left to right, G isomorphic to the dihedral groups of order 2,4,6, and 8, and $x = (1, 0)^T$ is on a line of reflection of each of these groups.

Proof. If $G \cong \mathbb{Z}/n\mathbb{Z}$, then G is precisely the cyclic group generated by

$$R_n = \begin{pmatrix} \cos\left(\frac{2\pi}{n}\right) & -\sin\left(\frac{2\pi}{n}\right) \\ \sin\left(\frac{2\pi}{n}\right) & \cos\left(\frac{2\pi}{n}\right) \end{pmatrix},$$

and $\mathrm{Orb}_G(x)$ is the set of vertices of a regular n-gon with one of the vertices being x.

If $G \cong D_{2n}$, then $G = \langle R_n, T\rangle$, where there are orthogonal vectors x_1, x_2 such that $Tx_1 = x_1$ and $Tx_2 = -x_2$, and where $\langle R_n, T\rangle$ is the group generated by the matrices R_n and T. Consider the set $\{TR_n^m : m = 0, 1, \ldots n-1\} \subset G$, which has n elements with determinant -1 corresponding to n reflections across the lines $\lambda R_n^m x_1$for $m = 0, 1, \ldots, n-1$. If x lies on a reflection line then $\mathrm{Orb}_G(x)$ is simply the vertices of a regular n-gon, see Figure 4.2. If x does not lie on a reflection line, then Gx is the union of 2 regular n-gons where the angle between the n-gons is twice the angle between x and the closest line of reflection, see Figure 4.3. □

We note for future comparison with Grassmannian frames, the set of any two vectors in $\mathbb{R}^2$ of equal length is a GU set, since this set is the union of the vertices of 2 regular 1-gons.

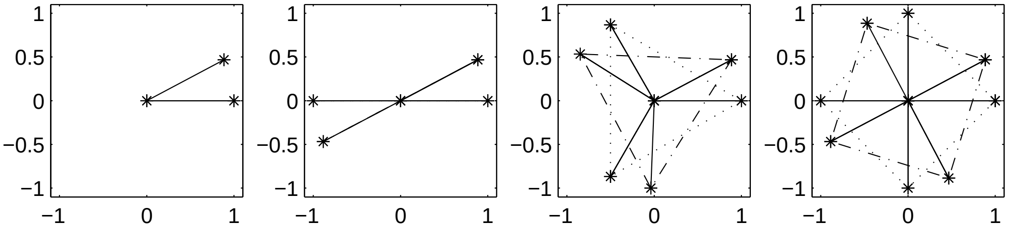


Figure 4.3: Four GU frames using G with the same isomorphism classes as in Figure 4.2 but with $x = (1, 0)^T$ **not** on any line of reflection of any of these groups. Notice there are twice as many elements in these frames as in the corresponding frames in Figure 4.2.

The GU frames in $\mathbb{R}^3$ are not as easy to classify. First, the finite subgroups of SO_3 are isomorphic to either $\mathbb{Z}/n\mathbb{Z}$, D_n, A_4, S_4, A_5, where

$\mathbb{Z}/n\mathbb{Z}$ = cyclic group of order n,

D_{2n} = dihedral group of order $2n$,

A_4 = symmetry group of a tetrahedron; order 12,

S_4 = symmetry group of cube and octahedron; order 24,

A_5 = symmetry group icosahedron and dodecahedron; order 60.

Furthermore, if two finite subgroups of SO_3 are isomorphic, then they are conjugate, [Wol84]. Thus, to classify GU frames with $G \in SO_3$, we may use a standard representation of G and conjugate by any matrix in SO_3.

The finite subgroups G of O_3 can be built from these rotation groups. If G is not equal to one of the above groups, then either $G \cong H \times \langle J \rangle$ where H is one of the above groups and J is reflection through the origin, i.e., $J = -I_3$, or G is a mixed group [GB85, Wol84], i.e.,

$$G = \{A \in O_3 : A \in H, \text{ or } A = JT \text{ where } T \in K \setminus H\},$$

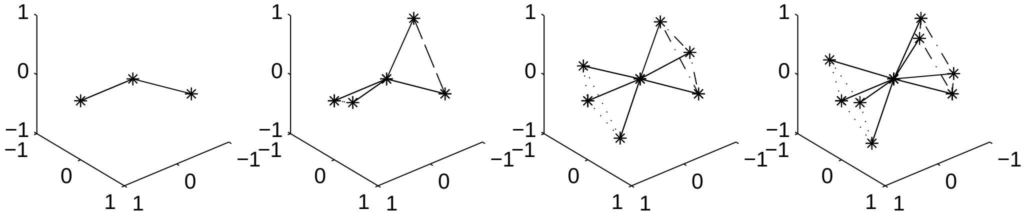


Figure 4.4: Four GU frames in $\mathbb{R}^3$ using G with the same isomorphism classes as in Figure 4.2 and with $x = (1, 0, 0)^T$.

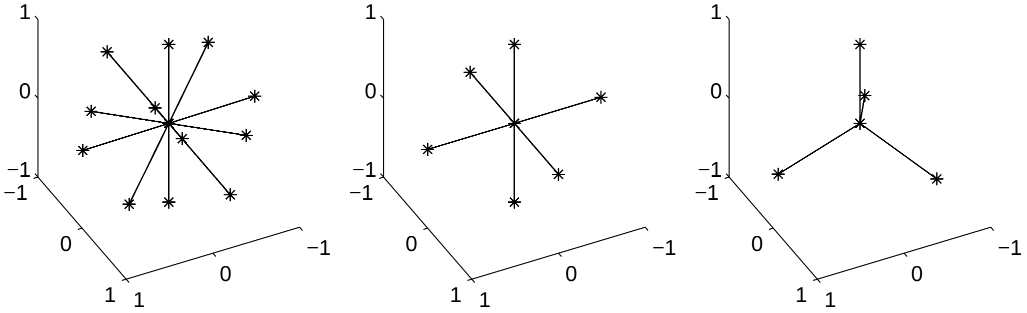


Figure 4.5: GU frames using the generating vector $x = (0, 0, 1)^T$ and the groups G isomorphic to the rotational symmetries of the icosahedron/dodecahedron, cube/octahedron, and tetrahedron, respectively.

where both H and K are finite subgroups of SO_3 with $H \leq K$ and with $[K : H] = 2$.

We conclude with some examples of GU frames where G is a finite subgroup of SO_3. If G is isomorphic to a dihedral group of order $n = 2$, 4, 6, and 8, and $x = (1, 0, 0)^T$, then the corresponding GU frames have symmetries related to a regular n-gon, see Figure 4.4. Also G is isomorphic to A_4, S_4, and A_5, then the corresponding GU frames have symmetries related to the platonic solids, see Figure 4.5 and Figure 4.6.

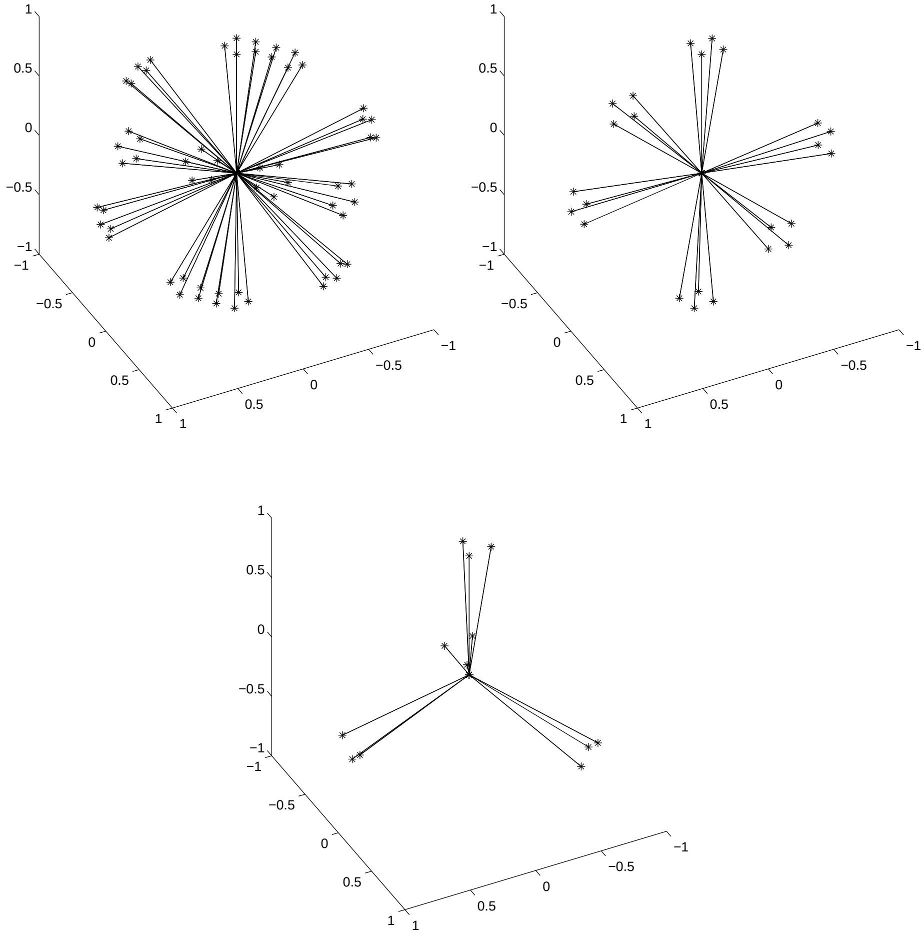


Figure 4.6: GU frames using the same isomorphism classes as in Figure 4.5 but conjugating G with the matrix $(1 + \varepsilon)I_3$, where I_3 is the 3 by 3 identity matrix, and $\varepsilon = 0.1$.

Chapter 5

Grassmannian Frames

Given a finite frame for $\mathbb{R}^d$ with N elements, we would like to measure the correlation between frame elements, and in particular decide when the correlation is small. We consider the following metric which is like an ℓ^∞ norm [SH03].

Definition 5.1. Let $N, d \in \mathbb{N}$ with $N \geq d$. Let $X_d^N = \{x_k\}_{k=1}^N$ be a subset of $\mathbb{R}^d$ with $\|x_k\| = 1$. The *maximum correlation* of X_d^N, $\mathcal{M}_\infty\left(X_d^N\right)$, is defined as

$$\mathcal{M}_\infty\left(X_d^N\right) = \max_{k \neq l} \left|\langle x_k, x_l \rangle\right|.$$

Notice that because we consider the absolute value of the inner product rather than just the inner product, if the angle between a pair of vectors is closer to 90°, then the pair is less correlated, while if the angle is closer to 0° or 180° then the pair is more correlated. We could instead consider an ℓ^1,ℓ^2, or ℓ^p-type norm to measure correlation i.e.,

$$\mathcal{M}_p\left(X_d^N\right) = \left(\sum_{k \neq l} \left|\langle x_k, x_l \rangle\right|^p\right)^{1/p}.$$

We next fix d and N with $N \geq d$ and we seek to find N element unit norm frames, X_d^N, with smallest ∞-correlation $\mathcal{M}_\infty\left(X_d^N\right)$, i.e., maximally spread apart. This relaxes condition (1.6) as discussed in Section 1.6. We make the following definition.

Definition 5.2. Let $N \geq d$. A sequence of unit norm vectors $U_d^N = \{u_k\}_{k=1}^N$ in $\mathbb{R}^d$ is called an (N, d)-*Grassmannian frame* if

$$\mathcal{M}_\infty \left(U_d^N\right) = \inf \left\{\mathcal{M}_\infty \left(X_d^N\right)\right\} \tag{5.1}$$

where the infimum is taken over all unit norm, N-element frames for $\mathbb{R}^d$.

First, we define the function

$$f : \underbrace{S^{d-1} \times \ldots \times S^{d-1}}_{N \text{ times}} \longrightarrow [0,1]$$

$$f(x_1, \ldots, x_N) = \mathcal{M}_\infty \left(\{x_k\}_{k=1}^N\right).$$

Next we check that f is continuous on $X := \mathbb{R}^d \times \ldots \times \mathbb{R}^d$ (N times). Consider a norm on X given by

$$\left\|\{x_k\}_{k=1}^N\right\|_X = \sum_{k=1}^N \|x_k\|_{\mathbb{R}^d},$$

let $\{x_k\}_{k=1}^N \in X$ be fixed, set $R - 1 = \max_k \{\|x_k\|_{\mathbb{R}^d}\}$ an let $\varepsilon > 0$ be given. So $R \geq 1$. Now choose δ such that $0 < \delta < \frac{\sqrt{1+\varepsilon}-1}{R}$, i.e., $R^2\delta^2 + 2R\delta < \varepsilon$. Then whenever $\left\|\{y_k\}_{k=1}^N - \{x_k\}_{k=1}^N\right\|_X < \delta$, we have that for every $j \in \{1, \ldots, N\}$,

$$\|y_j - x_j\|_{\mathbb{R}^d} \leq \sum_{k=1}^N \|y_k - x_k\|_{\mathbb{R}^d} = \left\|\{y_k\}_{k=1}^N - \{x_k\}_{k=1}^N\right\|_X < \delta$$

and therefore for each j, there is and $\alpha_j \in \mathbb{R}^d$ with $\|\alpha_j\| < \delta$ such that $y_j = x_j + \alpha_j$.

So,

$$
\begin{aligned}
&|f(y_1, \ldots, y_N) - f(x_1, \ldots, x_N)| \\
&\quad = \left|\mathcal{M}_\infty\left(\{y_k\}_{k=1}^N\right) - \mathcal{M}_\infty\left(\{x_k\}_{k=1}^N\right)\right| \\
&\quad = \left|\max_{k\neq l}\{|\langle x_k + \alpha_k, x_l + \alpha_l\rangle|\} - \max_{k\neq l}\{|\langle x_k, x_l\rangle|\}\right| \\
&\quad = \left|\max_{k\neq l}\{|\langle x_k, x_l\rangle| + |\langle x_k, \alpha_l\rangle| + |\langle \alpha_k, x_l\rangle| + |\langle \alpha_k, \alpha_l\rangle|\} - \max_{k\neq l}\{|\langle x_k, x_l\rangle|\}\right| \\
&\quad \leq \left|\max_{k\neq l}\{|\langle x_k, x_l\rangle| + \|x_k\|\,\|\alpha_l\| + \|\alpha_k\|\,\|x_l\| + \|\alpha_k\|\,\|\alpha_l\|\} - \max_{k\neq l}\{|\langle x_k, x_l\rangle|\}\right| \\
&\quad < \left|\max_{k\neq l}\{|\langle x_k, x_l\rangle|\} + 2R\delta + R\delta^2 - \max_{k\neq l}\{|\langle x_k, x_l\rangle|\}\right| \\
&\quad = 2R\delta + R\delta^2 < \varepsilon.
\end{aligned}
$$

Therefore f is continuous on the compact set $S^{d-1} \times \ldots \times S^{d-1}$ (N times), thus f achieves its absolute maximum and absolute minimum on this set. Thus we know that (N, d)-Grassmannian frames exist for any $N \geq d$. Next we must check that if U_d^N solves (5.1), then U_d^N is a unit norm frame for $\mathbb{R}^d$, but this a tautology since, by compactness, U_d^N is one of the frames over which we are taking the infimum.

5.1 Two dimensional Grassmannian frames

We now classify all $(N, 2)$-Grassmannian frames for any $N \geq 2$.

Theorem 5.3 (2 dimensional Grassmanian). *Let $X = X_2^N = \{x_k\}_{k=1}^N$ be a collection of N unit vectors in $\mathbb{R}^2$. Then we have the lower bound*

$$\cos(\pi/N) \leq \mathcal{M}_\infty\left(X_2^N\right).$$

Furthermore, X_2^N *is an* $(N,2)$*-Grassmannian frame if and only if there is an orthogonal matrix* P*, see* (4.2)*, and a sequence* $\{\varepsilon_k\}_{k=1}^N \subset \{\pm 1\}^N$ *such that*

$$P\varepsilon X_2^N := \left\{P(\varepsilon_k x_k) : x_k \in X_2^N\right\} = \left\{ \begin{pmatrix} \cos(\pi k/N) \\ \sin(\pi k/N) \end{pmatrix} : k = 1, \ldots, N \right\}.$$

Proof. First, since $|\langle x, y\rangle| = |\langle x, -y\rangle|$, we note that changing the sign of any $x_k \in X$ does not effect the value of $\mathcal{M}_\infty(X)$. So by changing the sign on x_k when necessary, we may assume $x_k \in \{v \in S^1 : \langle v, \delta_2\rangle \geq 0\}$. Also, since rotations preserve inner products, applying a rotation to all the vectors in X does not effect $\mathcal{M}_\infty(X)$. Thus rotating by $-\phi$ where $\phi = \min_{k=1,\ldots,N} \cos^{-1}(\langle x_k, \delta_1\rangle)$, and reordering if necessary, we may assume $x_1 = \delta_1 = (1,0)^T$, and

$$1 \geq \langle x_2, x_1\rangle \geq \langle x_3, x_1\rangle \geq \ldots \geq \langle x_N, x_1\rangle \geq -1. \tag{5.2}$$

For $k = 1, \ldots, N-1$, let θ_k be the angle between x_k and x_{k+1}, and let θ_N be the angle between x_N and the negative x-axis, i.e., $\theta_k = \cos^{-1}\left(\langle x_{k+1}, x_k\rangle\right)$ and $\theta_N = \cos^{-1}\left(\langle -\delta_1, x_N\rangle\right)$, see Figure 5.1 for an example when $N = 6$. Then because of the above reordering, $\theta_k \geq 0$ for $k = 1, \ldots, N$, and $\sum_{k=1}^N \theta_k = \pi$. Thus for $1 \leq l < k \leq N$,

$$|\langle x_k, x_l\rangle| = \left|\cos\left(\sum_{j=l}^{k-1} \theta_j\right)\right|, \quad \text{where} \min_{k=1,\ldots,N-1} \theta_k \leq \sum_{j=l}^{k-1} \theta_j \leq \pi - \theta_N.$$

Furthermore, $|\cos(\theta)|$ has a maximum on $[0,\pi]$, at $\theta = 0$, and $\theta = \pi$, and $|\cos(\theta)|$

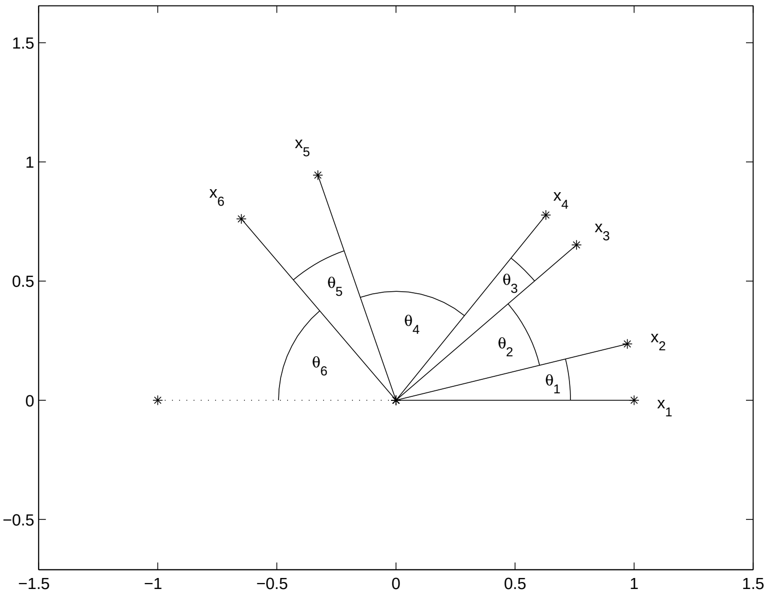


Figure 5.1: An example of the reordering induced by the inequalities on the inner products in (5.2). Note $N = 6$.

is monotone decreasing on $[0, \pi/2]$ and monotone increasing on $[\pi/2, \pi]$. Hence

$$\begin{aligned}\mathcal{M}_\infty(X) &= \max_{k\neq l} |\langle x_k, x_l\rangle| \\ &= \max_{k\neq l} \left|\cos\left(\sum_{j=l}^{k-1} \theta_j\right)\right| \\ &= \max\left\{|\cos(\pi - \theta_N)|, \left|\cos\left(\min_{k=1,\ldots,N-1} \theta_k\right)\right|\right\} \\ &= \left|\cos\left(\min_{k=1,\ldots,N} \theta_k\right)\right|.\end{aligned}$$

Thus, in order to minimize $\mathcal{M}_\infty(X)$ we must choose N positive numbers $\alpha_1, \ldots, \alpha_N$ which sum to π and which minimize $|\cos(\min_{k=1,\ldots,N} \alpha_k)|$, hence, which maximize the expression

$$\min_{k=1,\ldots,N} \alpha_k. \tag{5.3}$$

Now we claim that if $\alpha_1, \ldots, \alpha_N$ maximize (5.3) then $\alpha_1 = \ldots = \alpha_N$. We prove this impliction by contraposition, i.e., assume it is not the case that $\alpha_1 =$

$\ldots = \alpha_N$. Then there is an $m \in \{1, 2, \ldots, N-1\}$ so that if we list $\alpha_1 \leq \ldots \leq \alpha_N$ by size, then only the first m are equal, and the $(m+1)$st is strictly larger than the mth, i.e.,

$$\alpha_{k_1} = \alpha_{k_2} = \ldots = \alpha_{k_m} < \alpha_{k_{m+1}} \leq \ldots \leq \alpha_{k_N}.$$

Let $\varepsilon = \alpha_{k_{m+1}} - \alpha_{k_m}$ and for $j = 1, \ldots, N$, define the sequence β_{k_j} as

$$\beta_{k_j} = \begin{cases} \alpha_{k_j} + \frac{\varepsilon}{2m} & \text{for } j = 1, \ldots, m, \\ \alpha_{k_j} - \frac{\varepsilon}{2} & \text{for } j = m+1, \\ \alpha_{k_j} & \text{for } j = m+2, \ldots, N, \end{cases}$$

Now the new set

$$\beta_{k_1} = \beta_{k_2} = \ldots = \beta_{k_m} \leq \beta_{k_{m+1}} \leq \ldots \leq \beta_{k_N},$$

has a strictly larger minimum angle than the original since for $j = 1, \ldots N$,

$$\min_{k=1,\ldots,N} \alpha_k = \alpha_{k_1} < \alpha_{k_1} + \frac{\varepsilon}{m} \leq \beta_{k_1} \leq \beta_j.$$

We see that the original αs do not maximize (5.3). So by contraposition we have that if αs maximize (5.3), then they must all equal. Finally, if α is the common value, then $\sum_{k=1}^N \alpha_k = N\alpha = \pi$, and therefore $\alpha = \pi/N$. Thus $\pi/N \geq \min_{k=1,\ldots,N} \theta_K$, so

$$\cos(\pi/N) \leq \cos\left(\min_{k=1,\ldots,N} \theta_K\right) = \mathcal{M}_\infty\left(X_2^N\right).$$

Next we prove the equivalent characterization of $(N, 2)$-Grassmannian frames. If X_2^N is an $(N, 2)$-Grassmannian frame, then using the above argument, we see that we can choose $\{\varepsilon_k\} \subset \{\pm 1\}^N$ and $P \in SO_2$ so that the frame $P\varepsilon X_2^N =$

$\left\{P(\varepsilon_k x_k) : x_k \in X_2^N\right\}$ is in the closed upper halfplane with one of the vectors being $(1,0)^T$, and

$$\mathcal{M}_\infty\left(X_2^N\right) = \mathcal{M}_\infty\left(P\varepsilon X_2^N\right) = \cos\left(\min_{k=1,\ldots,N} \theta_k\right).$$

where θ_k is the angle between the kth and $(k+1)$st adjacent vectors in $P\varepsilon X_2^N$ (reindexing may be necessary). Since an $(N,2)$-Grassmannian frame minimizes the ∞-correlation $\mathcal{M}_\infty\left(X_2^N\right)$, the above argument also shows that $\theta_1 = \ldots = \theta_N = \pi/N$. Therefore, the angle between adjacent vectors in $P\varepsilon X_2^N$ is π/N, and we have shown the forward direction of the equivalence.

To show the reverse implication we note if

$$P\varepsilon X_2^N = \left\{\begin{pmatrix}\cos(\pi k/N)\\ \sin(\pi k/N)\end{pmatrix} : k = 1,\ldots,N\right\}$$

then

$$\mathcal{M}_\infty\left(X_2^N\right) = \mathcal{M}_\infty\left(P\varepsilon X_2^N\right) = \cos\left(\min_{k=1,\ldots,N} \theta_k\right) = \cos(\pi/N)$$

So X_2^N is $(N,2)$-Grassmannian since it achieves the lower bound. □

Notice that for N odd, if we change the sign on the the Nth roots of unity below the real axis, then we obtain the frame described in the above claim with $\varepsilon_k = 1$, i.e., with all vectors in the upper half plane, and a common angle of π/N between adjacent vectors. Hence for N odd, the Nth roots of unity are $(N,2)$-Grassmannian. Furthermore, for N even, the Nth roots of unity do not form an $(N,2)$-Grassmannian frame because ζ and $-\zeta$ are both Nth roots. If we identify ζ and $-\zeta$ then we obtain an $(N/2,2)$-Grassmannian frame.

5.2 A lower bound for $\mathcal{M}_\infty$

It is much harder to construct a Grassmannian frame in $\mathbb{R}^3$ for $N > 3$. Thus we first derive a lower bound for the maximum correlation between frame elements of an N-element frame for $\mathbb{R}^d$, [SH03, Ros97].

Theorem 5.4. *Let $N \geq d$ and let X_d^N be an N-element subset S^{d-1}, and let $d_0 = \dim\left(\operatorname{span}\left(X_d^N\right)\right)$. Then*

$$\mathcal{M}_\infty\left(X_d^N\right) \geq \sqrt{\frac{N-d_0}{d_0(N-1)}}, \tag{5.4}$$

where equality holds in (5.4) *if and only if*

1.) X_d^N is equiangular, and

2.) X_d^N is a tight frame for its span with frame bounds $A = B = \frac{N}{d_0}$.

Furthermore, if $N > \frac{d(d+1)}{2}$, then X_d^N is not equiangular, hence equality cannot hold in (5.4).

Proof. First we show the inequality (5.4). Since the $N \times N$ Grammian matrix G is hermitian, the spectral theorem applies, so G has N eigenvalues counted with multiplicity and ordered by size, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_N$. Furthermore, since $\operatorname{rank}(G) = d_0$, only the first d_0 of these eigenvalues are nonzero. So

$$\sum_{k=1}^{d_0} \lambda_k = \operatorname{Trace} G = \sum_{k=1}^{N} |\langle x_k, x_k\rangle| = \sum_{k=1}^{N} 1 = N$$

Now set $e_k = \lambda_k - \frac{N}{d_0}$, then

$$\sum_{k=1}^{d_0} e_k = \sum_{k=1}^{d_0}\left(\lambda_k - \frac{N}{d_0}\right) = N - d_0\frac{N}{d_0} = 0,$$

so

$$\begin{aligned}\sum_{k=1}^{d_0} \lambda_k^2 &= \sum_{k=1}^{d_0} \left(\frac{N}{d_0} + e_k \right)^2 \\ &= \sum_{k=1}^{d_0} \frac{N^2}{d_0^2} + \frac{2N}{d_0} \sum_{k=1}^{d_0} e_k + \sum_{k=1}^{d_0} e_k^2 \\ &= \frac{N^2}{d_0} + \sum_{k=1}^{d_0} e_k^2 \\ &\geq \frac{N^2}{d_0},\end{aligned}$$

with equality if and only if $e_k = 0$ for $k = 1, \ldots d_0$, i.e., $\lambda_k = \frac{N}{d_0}$ for $k = 1, \ldots d_0$. Now the eigenvalues of G^2 are $\lambda_1^2 \geq \lambda_2^2 \geq \ldots \geq \lambda_N^2$, so if g_k is the kth column of G, then by matrix multiplication we have

$$\frac{N^2}{d_0} \leq \sum_{k=1}^{d_0} \lambda_k^2 = \text{Trace}(G^2) = \sum_{k=1}^{N} g_k^* g_k = \sum_{k=1}^{N} \sum_{l=1}^{N} |\langle x_k, x_l \rangle|^2. \tag{5.5}$$

And since G is hermitian, $|\langle x_k, x_l \rangle| = |\langle x_l, x_k \rangle|$, so using the previous inequality we derive

$$\begin{aligned}\frac{N^2}{d_0} &\leq \sum_{k=1}^{N} \sum_{l=1}^{N} |\langle x_k, x_l \rangle|^2 \\ &= \sum_{k=l} |\langle x_k, x_l \rangle|^2 + \sum_{k<l} |\langle x_k, x_l \rangle|^2 + \sum_{k>l} |\langle x_k, x_l \rangle|^2 \\ &= N + 2 \sum_{k<l} |\langle x_k, x_l \rangle|^2 \\ &\leq N + 2 \frac{N(N-1)}{2} \max_{k \neq l} \{ |\langle x_k, x_l \rangle|^2 \}\end{aligned} \tag{5.6}$$

therefore, solving for the max in the above inequality, we have

$$\frac{N - d_0}{d_0(N-1)} \leq \mathcal{M}_\infty \left(X_d^N \right)^2. \tag{5.7}$$

For future reference we note that $d \geq d_0$ implies $\frac{N-d}{d(N-1)} \leq \frac{N-d_0}{d_0(N-1)}$ hence we have (5.4) with the d_0s replaces with ds.

Next we show that equality holds in (5.4) if and only if, X_d^N is equiangular and a tight frame for its span.

Proof $\Longrightarrow$. Say $\mathcal{M}_\infty(X_d^N) = \sqrt{\frac{N-d_0}{d_0(N-1)}}$. The (5.6) becomes

$$\sum_{k=1}^{N}\sum_{l=1}^{N} |\langle x_k, x_l\rangle|^2 = \frac{N^2}{d_0},$$

which implies that (5.5) becomes

$$\sum_{k=1}^{d_0} \lambda_k^2 = \frac{N^2}{d_0},$$

and as we saw above, equality in this sum implies that $\lambda_k = \frac{N}{d_0}$ for $k = 1, \dots, d_0$. But the frame bounds for X_d^N are the largest and smallest nonzero eigenvalues hence $A = N/d_0 = B$ and X_d^N is a tight frame for its span.

To see that X_d^N is also equiangular, we notice that (5.6) also implies that

$$N = 2\sum_{k<l} |\langle x_k, x_l\rangle|^2 = \frac{N^2}{d_0},$$

hence,

$$\sum_{k<l} |\langle x_k, x_l\rangle|^2 = \frac{N(N-d_0)}{2d_0}. \tag{5.8}$$

Now $\max_{k\neq l} |\langle x_k, x_l\rangle|^2 = \frac{N-d_0}{d_0(N-1)}$ implies that for any $k \neq l$,

$$|\langle x_k, x_l\rangle|^2 = \frac{N-d_0}{d_0(N-1)} - \varepsilon_{k,l},$$

where $\varepsilon_{k,l} \geq 0$. Thus (5.8) implies

$$\begin{aligned}\frac{N(N-d_0)}{2d_0} &= \sum_{k<l}\left(\frac{N-d_0}{d_0(N-1)} - \varepsilon_{k,l}\right)\\ &= \left(\frac{N(N-1)}{2}\right)\left(\frac{N-d_0}{d_0(N-1)}\right) - \sum_{k<l}\varepsilon_{k,l}\\ &= \frac{N(N-d_0)}{2d_0} - \sum_{k<l}\varepsilon_{k,l}\end{aligned}$$

hence $\sum_{k<l} \varepsilon_{k,l} = 0$, and since $\varepsilon_{k.l}$ are positive, $\varepsilon_{k,l} = 0$ for $k < l$. Also since G is symmetric, $\varepsilon_{k,l} = 0$ for all $k \neq l$, hence X_d^N is equiangular with $|\langle x_k, x_l \rangle|^2 = \frac{N-d_0}{d_0(N-1)}$.

Proof $\Longleftarrow$. Now assume X_d^N is equiangular and tight with $A = B = \frac{N}{d_0}$. Then there is an $\alpha \in [0,1]$, such that $|\langle x_k, x_l \rangle|^2 = \alpha$ for $k \neq l$. Now since X_d^N is tight, $\lambda_k = \frac{N}{d_0}$ for $k = 1, \ldots, d_0$, and zero otherwise. Hence (5.5) and (5.6) imply

$$\frac{N}{d_0} = \sum_{k=1}^{d_0} \lambda_k^2 = \sum_{k=1}^{N} \sum_{l=1}^{N} |\langle x_k, x_l \rangle|^2 = N + N(N-1)\alpha,$$

hence, solving for α we see that equality holds in (5.4).

Finally to prove $N > \frac{d(d+1)}{2}$ implies X_d^N is not equiangular, we need the following lemma

Lemma 5.5. *Let H_n be the $n \times n$ matrix with 1 on the main diagonal and β elsewhere, and let C_n be the $n \times n$ matrix defined by*

$$[C_n]_{i,j} = \begin{cases} \beta, & \text{if } (i,j) = (1,1) \\ [H_n]_{i,j}, & \text{otherwise.} \end{cases}$$

Then

$$\det(H_n) = (1 + (n-1)\beta)(1-\beta)^{n-1} \tag{5.9}$$

$$\det(C_n) = \beta(1-\beta)^{n-1}. \tag{5.10}$$

Proof of Lemma 5.5. We proceed by induction. Let $P(n)$ be the statement

$$\det(H_n) = (1 + (n-1)\beta)(1-\beta)^{n-1} \text{ and } \det(C_n) = \beta(1-\beta)^{n-1}.$$

Now for $n = 1$, $H_1 = 1$ and $C_1 = \beta$, so $\det(H_1) = 1$ and $\det(C_1) = \beta$, hence $P(1)$.

Next assume $P(n)$. Now using the cofactor expansion of the determinant, we first note that the $(1,1)$-cofactor of H_{n+1} and C_{n+1} is $\det(H_n)$. Also note that for $j = 2, \ldots, n+1$, the $(1,j)$-cofactor of both H_{n+1} and C_{n+1} is

$$(-1)^{1+j} \det\left(B_n^{(j)}\right)$$

where $B_n^{(j)}$ can be defined recursively as

$$\begin{aligned} B_n^{(1)} &= C_n \\ B_n^{(j)} &= \tilde{B}_n^{(j-1)} \quad \text{for } j = 2, \ldots, n+1. \end{aligned}$$

where $\tilde{B}_n^{(j-1)}$ is $B_n^{(j-1)}$ with the jth and $(j-1)$st rows interchanged. Now since det is multilinear, interchanging a row changes the sign of the determinant, hence

$$(-1)^{1+j} \det(B_n^{(j)}) = -\det(C_n) \quad \text{for } j = 2, \ldots n+1.$$

Therefore we compute using the induction hypothesis and the cofactor expansion,

$$\begin{aligned} \det(H_{n+1}) &= 1 \cdot \det(H_n) + \sum_{j=2}^{n+1} \left(\beta \cdot (-1)^{1+j} \det(B_n^{(j)})\right) \\ &= \det(H_n) - n\beta \det(C_n) \\ &= (1 + (n-1)\beta)(1-\beta)^{n-1} - n\beta^2(1-\beta)^{n-1} \\ &= (1 + n\beta)(1-\beta)(1-\beta)^{n-1} \end{aligned}$$

and,

$$\begin{aligned} \det(C_{n+1}) &= \beta \det(H_n) - n\beta \det(C_n) \\ &= \beta(1 + (n-1)\beta)(1-\beta)^{n-1} - n\beta^2(1-\beta)^{n-1} \\ &= (\beta - \beta^2)(1-\beta)^{n-1}, \end{aligned}$$

hence by induction, $P(n)$ for all $n \in \mathbb{N}$. □

Thus to prove $N > \frac{d(d+1)}{2}$ implies X_d^N is not equiangular, we prove the contrapositive using the above lemma and the following argument, [LS73]. Assume X_d^N is equiangular. Let $P_k : \mathbb{R}^d \to \mathbb{R}^d$ be the projection of x onto the line spanned by x_k i.e., $P_k x = \langle x_k, x \rangle\, x_k$. Let V be the vector space of symmetric linear mappings from $\mathbb{R}^d \to \mathbb{R}^d$. Then $\dim(V) = \frac{d(d+1)}{2}$, and the map $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$ given by $\langle A, B \rangle = \text{Trace}(AB)$ is an inner product on V. Now, since X_d^N is equiangular, there is an $\alpha \in [0,1]$ such that $\langle x_k, x_l \rangle = \pm\alpha$ for $k \neq l$. Furthermore, $\alpha = 1$ implies $N = 2$, since the elements of X_d^N are assumed to be distinct and of unit norm. Thus, for $d \geq 2$, $N = 2 < 3 \leq \frac{d(d+1)}{2}$. So we may assume $\alpha \in [0,1)$. Now,

$$\langle P_k, P_l \rangle = \langle x_k, x_l \rangle^2 = \begin{cases} 1, & \text{if } k = l \\ \alpha^2, & \text{if } k \neq l. \end{cases}$$

so the Grammian of the set $\{P_1, \ldots P_N\} \subset V$ is

$$[G]_{k,l} = [\langle P_k, P_l \rangle]_{k,l} = \begin{cases} 1, & \text{if } k = l \\ \alpha^2, & \text{if } k \neq l. \end{cases}$$

Thus Lemma 5.5 applies with $G = H_N$ and $\beta = \alpha^2$, thus, if $\alpha \in [0,1)$, then

$$\det G = \left(1 + (N-1)\alpha^2\right)(1 - \alpha^2)^{N-1} \neq 0.$$

Hence G is invertible and therfore has full rank. Finally, since $\text{rank}(G) = \text{rank}(S) = N$, so,

$$N = \text{rank}(G) = \dim\left(\text{span}\left\{P_1, \ldots, P_N\right\}\right) \leq \dim(V) = \frac{d(d+1)}{2}.$$

So, X_d^N equiangular implies $N \leq d(d+1)/2$, hence by contraposition we have proven the result. □

Remark Theorem 5.3 shows that $\mathcal{M}_\infty\left(X_2^N\right) = \cos\left(\pi/N\right)$, while Theorem 5.4 shows $\mathcal{M}_\infty\left(X_2^N\right) \geq \sqrt{\frac{N-2}{2(N-1)}}$. Using standard calculus techinques, we can show that the bound in Theorem 5.3 is an improvement over the bound in Theorem 5.4 for all $N > 3$. Let

$$f(x) = \cos^2(\pi/x) - \frac{x-2}{2(x-1)},$$

then

$$f'(x) = \frac{\pi}{x^2}\sin\left(\frac{2\pi}{x}\right) - \frac{1}{2(x-1)^2}$$

and

$$f''(x) = -\frac{2\pi}{x^3}\left(\frac{\pi}{x}\cos\left(\frac{2\pi}{x}\right) + \sin\left(\frac{2\pi}{x}\right)\right) + \frac{1}{(x-1)^3}.$$

So, $f'(x) > 0 \iff \sin(2\pi/x) > \frac{1}{2\pi}\left(\frac{x}{x-1}\right)^2$. But for $x \in [3,6]$,

$$\sin\left(\frac{2\pi}{x}\right) \geq \frac{\sqrt{3}}{2} \geq \frac{9}{8\pi} \geq \frac{1}{2\pi}\left(\frac{x}{x-1}\right)^2,$$

so $f(x)$ is increasing for $x \in [3,6]$, and since $f(3) = 0$, we have that $f(x) \geq 0$ for $x \in [3,6]$. Furthermore, for $x \in [6,\infty)$, $\frac{36}{50\pi} \geq \frac{1}{2\pi}\left(\frac{x}{x-1}\right)^2$, and $\sin\left(\frac{2\pi}{x}\right) \geq \frac{36}{50\pi}$ if and only if

$$x \leq \frac{2\pi}{\sin^{-1}\left(\frac{36}{50\pi}\right)} \approx 27.1719.$$

So $f(x)$ is increasing for $x \in [3,27]$, hence greater that zero on that same interval We also note that

$$\sin\left(2\pi x\right) \geq \frac{1}{2\pi} < \frac{1}{2\pi}\left(\frac{x}{x-1}\right)^2$$

and $\sin\left(2\pi x\right) \geq \frac{1}{2\pi}$ when

$$x \geq \frac{2\pi}{\sin^{-1}\left(\frac{1}{2\pi}\right)} \approx 39.3105.$$

Hence f is decreasing on the interval $[40,\infty)$, and since $\lim_{x\to\infty} f(x) = \frac{1}{2}$, we have that $f(x) > \frac{1}{2}$ for $x \in [40,\infty)$. Finally, we check that $f'' < 0$ on the interval

N	Optimal bound	Bound from Theorem 5.3
	$= \sqrt{\frac{N-2}{2(N-1)}}$	$\cos(\pi/N)$
3	0.5000	0.5000
4	0.5774	0.7071
5	0.6124	0.8090
6	0.6325	0.8660
7	0.6455	0.9010
8	0.6547	0.9239
9	0.6614	0.9397
10	0.6667	0.9511

Table 5.1: Improvment of the optimal bound derived in Theorem 5.4 for the case of $(N, 2)$-Grassmannian frames.

$[27, 40]$ and $f(27), f(40) > \frac{1}{2}$, so $f(x) > \frac{1}{2}$ on $[27, 40]$. In summary we have shown that $f(x) > 0$ on $(3, \infty)$ and that $f(x) > \frac{1}{2}$ on $[27, \infty)$. Therefore

$$\cos\left(\frac{\pi}{N}\right) > \sqrt{\frac{N-2}{2(N-1)}} \quad \text{for } N > 3.$$

In light of Theorem 5.4, we make the following definition,

Definition 5.6. Let $N, d \in \mathbb{N}$ with $d \leq N \leq \frac{d(d+1)}{2}$. Let $X_d^N = \{x_k\}_{k=1}^N$ be a frame for $\mathbb{R}^d$ with $\|x_k\| = 1$. We call X_d^N an *optimal Grassmannian* frame if X_d^N satisfies (5.4) with equality, i.e

$$\mathcal{M}_\infty\left(X_d^N\right) = \sqrt{\frac{N-d}{d(N-1)}}.$$

N	Optimal bound	Grassmannian bound
	$=\sqrt{\frac{N-3}{3(N-1)}}$	$= \min \mathcal{M}_\infty\left(X_3^N\right)$
3	0	0
4	$0.333\overline{3}$	$0.333\overline{3}$
5	0.4082	0.4472
6	0.4472	0.4472

Table 5.2: Bounds for N-element frames in $\mathbb{R}^3$ with potential of being optimal Grassmannian.

In $\mathbb{R}^2$, since $d = 2$ and $\frac{d(d+1)}{2} = 3$, only frames with $N = 2$ and $N = 3$ elements can be optimal Grassmannian. Since $\cos(\pi/2) = 0 = \sqrt{(2-2)/(2(2-1)}$, and $\cos(\pi/3) = 1/2 = \sqrt{(3-2)/(2(3-1)}$, both $(2,2)$- and $(3,2)$-Grassmannian frames are optimal. The same phenomenon does not happen in three dimensions. Table 5.2 lists the Grassmannian bound which will be proven below and the optimal bound for $N = 3, 4, 5, 6$, (the only Ns with the possibility of being optimal). By inspecting Table 5.2, we notice that $(5,3)$-Grassmannian frames are not optimal, while $(3,3)$, $(4,3)$ and $(6,3)$-Grassmanians are optimal.

5.3 $(4,3)$-Grassmannian frames

In this section and the next we will derive the bounds for three dimensional Grassmannian frames with $N = 3, 4, 5$ and 6. First note that if $N = 3$, and if X is any orthonormal basis for $\mathbb{R}^3$, then $0 \leq \mathcal{M}_\infty(X) = 0$. Hence any orthonormal basis is Grassmannian, in fact, X is trivially optimal Grassmannian.

Next consider $N = 4$. We need the following lemma,

Lemma 5.7. *Let $a \in \mathbb{R}^d$, and let $\{v_1, v_2, \ldots, v_d\} \subset \mathbb{R}^d$. Set*

$$Q = \left\{ a + \sum_{j=1}^{d} s_j v_j : s_j \in [0,1] \right\}$$
$$C = \left\{ a + \sum_{j=1}^{d} \varepsilon_j v_j : \varepsilon_j \in \{0,1\} \right\},$$

and choose $c \in C$ such that $\|c\| = \max\{\|c_l\| : c_l \in C\}$ where $l = 1, \ldots, 2^d$. Then for any $v \in Q \setminus C$, $\|v\| < \|c\|$.

Proof. Let $v \in Q \setminus C$, so $v = a + \sum_{j=1}^{d} s_j v_j$. Since $v \notin C$, there is an $m \geq 1$, such that $s_{j_1}, \ldots, s_{j_m} \in (0,1)$ and $s_{j_{m+1}}, \ldots, s_{j_d} \in \{0,1\}$. For $i \leq m$ set $t_i = s_{j_i}$, and for $i > m$ set $\varepsilon_i = s_{j_i}$, so $t_i \in (0,1)$ and $\varepsilon_i \in \{0,1\}$. Now, let $w_0 = v$, and for each $i = 1, \ldots, m$, recursively let $w_i = w_{i-1} + (\tilde{\varepsilon}_i - t_i)v_{j_i}$, where

$$\tilde{\varepsilon}_i = \begin{cases} 1, & \text{if } \langle v_{j_i}, w_{i-1} \rangle > 0 \\ 0, & \text{if } \langle v_{j_i}, w_{i-1} \rangle \leq 0. \end{cases}$$

By inducting on i, we show that $\|v\| = \|w_0\| < \|w_1\| < \ldots < \|w_m\|$. Note that by construction of $\tilde{\varepsilon}_i$, we have that $w_m \in C$ hence $\|w_m\| \leq \|c\|$. Also note for $i = 1, \ldots, m$,

$$\|w_i\|^2 = \|w_{i-1}\|^2 + 2(\tilde{\varepsilon}_i - t_i)\langle w_{i-1}, v_{j_i} \rangle + (\tilde{\varepsilon}_i - t_i)^2 \|v_{j_i}\|^2 \tag{5.11}$$

We begin the induction with the base case $i = 1$. Inspecting (5.11) with $i = 1$, we have 3 cases:

Case 1. $\langle w_0, v_{j_1} \rangle > 0$.
Then $\tilde{\varepsilon}_1 = 1$, so $\tilde{\varepsilon}_1 - t_1 > 1 - t_1 > 0$, and $2(\tilde{\varepsilon}_1 - t_1)\langle w_0, v_{j_1} \rangle > 0$. Hence by (5.11), $\|w_1\|^2 > \|w_0\|^2$.

Case 2. $\langle w_0, v_{j_1}\rangle < 0$.

Then $\tilde{\varepsilon}_1 = 0$, so $\tilde{\varepsilon}_1 - t_1 < -t_1 < 0$, and therfore $2\left(\tilde{\varepsilon}_1 - t_1\right)\langle w_0, v_{j_1}\rangle > 0$. Hence by (5.11), $\|w_1\|^2 > \|w_0\|^2$.

Case 3. $\langle w_0, v_{j_1}\rangle = 0$.

Then $\tilde{\varepsilon}_1 = 0$, and $w_0 + (-t_1 v_{j_1}) = y_2$, where for $k = 2, \dots m$,

$$y_k = a + \sum_{i=k}^{d} s_{j_i} v_{j_i}.$$

So by the Pythagorean theorem, $\|w_0\|^2 + \|-t_1 v_{j_1}\|^2 = \|y_2\|^2$, hence

$$\|w_0\|^2 = \|y_2\|^2 - t_1 \|v_{j_1}\|^2 . \tag{5.12}$$

So, $\tilde{\varepsilon}_1 = 0$ and $-t_1^2 < 0$ imply $-t_1^2 \|v_{j_1}\| < -\tilde{\varepsilon}_1^2 \|v_{j_1}\|$ and by equation (5.12),

$$\|w_0\|^2 < \|y_2\|^2 + 0 = \|w_0 + (\tilde{\varepsilon}_1 - t_1)v_{j_1}\|^2 = \|w_1\|^2$$

So in every case, $\|w_0\| < \|w_1\|$.

Now for the induction step, if $1 < i \leq m$, and if we have that

$$\|w_0\| < \|w_1\| < \dots < \|w_{1-1}\| ,$$

then repeating the above with w_0, w_1, y_2 replaced with w_{i-1}, w_i, y_{i+1} respectively, we have that $\|w_{i-1}\| < \|w_i\|$. Finally, since $w_m \in C$, we have $\|v\| < \|w_m\| \leq \|c\|$. □

With this lemma we can prove the following.

Theorem 5.8 ((4,3)-Grassmanian). *Let $U = \{u_1, u_2, u_3, u_4\} \subset S^2 \subset \mathbb{R}^3$. If U is $(4,3)$-Grassmanian, then U is equiangular, i.e., $|\langle u_k, u_l\rangle| = c$ for $k \neq l$.*

Proof. We show the contrapositive of the above implication. Namely, if U is not equiangular, then there is a 4 element set $X \subset S^2$ such that

$$\mathcal{M}_\infty(X) < \mathcal{M}_\infty(U),$$

hence U does not have minimal ∞-correlation and is therefore not $(4,3)$-Grassmannian. We need the following lemma

Lemma 5.9. *Let $\{b, y_1, y_2, y_3\} \subset S^2$. If $|\langle b, y_1\rangle|$, $|\langle b, y_2\rangle|$, and $|\langle b, y_3\rangle|$ are not all equal, then there is a constructible $c \in \mathbb{R}^3$ such that*

$$\max\{|\langle b, y_k\rangle| : k = 1, 2, 3\} > \max\left\{\left|\left\langle \frac{c}{\|c\|}, y_k\right\rangle\right| : k = 1, 2, 3\right\}.$$

Furthermore, $\left|\left\langle \frac{c}{\|c\|}, y_1\right\rangle\right| = \left|\left\langle \frac{c}{\|c\|}, y_2\right\rangle\right| = \left|\left\langle \frac{c}{\|c\|}, y_3\right\rangle\right|$.

Proof of Lemma 5.9: Case 1. $y_1, y_2, y_3 \subset S^2$ are linearly dependent.

Then there is $a_1, a_2, a_3 \in \mathbb{R}^3$ with at least one (actually two) $a_k \neq 0$ such that

$$a_1 y_1 + a_2 y_2 + a_3 y_3 = 0.$$

So $\dim(\ker Y) \geq 1$, where Y is a 3×3 matrix with columns y_j. Hence, $\dim(\operatorname{span} Y) = \operatorname{rank} Y \leq 2$. So take $c \in (\operatorname{span} y)^\perp$, then

$$|\langle b, y_k\rangle| > 0 = \left|\left\langle \frac{c}{\|c\|}, y_k\right\rangle\right|, \quad \forall k,$$

since by assumption we know that $|\langle b, y_k\rangle|$ cannot all be equal, hence cannot all equal zero.

Proof of Lemma 5.9: Case 2. $y_1, y_2, y_3 \subset S^2$ are linearly independent.

Let Y be the 3×3 matrix whose columns are y_j. Then Y^T is invertible. Let

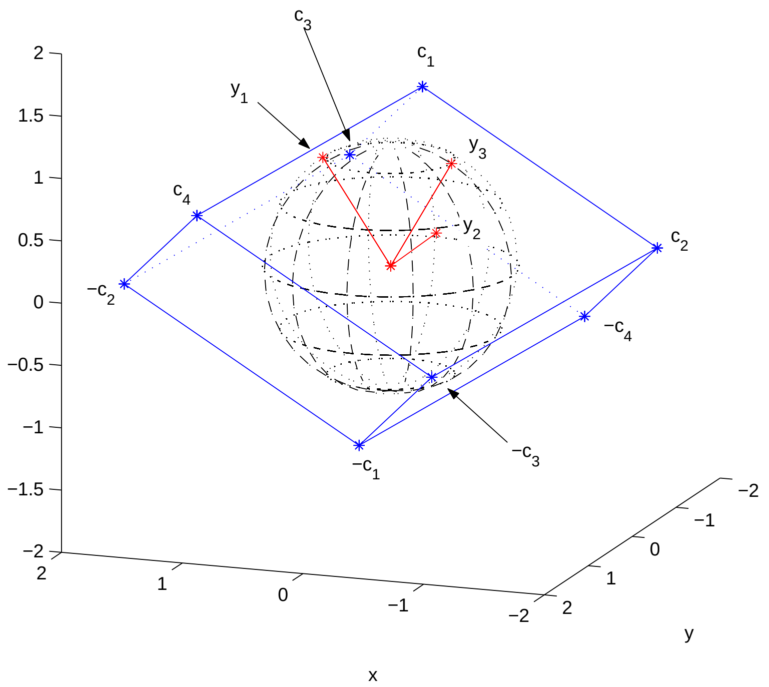


Figure 5.2: An example showing the points $\pm c_k$, $k = 1, \dots, 4$, and their relationship to the vectors y_1, y_2, y_3. Note, y_2 lies on the plane with vertices $\{c_1, c_2, -c_3, c_4\}$.

$c_1, \dots, c_4$ be the columns of the following matrix product

$$\begin{bmatrix} | & | & | & | \\ c_1 & c_2 & c_3 & c_4 \\ | & | & | & | \end{bmatrix} = \left(Y^T\right)^{-1} \begin{bmatrix} 1 & -1 & 1 & 1 \\ 1 & 1 & -1 & 1 \\ 1 & 1 & 1 & -1 \end{bmatrix}, \tag{5.13}$$

see Figure 5.2. Notice that

$$\begin{aligned} c_1 = \left(Y^T\right)^{-1} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} &= \left(Y^T\right)^{-1} \left(\begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix} + \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix} \right) \\ &= c_2 + c_3 + c_4 \end{aligned}$$

Let $c \in \{c_1, \ldots, c_4\}$ such that $\|c\| = \max\{\|c_1\|, \ldots, \|c_4\|\}$, and for $j = 1, 2, 3$, let $v_j = c_{j+1} - c_1$. Now set

$$Q = \left\{ c_1 + \sum_{j=1}^{3} s_j v_j : s_j \in [0,1] \right\}, \quad \text{and} \quad C = \left\{ c_1 + \sum_{j=1}^{3} \varepsilon_j v_j : \varepsilon_j \in \{0,1\} \right\}$$

Next, identify a point in C with a vector $(\varepsilon_1, \varepsilon_2, \varepsilon_3)$. So, for example if $(\varepsilon_1, \varepsilon_2, \varepsilon_3) = (1,0,1)$, then $v = c_1 + v_1 + v_3$. Then, observe that since $c_1 = c_2 + c_3 + c_4$, we have the following bijection between C and $\{\pm c_1, \pm c_2, \pm c_3, \pm c_4\}$,

$$\begin{aligned}
(0,0,0) &\longleftrightarrow & c_1\\
(1,0,0) &\longleftrightarrow & c_2\\
(0,1,0) &\longleftrightarrow & c_3\\
(0,0,1) &\longleftrightarrow & c_4\\
(1,1,0) &\longleftrightarrow & -c_4\\
(1,0,1) &\longleftrightarrow & -c_3\\
(0,1,1) &\longleftrightarrow & -c_2\\
(1,1,1) &\longleftrightarrow & -c_1
\end{aligned}$$

So $\|c\| = \max\{\|\tilde{c}\| : \tilde{c} \in C\}$.

Now, if

$$H = \left\{ v \in \mathbb{R}^3 : |\langle v, y_k \rangle| \leq 1 \text{ for } k = 1,2,3 \right\}$$

then $Q = H$. To see this, check both containments. First we note that (5.13) with $j = 2$ implies

$$Y^T c_2 = \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix} \implies \begin{bmatrix} y_1^T c_2 \\ y_2^T c_2 \\ y_3^T c_2 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix},$$

i.e., $\langle y_1, c_2 \rangle = -1$, $\langle y_2, c_2 \rangle = 1$, and $\langle y_3, c_2 \rangle = 1$. And similarly for the other c_j.

$Q \subset H$. Let $v \in Q$, then

$$|\langle v, y_k \rangle| = \left| \langle c_1, y_k \rangle + \sum_{j=1}^{3} s_j \langle v_j, y_k \rangle \right| = \left| 1 + \sum_{j=1}^{3} s_j \langle v_j, y_k \rangle \right|,$$

but

$$\langle v_j, y_k \rangle = \langle c_{j+1} - c_1, y_k \rangle = \begin{cases} 1 - 1, & \text{if } k = j, \\ -1 - 1, & \text{if } k \neq j, \end{cases}$$

so $|\langle v_j, y_k \rangle| = |1 - 2s_k|$, and $s_k \in [0,1]$ implies $1 - 2s_k \in [-1, 0]$, so $|\langle v_j, y_k \rangle| \leq 1$

$H \subset Q$. Or equivalently, we show $Q^C \subset H^C$. Let $v \notin Q$. Now v_1, v_2, v_3 are the image of $(-2,0,0)^T, (0,-2,0)^T, (0,0,-2)^T$ respectively under the transformation $\left(Y^T\right)^{-1}$, so $\{v_1, v_2, v_3\}$ is a basis for $\mathbb{R}^3$. Thus, there are unique $s_1, s_2, s_3 \in \mathbb{R}^3$ such that

$$v - c_1 = s_1 v_1 + s_2 v_2 + s_3 v_3.$$

So $v = c_1 + s_1 v_1 + s_2 v_2 + s_3 v_3$. Now because of the uniqueness of s_js, $v \notin Q$ implies there is a $j_0 \in \{1,2,3\}$ such that $s_{j_0} \notin [0,1]$. Now $|\langle v, y_{j_0} \rangle| = |1 - 2s_{j_0}|$, and

$$\begin{aligned} s_{j_0} \notin [0,1] &\implies s_{j_0} \in (-\infty, 0) \cup (1, \infty) \\ &\implies -2s_{j_0} \in (-\infty, -2) \cup (0, \infty) \\ &\implies 1 - 2s_{j_0} \in (-\infty, -1) \cup (1, \infty), \end{aligned}$$

so $|\langle v_{j_0} \rangle| > 1$, so $v \notin H$, and we have shown both containments. □

Now to finish the proof of Lemma 5.9, let

$$|\langle b, y_{k_b} \rangle| = \max \{ |\langle b, y_1 \rangle|, |\langle b, y_2 \rangle|, |\langle b, y_3 \rangle| \},$$

and set $\lambda_b = \langle b, y_{k_b}\rangle$. Then for any $k = 1, 2, 3$,

$$\left|\left\langle \frac{b}{\lambda_b}, y_k \right\rangle\right| = \frac{|\langle b, y_k\rangle|}{|\langle b, y_{k_b}\rangle|} \leq \frac{|\langle b, y_{k_b}\rangle|}{|\langle b, y_{k_b}\rangle|} = 1,$$

so $\frac{b}{\lambda_b} \in Q$. Now $v \in C$ implies $\{|\langle v, y_k\rangle| : k = 1, 2, 3\}$ are all equal, so by the contrapositive of this implication, we see that the assumption, $\{|\langle b, y_k\rangle| : k = 1, 2, 3\}$ are not all equal, implies $\frac{b}{\lambda_b} \notin C$. Thus we have shown $\frac{b}{\lambda_b} \in Q \setminus C$. So by Lemma 5.7, $\left\|\frac{b}{\lambda_b}\right\| < \|c\|$, hence $b \in S^2$ implies

$$\frac{1}{|\lambda_b|} = \frac{1}{|\lambda_b|}\|b\| = \left\|\frac{b}{\lambda_b}\right\| < \|c\|$$

and therefore

$$\max\{|\langle b, y_k\rangle| : k = 1, 2, 3\} = |\lambda_b| > \frac{1}{\|c\|} = \max\left\{\left|\left\langle \frac{c}{\|c\|}, y_k \right\rangle\right| : k = 1, 2, 3\right\}$$

Thus we have proven Lemma 5.9

Thus to complete that proof of Theorem 5.8, we suppose $U = \{u_1, u_2, u_3, u_4\}$ is not equiangular. Because U is not equiangular, there is an $m_1 \in \{1, 2, 3, 4\}$ such that if k_1, k_2, k_3 are the remaining indicies, then

1.) $\max\{|\langle u_{m_1}, u_{k_1}\rangle|, |\langle u_{m_1}, u_{k_2}\rangle|, |\langle u_{m_1}, u_{k_3}\rangle|\} = \mathcal{M}_\infty(U)$

2.) $|\langle u_{m_1}, u_{k_1}\rangle|, |\langle u_{m_1}, u_{k_2}\rangle|, |\langle u_{m_1}, u_{k_3}\rangle|$ are not all equal.

Then applying Lemma 5.9 with $b = u_{m_1}$, and $\{y_1, y_2, y_3\} = \{u_{k_1}, u_{k_2}, u_{k_3}\}$, there is a $c \in \mathbb{R}^3$ such that

$$\max\{|\langle u_{m_1}, u_{k_i}\rangle| : i = 1, 2, 3, \} > \max\left\{\left|\left\langle \frac{c}{\|c\|}, u_{k_i} \right\rangle\right| : i = 1, 2, 3, \right\}.$$

Let $x_{m_1} = \frac{c}{\|c\|}$, see step 2 in Figure 5.3. Now since we have only moved the point u_{m_1} to x_{m_1}, the remaining correlations are uneffected since they do not involve

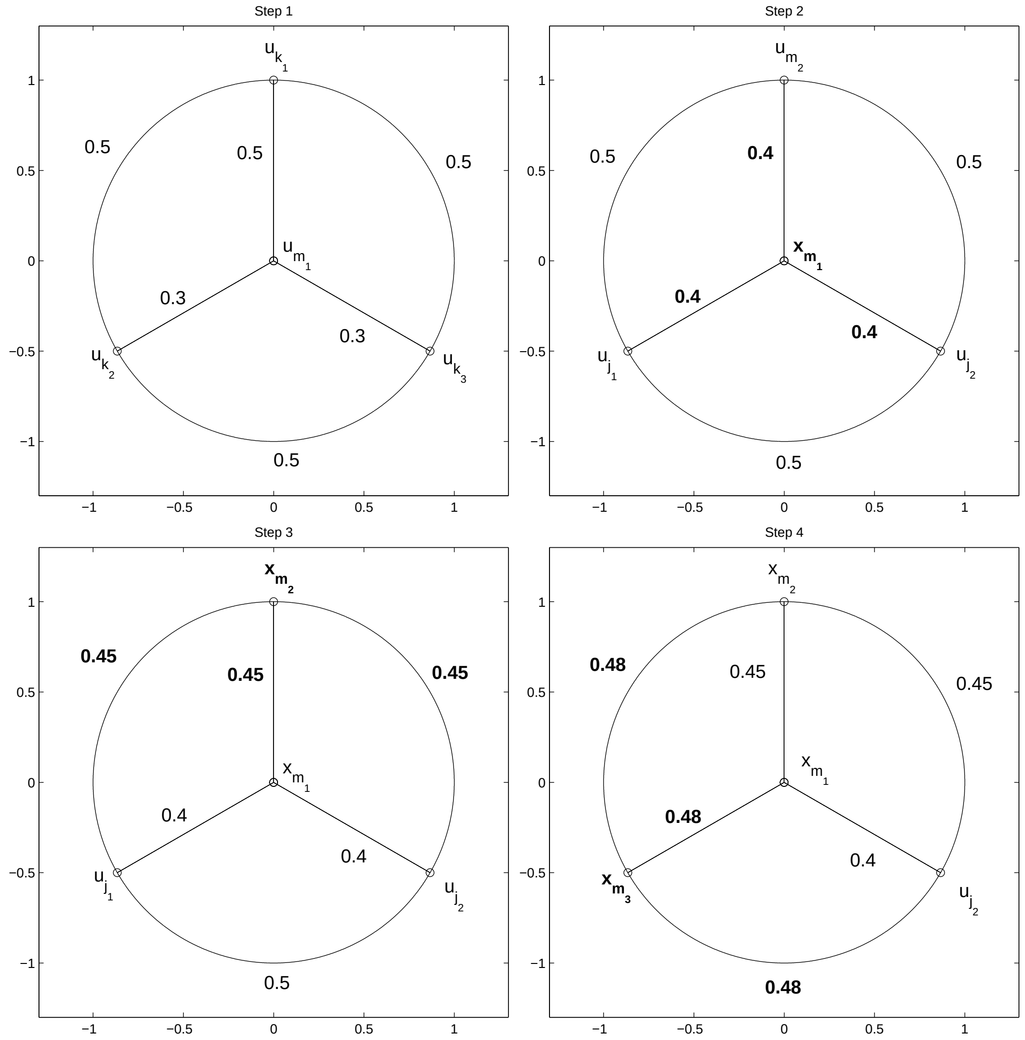


Figure 5.3: An example of the four steps in proving Theorem 5.8. A number next to an edge represents the inner product of the two boundary points of the edge.

u_{m_1}, thus

$$\begin{aligned}\mathcal{M}_\infty(U) &= \max\{|\langle u_{m_1}, u_{k_i}\rangle| : i = 1,2,3,\} \\ &> \max\{|\langle x_{m_1}, u_{k_i}\rangle| : i = 1,2,3,\} \\ &=: \alpha\end{aligned} \tag{5.14}$$

Now either $\mathcal{M}_\infty(\{x_{m_1}, u_{k_1}, u_{k_2}, u_{k_3}\}) = \alpha$, or there is an $m_2 \in \{1,2,3,4\} \setminus \{m_1\}$ such that if m_1, j_1, j_2 are the remaining indicies, then

$$\begin{aligned}&1.)\quad \mathcal{M}_\infty(U) = \max\{|\langle u_{m_2}, u_{j_1}\rangle|, |\langle u_{m_2}, u_{j_2}\rangle|\} \\ &2.)\quad |\langle u_{m_2}, x_{m_1}\rangle|, |\langle u_{m_2}, u_{j_1}\rangle|, |\langle u_{m_2}, u_{j_2}\rangle| \text{ are not all equal,}\end{aligned} \tag{5.15}$$

where (5.15) follows from (5.14). In this case we apply Lemma 5.9 to $b = u_{m_2}$, and $\{y_1, y_2, y_3\} = \{u_{j_1}, u_{j_2}, x_{m_1}\}$. So there is a $c' \in \mathbb{R}^3$ such that

$$\begin{aligned}&\max\{|\langle u_{m_2}, x_{m_1}\rangle|, |\langle u_{m_2}, u_{j_1}\rangle|, |\langle u_{m_2}, u_{j_2}\rangle|\} \\ &\quad = \max\{|\langle u_{m_2}, u_{j_1}\rangle|, |\langle u_{m_2}, u_{j_2}\rangle|\} \\ &\quad > \max\left\{\left|\left\langle \frac{c'}{\|c'\|}, x_{m_1}\right\rangle\right|, \left|\left\langle \frac{c'}{\|c'\|}, u_{j_1}\right\rangle\right|, \left|\left\langle \frac{c'}{\|c'\|}, u_{j_2}\right\rangle\right|\right\}.\end{aligned}$$

Let $x_{m_2} = \frac{c'}{\|c'\|}$, see step 3 in Figure 5.3. Thus

$$\begin{aligned}\mathcal{M}_\infty(U) &= \max\{|\langle u_{m_2}, u_{j_1}\rangle|, |\langle u_{m_2}, u_{j_2}\rangle|\} \\ &> \max\{|\langle x_{m_2}, x_{m_1}\rangle|, |\langle x_{m_2}, u_{j_1}\rangle|, |\langle x_{m_2}, u_{j_2}\rangle|\} \\ &=: \alpha'\end{aligned} \tag{5.16}$$

Therefore (5.14) and (5.16) imply

$$\begin{aligned}&\mathcal{M}_\infty(U) \\ &\quad > \max\left\{\begin{array}{ccc} |\langle x_{m_1}, u_{j_1}\rangle|, & |\langle x_{m_1}, u_{j_2}\rangle|, & \\ |\langle x_{m_2}, x_{m_1}\rangle|, & |\langle x_{m_2}, u_{j_1}\rangle|, & |\langle x_{m_2}, u_{j_2}\rangle| \end{array}\right\} \\ &\quad = \max\{\alpha, \alpha'\},\end{aligned} \tag{5.17}$$

because $j_1, j_2 \in \{k_1, k_2, k_3\}$.

So either $\mathcal{M}_\infty(\{x_{m_1}, x_{m_2}, u_{j_1}, u_{j_2}\}) = \max\{\alpha, \alpha'\}$ or else $\mathcal{M}_\infty(U) = |\langle u_{j_1}, u_{j_2}\rangle|$.

In the latter case, (5.16) implies that $|\langle u_{j_1}, u_{j_2}\rangle|, |\langle u_{j_1}, x_{m_1}\rangle|, |\langle u_{j_1}, x_{m_2}\rangle|$ are not all equal, so we apply Lemma 5.9 to $b = u_{j_1}$, and $\{y_1, y_2, y_3\} = \{u_{j_2}, x_{m_1}, x_{m_2}\}$. Thus there is a $c'' \in \mathbb{R}^3$ such that

$$\begin{aligned}\max&\{|\langle u_{j_1}, x_{m_1}\rangle|, |\langle u_{j_1}, x_{m_2}\rangle|, |\langle u_{j_1}, u_{j_2}\rangle|\}\\ &= |\langle u_{j_2}, u_{j_1}\rangle|\\ &> \max\left\{\left|\left\langle \frac{c''}{\|c''\|}, x_{m_1}\right\rangle\right|, \left|\left\langle \frac{c''}{\|c''\|}, x_{m_2}\right\rangle\right|, \left|\left\langle \frac{c''}{\|c''\|}, u_{j_2}\right\rangle\right|\right\}. \qquad (5.18)\end{aligned}$$

Let $x_{m_3} = \frac{c''}{\|c''\|}$ and let $X = x_{m_1}, x_{m_2}, x_{m_3}, u_{j_2}$. Then (5.17) and (5.18) imply

$$\begin{aligned}\mathcal{M}_\infty(U) &> \max\left\{\begin{matrix}|\langle x_{m_3}, x_{m_1}\rangle|, & |\langle x_{m_3}, x_{m_2}\rangle|, & |\langle x_{m_3}, u_{j_2}\rangle|,\\ |\langle x_{m_2}, x_{m_1}\rangle|, & |\langle x_{m_2}, u_{j_2}\rangle|, & |\langle x_{m_1}, u_{j_2}\rangle|\end{matrix}\right\}\\ &= \mathcal{M}_\infty(X)\end{aligned}$$

□

Next we show that if a 4 element set is equiangular then the vectors are parallel to the diagonals of a cube or 4 of the diagonals of an icosahedron.

Theorem 5.10. *If $u_1, u_2, u_3, u_4 \in S^2$ and $|\langle u_k, u_l\rangle| = \alpha$ for $k, l \in \{1, \ldots, 4\}$ with $k \neq l$, then*

$$\alpha = \frac{1}{3} \text{ or } \frac{1}{\sqrt{5}}$$

Proof. Since sign changes and rotations do not effect inner products, let P be an element of SO_3 which rotates u_1 to δ_3, for $k = 1, 2, 3, 4$ let

$$\varepsilon_k = \operatorname{sign}\langle Px_k, \delta_3\rangle,$$

	$\langle w_2, w_3\rangle$	$\langle w_2, w_4\rangle$	$\langle w_3, w_4\rangle$	
case 1	α	α	α	impossible
case 2	$-\alpha$	α	α	$\alpha = 1/\sqrt{5}$
case 3	$-\alpha$	$-\alpha$	α	$\alpha = 1/\sqrt{5}$
case 4	$-\alpha$	$-\alpha$	$-\alpha$	$\alpha = 1/3$

Table 5.3: Four main cases in the proof of Theorem 5.10.

and let $Q \in SO_3$ so that Q fixes δ_3 and Q rotates $\varepsilon_k P x_2$ to the positive xz-plane, i.e $\langle Q\varepsilon_k P x_2, \delta_1\rangle \geq 0$ and $\langle Q\varepsilon_k P x_2, \delta_2\rangle = 0$. Then for $k \neq l$

$$\alpha = |\langle u_k, u_l\rangle| = |\langle \varepsilon_k Q P u_k, \varepsilon_l Q P u_l\rangle| = |\langle w_k, w_l\rangle|$$

where $w_k = \varepsilon_k Q P u_k$. Now by the choice of ε_k, for $k = 2, 3, 4$,

$$\alpha = |\langle w_1, w_k\rangle| = \langle \delta_3, w_k\rangle\,,$$

so the third component of w_k is α. Also $0 = \langle \delta_2, w_2\rangle$, and $w_2 \in S^2$, so the first component of w_2 is $\sqrt{1-\alpha^2}$. Therefore, we have

$$\begin{aligned} w_1 &= (0,0,1)^T \\ w_2 &= (\sqrt{1-\alpha^2}, 0, \alpha)^T \\ w_3 &= (x_3, y_3, \alpha)^T \\ w_4 &= (x_4, y_4, \alpha)^T. \end{aligned} \tag{5.19}$$

Now we have four cases, see Table 5.3, where both case 2 and 3 have three subcases which by relabeling can be reduced to the considered case.

Case 1. For $k = 3, 4$, $\langle w_2, w_k\rangle = \alpha$ implies

$$x_k = \frac{\alpha - \alpha^2}{\sqrt{1-\alpha^2}} = \alpha\sqrt{\frac{1-\alpha}{1+\alpha}}. \tag{5.20}$$

Then (5.20) and $\|w_k\|^2 = 1$ implies

$$y_k = \pm\sqrt{\frac{1+\alpha-2\alpha^2}{1+\alpha}}. \tag{5.21}$$

In addition (5.20) and $\langle w_3, w_4\rangle = \alpha$ implies

$$y_3 \cdot y_4 = \frac{\alpha}{1+\alpha} \tag{5.22}$$

Now combining (5.21) and (5.22), we have

$$-\frac{1+\alpha-2\alpha^2}{1+\alpha} = \frac{\alpha}{1+\alpha} \implies 2\alpha^2 - 2\alpha - 1 = 0 \implies \alpha \in \mathbb{C},$$

hence case 1 is impossible.

Case 2. Now $\langle w_2, w_3\rangle = \alpha$ implies

$$x_3 = \frac{\alpha - \alpha^2}{\sqrt{1-\alpha^2}}, \tag{5.23}$$

and $\langle w_2, w_4\rangle = -\alpha$ implies

$$x_4 = \frac{-\alpha - \alpha^2}{\sqrt{1-\alpha^2}}, \tag{5.24}$$

and (5.23), and (5.24) imply

$$y_3 \cdot y_4 = \alpha. \tag{5.25}$$

Then (5.23) and $\|w_3\|^2 = 1$ implies

$$y_3^2 = -\frac{2\alpha^2 - \alpha - 1}{\alpha + 1} = \frac{(2\alpha+1)(\alpha-1)}{\alpha+1}, \tag{5.26}$$

and (5.24) and $\|w_4\|^2 = 1$ implies

$$y_4^2 = \frac{2\alpha^2 + \alpha - 1}{\alpha - 1} = \frac{(2\alpha-1)(\alpha+1)}{\alpha-1}. \tag{5.27}$$

Finally, (5.25), (5.26), and (5.27) imply

$$-\alpha^2 = (2\alpha+1)(2\alpha-1) \implies \alpha = \pm\frac{1}{\sqrt{5}}.$$

Since α is assumed to be positive, we have proven case 2.

Case 3. For $k = 3, 4$, $\langle w_2, w_k \rangle = -\alpha$ implies

$$x_k = \frac{-\alpha - \alpha^2}{\sqrt{1-\alpha^2}} = -\alpha\sqrt{\frac{1+\alpha}{1-\alpha}}. \tag{5.28}$$

Then (5.28) and $\|w_k\|^2 = 1$ imply

$$y_k^2 = \frac{2\alpha^2 + \alpha - 1}{\alpha - 1} = \frac{(2\alpha-1)(\alpha+1)}{\alpha-1}. \tag{5.29}$$

So (5.28) and $\langle w_3, w_4 \rangle = \alpha$ imply

$$y_3 \cdot y_4 = \frac{\alpha - 3\alpha^2}{1-\alpha} \tag{5.30}$$

Now combining (5.29) and (5.30), we have

$$-\frac{(2\alpha-1)(\alpha+1)}{\alpha-1} = -y_k^2 = y_3 \cdot y_4 = \frac{\alpha - 3\alpha^2}{1-\alpha} \implies \alpha = \pm\frac{1}{\sqrt{5}},$$

and since α is positive, we have shown case 3.

Case 4. This is the same as case 3 except $\langle w_3, w_4 \rangle = \alpha$ and (5.28) imply

$$y_3 \cdot y_4 = \frac{\alpha(\alpha+1)}{\alpha-1} \tag{5.31}$$

Now combining (5.29) and (5.31), we have

$$-\frac{(2\alpha-1)(\alpha+1)}{\alpha-1} = -y_k^2 = y_3 \cdot y_4 = \frac{\alpha(\alpha+1)}{\alpha-1} \implies \alpha = \frac{1}{3},$$

and we have proven case 4.

Hence we have proven the theorem. □

Now by Theorem 5.10, since $\frac{1}{\sqrt{5}} > \frac{1}{3}$ we see that the $(4, 3)$-Grassmannian bound is $\frac{1}{3}$ which is also seen to be optimal by inspection.

5.4 $(5, 3)$-Grassmannian frames

We first introduce some ideas from convex analysis, [Pan93, Lay82, Web94].

Definition 5.11. A set $A \subset \mathbb{R}^n$ is *convex* if for any $x_1, x_2 \in A$, and for any $\lambda \in [0, 1]$,

$$\lambda x_1 + (1 - \lambda)x_2 \in A.$$

A point $x \in A$ is an *extreme point* of A if whenever $x = \lambda x_1 + (1 - \lambda)x_2$, where $0 < \lambda < 1$ and $x_1, x_2 \in A$, then $x = x_1 = x_2$. Given a set $A \subset \mathbb{R}^n$, the *convex hull* of A is

$$\operatorname{Hull}(A) = \left\{ \sum_{j=1}^{m} \lambda_j x_j : \sum_{j=1}^{m} \lambda_j = 1, \lambda_j > 0, x_j \in A, m \in \mathbb{N} \right\}.$$

There is the following relationship between extreme points, convex hulls and convex sets, [KM40].

Theorem 5.12. *A nonempty bounded convex set in $\mathbb{R}^d$ is the convex hull of its set of extreme points.*

We need the following 2 Propositions.

Proposition 5.13. *Let $N \geq d$, let $Y = \{y_1, \ldots, y_N\} \subset S^{d-1} \subset \mathbb{R}^d$, and assume* $\operatorname{span}(Y) = \mathbb{R}^d$. *Let*

$$Q = \left\{ v \in \mathbb{R}^d : |\langle v, y_k \rangle| \leq 1, \quad \text{for } k = 1, \ldots, N \right\}$$

and let C be the set of extreme points of Q. Then

1.) Q is a bounded convex set,

2.) If $v_0 \in C$ then there are at least d distinct $k_1, \ldots, k_d \in \{1, \ldots, N\}$ such that $|\langle v_0, y_{k_i} \rangle| = 1$ for $i = 1, \ldots, d$,

3.) $|C| \leq \binom{N}{d} 2^d < \infty$.

Proof of 1. First, to show Q is convex, let $x_1, x_2 \in Q$. Then for any $\lambda \in [0,1]$, and for any $k \in \{1, \dots, N\}$,

$$\begin{aligned} |\langle \lambda x_1 + (1-\lambda)x_2, y_k\rangle| &\leq \lambda \, |\langle x_1, y_k\rangle| + (1-\lambda)\, |\langle x_2, y_k\rangle| \\ &\leq \lambda + (1-\lambda) \\ &= 1. \end{aligned}$$

Next, we show Q is bounded. Now since $\operatorname{span}(Y) = \mathbb{R}^d$, Proposition 4.2 implies Y is a frame for $\mathbb{R}^d$. Let S be the associated frame operator, and A and B be the lower and upper frame bounds respectively. Now S is invertible, so we can set $v_j = S^{-1}y_j$ for $j = 1, \dots N$. Then we have

$$\|v_j\| = \left\|S^{-1}y_j\right\| \leq \left\|S^{-1}\right\| \|y_j\| = \frac{1}{A}.$$

Now, for any $x \in \mathbb{R}^d$,

$$x = S^{-1}Sx = \sum_{j=1}^{N} \langle x, y_j\rangle \, S^{-1}(y_j) = \sum_{j=1}^{N} \langle x, y_j\rangle \, v_j.$$

Thus, given $x \in Q$,

$$\|x\| = \left\|\sum_{j=1}^{N} \langle x, y_j\rangle \, v_j\right\| \leq \sum_{j=1}^{N} |\langle x, y_j\rangle| \, \|v_j\| = \sum_{j=1}^{N} \|v_j\| \leq \frac{N}{A}$$

□

Proof of 2. We prove the contrapositive. Assume $|\langle v_0, y_k\rangle| = 1$ for less than d vectors in Y, i.e, by relabeling, assume there is an $m \geq 0$ such that

$$\begin{aligned} |\langle v_0, y_k\rangle| &= 1, \quad \text{for } k \in \mathbb{N} \text{ with } k \leq m, \\ |\langle v_0, y_k\rangle| &< 1, \quad \text{for } k = m+1, \dots, N. \end{aligned}$$

We now show that v_0 is not an extreme point by constructing $x_1, x_2 \in Q$ with $x_1 \neq x_2$ such that there is a $\lambda \in (0,1)$ for which

$$v_0 = \lambda x_1 + (1-\lambda)x_2.$$

Let $\tilde{Y} = \operatorname{span}\{y_1, \ldots, y_m\}$, where $\tilde{Y}$ is empty if $m = 0$. Since by assumption $m < d$, we have $\dim(\tilde{Y}) < d$. Hence let $z \in \tilde{Y}^{\perp} \cap S^{d-1}$. Set

$$\beta = \max\{|\langle v_0, y_k\rangle| : k = m+1, \ldots, N\}.$$

Then by the choice of m, we have $\beta < 1$. Now set

$$x_1 = v_0 + \frac{1-\beta}{2}z, \quad \text{and} \quad x_2 = v_0 - \frac{1-\beta}{2}z.$$

Notice $\beta < 1$ implies $\|x_1 - x_2\| = (1-\beta)\|z\| = 1-\beta > 0$, hence $x_1 \neq x_2$. Furthermore if $\lambda = \frac{1}{2}$, then

$$\lambda x_1 + (1-\lambda)\, x_2 = \frac{1}{2}v_0 + \frac{1-\beta}{4}z + \frac{1}{2}v_0 - \frac{1-\beta}{4}z = v_0.$$

Finally, we check that x_1 and x_2 are in Q. For $k = 1, \ldots, m$, and $l = 1, 2$,

$$|\langle x_l, y_k\rangle| = \left|\langle v_0, y_k\rangle \pm \frac{1-\beta}{2}\langle z, y_k\rangle\right| = |\langle v_0, y_k\rangle| = 1$$

and for $k = m+1, \ldots, N$, and $l = 1, 2$,

$$\begin{aligned} |\langle x_l, y_k\rangle| &= \left|\langle v_0, y_k\rangle \pm \frac{1-\beta}{2}\langle z, y_k\rangle\right| \\ &\leq |\langle v_0, y_k\rangle| + \frac{1-\beta}{2}|\langle z, y_k\rangle| \\ &\leq \beta + \frac{1-\beta}{2}\|z\|\, y_k \\ &= \frac{1+\beta}{2} < 1. \end{aligned}$$

Hence, $v_0 \in Q \setminus C$. □

Proof of 3. If v_0 is an extreme point then v_0 must satisfy at least d of the N equations which define Q. Therefore we count the number of ways we can pick d distinct elements, y_k, from Y to satisfy the d equations $|\langle v_0, y_k \rangle| = 1$. There are $\binom{N}{d}$ d-element subsets of $\{1, \ldots, N\}$, and because of the absolute value there are two choices for the equation each v_0 can satsify, namely $\langle v_0, y_k \rangle = 1$ or $\langle v_0, y_k \rangle = -1$. Note, if any one of the remaining $N - d$ inequalities is not satisfied by v_0, then v_0 is not an extreme point. Which shows we can have less than $\binom{N}{d} 2^d$ extreme points for a given arrangement of y_ks. □

Under the same hypotheses of Proposition 5.13 we have

Proposition 5.14. *Let $N, d, Y, Q,$ and C be as in Proposition 5.13, and let $c \in C$ such that $\|c\| = \max\{\|\tilde{c}\| : \tilde{c} \in C\}$. Then for any $v \in Q \setminus C$,*

$$\|v\| < \|c\| .$$

Proof. Let $v \in Q \setminus C$. Then there is a $\lambda \in (0,1)$, and there are $x_1, x_2 \in Q$ with $x_1 \neq x_2$ such that $v = \lambda x_1 + (1-\lambda)x_2$. Consider the function $f : \mathbb{R} \to \mathbb{R}$ by

$$f(\lambda) = \|\lambda x_1 + (1-\lambda)x_2\| .$$

Now we check that f is continuous on $\mathbb{R}$. Let $\lambda_0 \in \mathbb{R}$, let $\varepsilon > 0$ be given, and choose $\delta < \frac{\varepsilon}{\|x_1 - x_2\|}$. Then whenever $|\lambda - \lambda_0| < \delta$, we have

$$\begin{aligned}
|f(\lambda) - f(\lambda_0)| &= \Big| \|\lambda x_1 + (1-\lambda)x_2\| - \|\lambda_0 x_1 + (1-\lambda)_0 x_2\| \Big| \\
&\leq \|\lambda x_1 + (1-\lambda)x_2 - \lambda_0 x_1 - (1-\lambda)_0 x_2\| \\
&= \|(\lambda - \lambda_0)(x_1 - x_2)\| \\
&= |\lambda - \lambda_0| \, \|x_1 - x_2\| \\
&< \delta \, \|x_1 - x_2\| < \varepsilon.
\end{aligned}$$

Now set $g(\lambda) = f(\lambda)^2$. Then $g(\lambda)$ is also continuous on $\mathbb{R}$ and

$$\begin{aligned} g(\lambda) &= \|\lambda x_1 + (1-\lambda)x_2\|^2 \\ &= \lambda^2 \|x_1\|^2 + 2(\lambda - \lambda^2)\langle x_1, x_2\rangle - 2(1-\lambda)\|x_2\|^2, \end{aligned}$$

so

$$\begin{aligned} g'(\lambda) &= 2\lambda \|x_1\|^2 + (2-4\lambda)\langle x_1, x_2\rangle - 2(1-\lambda)\|x_2\|^2 \\ &= 2\lambda\left(\|x_1\|^2 + 2\langle x_1, x_2\rangle \|x_2\|^2\right) + 2\langle x_1, x_2\rangle - 2\|x_2\|^2 \\ &= 2\lambda\|x_1 - x_2\|^2 + 2\langle x_1 - x_2, x_2\rangle, \end{aligned}$$

and $g'(\lambda) = 0$ at

$$\lambda_* = -\frac{\langle x_1 - x_2, x_2\rangle}{\|x_1 - x_2\|^2}.$$

Furthermore, for all $\lambda \in \mathbb{R}$

$$g''(\lambda) = 2\|x_1 - x_2\|^2 > 0, \tag{5.32}$$

so g attains a minimum at λ_*, and for all $\lambda \neq \lambda_*$, we have $g(\lambda) > g(\lambda_*)$. Now if we restrict g to $[0,1]$, then g achieves its maximum and minimum on $[0,1]$. Thus if $\lambda_* \in [0,1]$, then by(5.32),

$$\min_{\lambda\in[0,1]} g(\lambda) = g(\lambda_*) \quad \text{and} \quad \max_{\lambda\in[0,1]} g(\lambda) = \max\{g(0), g(1)\}.$$

If $\lambda_* \notin [0,1]$, then

$$\min_{\lambda\in[0,1]} g(\lambda) = \min\{g(0), g(1)\} \quad \text{and} \quad \max_{\lambda\in[0,1]} g(\lambda) = \max\{g(0), g(1)\}.$$

In either case the maximum of g occurs at the at one of the end points. Furthermore at interior points, g is strictly less that the maximum value.

Now since $\|v\|^2 = g(\lambda_0)$ for some $\lambda_0 \in (0,1)$, (5.32) implies

$$\|v\|^2 = g(\lambda_0) < \max_{\lambda\in[0,1]} g(\lambda) = \max\{g(0), g(1)\} = \max\left\{\|x_1\|^2, \|x_2\|^2\right\}.$$

Thus, we have shown that

$$v \in Q \setminus C \Longrightarrow \text{ there is an } x \in Q \text{ such that } \|v\| < \|x\| \tag{5.33}$$

$$\Longrightarrow \|v\| < \max\{\|x\| : x \in Q\}. \tag{5.34}$$

Now Q is a bounded closed set so by continuity of $\|\cdot\|$, the maximum norm is achieved on Q, but (5.34) shows that this maximum norm is not achieved on $Q \setminus C$. Thus

$$\max\{\|x\| : x \in Q\} = \max\{\|x\| : x \in C\} = \|c\|.$$

so for any $v \in Q \setminus C$ we have, $\|v\| < \max\{\|x\| : x \in Q\} = \|c\|$. □

Using these Propositions we can compute the $(5,3)$-Grassmannian bound. Again we follow the basic geometric idea in [Tot65], but we use the propositions above, which can be implemented as explicit algorithms, to reduce the correlation of a given frame. To compute the $(5,3)$ case, we need the following two lemmas,

Lemma 5.15. *Let $U = \{b, y_1, y_2, y_3, y_4\} \subset S^2 \subset \mathbb{R}^3$, and let $\alpha = \mathcal{M}_\infty(U)$. Assume $|\langle b, y_1\rangle| < \alpha$ and $|\langle b, y_2\rangle| < \alpha$. Then there exists a $c \in \mathbb{R}^3$ such that*

$$\left|\left\langle \frac{c}{\|c\|}, y_k \right\rangle\right| < \alpha \quad \textit{for } k = 1, 2, 3, 4.$$

Proof. If both $|\langle b, y_3\rangle| < \alpha$ and $|\langle b, y_4\rangle| < \alpha$, then take $c = b$. Otherwise, without loss of generality, assume $|\langle b, y_3\rangle| = \alpha$. We have 2 cases:

Case 1. $\dim(\operatorname{span}\{y_1, \ldots, y_4\}) < 3$.

Then similar to Lemma 5.9, choose $c \in (\operatorname{span}\{y_1, \ldots, y_4\})^\perp$. Then by Theorem 5.4

$$\left\langle \frac{c}{\|c\|}, y_k \right\rangle = 0 < \frac{1}{\sqrt{6}} \leq \alpha.$$

Case 2. $\operatorname{span}\{y_1, \dots y_4\} = \mathbb{R}^3$.

Let

$$Q = \left\{v \in \mathbb{R}^d : |\langle v, y_k\rangle| \le 1, k = 1, \dots, 4\right\}$$

and let C be the set of extreme points of Q. Then by Proposition 5.13, Q is bounded and convex and C is finite. Let c be a point in C of maximum norm. Then by assumption

$$\left|\left\langle \frac{b}{\alpha}, y_1\right\rangle\right| < 1, \quad \left|\left\langle \frac{b}{\alpha}, y_2\right\rangle\right| < 1, \quad \left|\left\langle \frac{b}{\alpha}, y_3\right\rangle\right| = 1, \quad \left|\left\langle \frac{b}{\alpha}, y_4\right\rangle\right| \le 1,$$

which shows that $\frac{b}{\alpha}$ can satisfy with equality at most two of the four equations which define Q. Then by Proposition 5.13, $\frac{b}{\alpha}$ is not an extreme point of Q. Hence by Proposition 5.14,

$$\frac{1}{\alpha} = \left\|\frac{b}{\alpha}\right\| < \|c\|$$

Therefore, since $c \in C \subset Q$, we have $|\langle c, y_k\rangle| \le 1$ and

$$\left|\left\langle \frac{c}{\|c\|}, y_k\right\rangle\right| \le \frac{1}{\|c\|} < \alpha$$

for $k = 1, 2, 3, 4$.

□

Lemma 5.16. *Let $U = \{u_1, \dots, u_5\}$ be a $(5,3)$-Grassmannian frame, and let $\alpha = \mathcal{M}_\infty(U)$. Then for any j, there are distinct $j_1, j_2, j_3 \in \{1, \dots, 5\} \setminus \{j\}$ such that*

$$|\langle u_j, u_{j_k}\rangle| = \alpha \quad \textit{for } k = 1, 2, 3.$$

Proof. We prove the contrapositive. By relabeling if necessary, without loss of generality, assume $|\langle u_1, u_2\rangle| < \alpha$ and $|\langle u_1, u_3\rangle| < \alpha$. We use Lemma 5.15 to construct a new set W for which $\mathcal{M}_\infty(W) < \alpha$. This shows U is not $(5,3)$-Grassmannian.

First let $b = u_1$ and $\{y_1, \ldots, y_4\} = \{u_2, \ldots, u_5\}$ and apply Lemma 5.15. Then there is a $c_1 \in \mathbb{R}^3$ such that

$$\left|\left\langle \frac{c_1}{\|c_1\|}, u_k \right\rangle\right| < \alpha \quad \text{for } k = 2, 3, 4, 5.$$

Second consider the set $\tilde{U} := \{u_2, \ldots, u_5\}$. We have two cases,

Case 1. There exist $j_0, k_0 \in \{2, 3, 4, 5\}$ with $j_0 \neq k_0$, for which $|\langle u_{j_0}, u_{k_0}\rangle| < \alpha$.

For ease in notation, by relabeling if necessary, we assume $j_0 = 2$ and $k_0 = 3$. In this case, we can apply Lemma 5.15 with $b = u_2$ and $\{y_1, \ldots, y_4\} = \left\{\frac{c_1}{\|c_1\|}, u_3, u_4, u_5\right\}$, and construct $c_2 \in \mathbb{R}^3$ such that

$$\left|\left\langle \frac{c_1}{\|c_1\|}, \frac{c_2}{\|c_2\|} \right\rangle\right| < \alpha$$

and

$$\max\left\{\left|\left\langle \frac{c_2}{\|c_2\|}, u_k \right\rangle\right| : k = 3, 4, 5\right\} < \alpha$$

Now we can apply Lemma 5.15 to the remaining points and produce a frame with strictly smaller ∞-correlation. Namely, since $\left|\left\langle \frac{c_i}{\|c_i\|}, u_3 \right\rangle\right| < \alpha$ for $i = 1, 2$, we let $b = u_3$ and $\{y_1, \ldots, y_4\} = \left\{\frac{c_1}{\|c_1\|}, \frac{c_2}{\|c_2\|}, u_4, u_5\right\}$. Then, by Lemma 5.15, there is a $c_3 \in \mathbb{R}^d$ such that

$$\left|\left\langle \frac{c_3}{\|c_3\|}, \frac{c_i}{\|c_i\|} \right\rangle\right| < \alpha \quad \text{for } i = 1, 2,$$

and

$$\max\left\{\left|\left\langle \frac{c_3}{\|c_3\|}, u_k \right\rangle\right| : k = 4, 5\right\} < \alpha.$$

Finally, apply Lemma 5.15 one last time to $b = u_4$ and

$$\{y_1, \ldots, y_4\} = \left\{\frac{c_1}{\|c_1\|}, \frac{c_2}{\|c_2\|}, \frac{c_3}{\|c_3\|}, u_5\right\},$$

and obtain $c_4 \in \mathbb{R}^d$ for which

$$\left|\left\langle \frac{c_4}{\|c_4\|}, \frac{c_i}{\|c_i\|} \right\rangle\right| < \alpha \quad \text{for } i = 1, 2, 3,$$

and

$$\left|\left\langle \frac{c_4}{\|c_4\|}, u_5 \right\rangle\right| < \alpha.$$

Thus, if we let $W = \left\{\frac{c_1}{\|c_1\|}, \frac{c_2}{\|c_2\|}, \frac{c_3}{\|c_3\|}, \frac{c_4}{\|c_4\|}, u_5\right\}$, then by construction, for any $i, j \in \{1, \ldots, 4\}$, if $i \neq j$, then we have $\left|\left\langle \frac{c_i}{\|c_i\|}, \frac{c_1}{\|c_1\|} \right\rangle\right| < \alpha$ and $\left|\left\langle \frac{c_i}{\|c_i\|}, u_5 \right\rangle\right| < \alpha$. Hence,

$$\mathcal{M}_\infty(W) < \alpha = \mathcal{M}_\infty(U),$$

so U is not $(5,3)$-Grassmannian. This finishes case 1. □

Case 2. $\tilde{U}$ is equiangular.

Since $\tilde{U}$ has four elements, Theorem 5.10 implies $\alpha = 1/3$ or $\alpha = 1/\sqrt{5}$. If $\alpha = 1/3$, then set

$$\beta = \max\left\{\left|\left\langle \frac{c_1}{\|c_1\|}, u_k \right\rangle\right| : k = 2, 3, 4, 5\right\}.$$

Thus by construction of c_1, we have $\beta < \frac{1}{3}$ and

$$\mathcal{M}_\infty\left(\left\{\frac{c_1}{\|c_1\|}\right\} \cup \tilde{U}\right) = \max\left\{\frac{1}{3}, \beta\right\} = \frac{1}{3},$$

but Theorem 5.4 with $N = 5$ and $d = 3$ implies

$$\frac{1}{\sqrt{6}} \leq \mathcal{M}_\infty\left(\left\{\frac{c_1}{\|c_1\|}\right\} \cup \tilde{U}\right) = \frac{1}{3}$$

which is a contradiction.

Thus, $\alpha = \frac{1}{\sqrt{5}}$, and $|\langle u_1, u_k \rangle| < \alpha$, for $k = 2, 3, 4, 5$ and $|\langle u_k, u_j \rangle| = \alpha$, for $k \neq j$ and $k, j \in \{2, 3, 4, 5\}$.

We seek to find a contradiction. We can reduce to the following general position by using rotations and sign changes as in Theorem 5.10. Thus, without loss of generality

$$\begin{aligned} u_2 &= (0,0,1)^T \\ u_3 &= (\sqrt{1-\alpha^2}, 0, \alpha)^T \\ u_4, u_5 &\in \{p_1, p_2, p_3, p_4\}, \end{aligned}$$

where

$$\begin{aligned} p_1 &= \left(\alpha\sqrt{\frac{1-\alpha}{1+\alpha}}, \sqrt{\frac{(1+2\alpha)(1-\alpha)}{1+\alpha}}, \alpha\right)^T \\ &= \left(\sqrt{1-\alpha^2}\cos\left(\frac{2\pi}{5}\right), \sqrt{1-\alpha^2}\sin\left(\frac{2\pi}{5}\right), \alpha\right)^T \end{aligned}$$

and

$$\begin{aligned} p_2 &= \left(-\alpha\sqrt{\frac{1+\alpha}{1-\alpha}}, \sqrt{\frac{(1-2\alpha)(1+\alpha)}{1-\alpha}}, \alpha\right)^T \\ &= \left(\sqrt{1-\alpha^2}\cos\left(\frac{4\pi}{5}\right), \sqrt{1-\alpha^2}\sin\left(\frac{4\pi}{5}\right), \alpha\right)^T \end{aligned}$$

and

$$\begin{aligned} p_3 &= \left(-\alpha\sqrt{\frac{1+\alpha}{1-\alpha}}, -\sqrt{\frac{(1-2\alpha)(1+\alpha)}{1-\alpha}}, \alpha\right)^T \\ &= \left(\sqrt{1-\alpha^2}\cos\left(-\frac{4\pi}{5}\right), \sqrt{1-\alpha^2}\sin\left(-\frac{4\pi}{5}\right), \alpha\right)^T \end{aligned}$$

and

$$\begin{aligned} p_4 &= \left(\alpha\sqrt{\frac{1-\alpha}{1+\alpha}}, -\sqrt{\frac{(1+2\alpha)(1-\alpha)}{1+\alpha}}, \alpha\right)^T \\ &= \left(\sqrt{1-\alpha^2}\cos\left(-\frac{2\pi}{5}\right), \sqrt{1-\alpha^2}\sin\left(-\frac{2\pi}{5}\right), \alpha\right)^T \end{aligned}$$

Therefore, if

$$A = \begin{pmatrix} \cos(2\pi/5) & -\sin(2\pi/5) & 0 \\ \sin(2\pi/5) & \cos(2\pi/5) & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

and $p_0 = u_3$, then

$$A^k(p_0, p_1, p_2, p_3, p_4) = (p_{\sigma(0)}, p_{\sigma(1)}, p_{\sigma(2)}, p_{\sigma(3)}, p_{\sigma(4)})$$

where $\sigma(n) = n + k \mod 5$.

Now if we set $\beta = |\langle u_1, u_2\rangle| < \alpha$, then by changing the sign of u_1 if necessary and since $\|u_1\| = 1$, we may assume

$$u_1 = \left(\sqrt{1-\beta^2}\cos t_0, \sqrt{1-\beta^2}\sin t_0, \beta\right)^T,$$

for some fixed $t_0 \in [-\pi, \pi)$. Hence $|\langle u_1, u_3\rangle| < \alpha$

$$\begin{aligned}
&\iff \left|\sqrt{1-\alpha^2}\sqrt{1-\beta^2}\cos t_0 + \alpha\beta\right| < \alpha \\
&\iff \frac{-\alpha}{\sqrt{1-\alpha^2}}\frac{1+\beta}{\sqrt{1-\beta^2}} < \cos t_0 < \frac{\alpha}{\sqrt{1-\alpha^2}}\frac{1-\beta}{\sqrt{1-\beta^2}} \\
&\iff \underbrace{\cos^{-1}\left(\frac{1}{2}\sqrt{\frac{1-\beta}{1+\beta}}\right)}_{\gamma_1(\beta)} < |t_0| < \underbrace{\cos^{-1}\left(-\frac{1}{2}\sqrt{\frac{1+\beta}{1-\beta}}\right)}_{\gamma_2(\beta)}. \qquad (5.35)
\end{aligned}$$

We notice that

$$\gamma_1(\beta) = \begin{cases} \frac{6\pi}{15} = \frac{2\pi}{5}, & \text{if } \beta = \alpha, \\ \frac{5\pi}{15} = \frac{\pi}{3}, & \text{if } \beta = 0, \end{cases}$$

and

$$\gamma_2(\beta) = \begin{cases} \frac{12\pi}{15} = \frac{4\pi}{5}, & \text{if } \beta = \alpha, \\ \frac{10\pi}{15} = \frac{2\pi}{3}, & \text{if } \beta = 0, \end{cases}$$

and that $\frac{d}{d\beta}(\gamma_2 - \gamma_1)(\beta) > 0$ for $\beta \in (0, \alpha)$, see Figure 5.4. Thus $\frac{5\pi}{15} \le \gamma_2(\beta) - \gamma_1(\beta) < \frac{6\pi}{15}$, when $\beta \in [0, \alpha)$.

Now fix a $\beta \in [0, \alpha)$, then,

$$\gamma_2(\beta) < \gamma_1(\beta) + \frac{6\pi}{15},$$

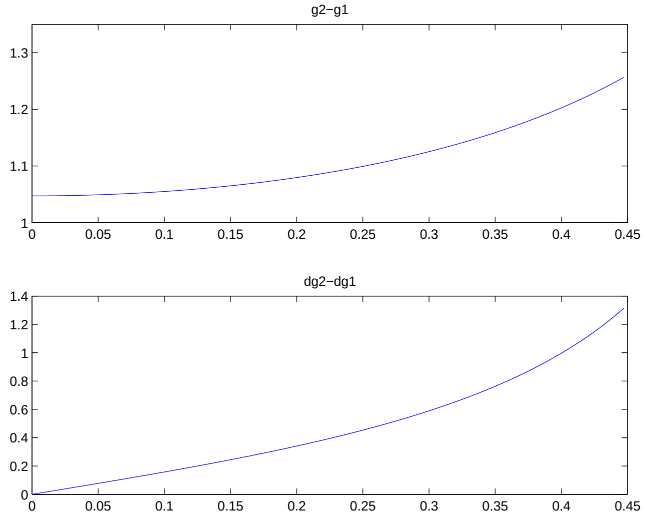


Figure 5.4: Top figure is the function $\gamma_2 - \gamma_1$, bottom figure is the function $\frac{d}{d\beta}(\gamma_2 - \gamma_1)$. We can see that the original function is strictly increasing.

and for $k = 1, 2, 3, 4$, we have

$$\alpha > |\langle u_1, p_k\rangle| = \left|\left\langle A^{-k}u_1, A^{-k}p_k\right\rangle\right| = \left|\left\langle A^{-k}u_1, p_0\right\rangle\right| = \left|\left\langle A^{-k}u_1, u_3\right\rangle\right|,$$

where

$$A^{-k}u_1 = \left(\sqrt{1-\beta^2}\cos\left(t_0 - \frac{2\pi k}{5}\right), \sqrt{1-\beta^2}\sin\left(t_0 - \frac{2\pi k}{5}\right), \beta\right)^T.$$

Therefore by (5.35),

$$\alpha > |\langle u_1, p_k\rangle| \iff \gamma_1(\beta) \leq \left|t_0 - \frac{2\pi k}{5}\right| \leq \gamma_2(\beta), \tag{5.36}$$

for $k = 0, 1, 2, 3, 4$. These inequalities define ten intervals on the torus $\mathbb{T}_{2\pi}$. If we plot these ten intervals on $\mathbb{T}_{2\pi}$, we see that no set of three of them overlap, see Figure 5.5.

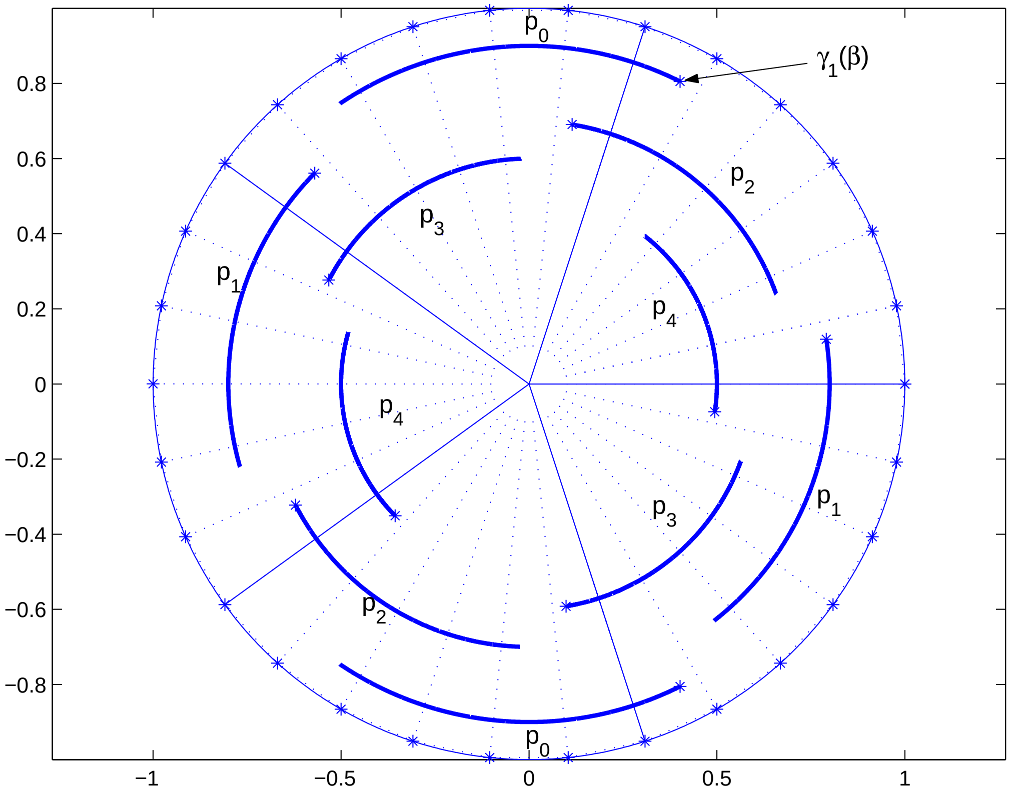


Figure 5.5: Ten intervals on $\mathbb{T}_{2\pi}$ coreesponding to the points $p_0 = u_3$, and $p_1, \ldots, p_4$.

This can also be seen since

$$\begin{aligned}\gamma_1(\beta) \le \left|t_0 - \frac{2\pi k}{5}\right| \le \gamma_2(\beta) \implies \gamma_1(\beta) \le \left|t_0 - \frac{2\pi k}{5}\right| < \gamma_1(\beta) + \frac{2\pi}{5}\\ \implies t_0 \in \underbrace{\big[\gamma + \varepsilon k, \gamma + \varepsilon(k+1)\big)}_{P_k} \cup \underbrace{\big[-\gamma + \varepsilon(k-1), -\gamma + \varepsilon k\big)}_{N_k}\end{aligned}$$

where $\gamma = \gamma_1(\beta)$, $\varepsilon = \frac{2\pi}{5}$ and $k = 0, 1, 2, 3, 4$. So $\cup_{k=0}^{4} P_k$ is a disjoint cover of $\mathbb{T}_{2\pi} \setminus [\gamma - \varepsilon, \gamma)$, and $\cup_{k=0}^{4} N_k$is a disjoint cover of $\mathbb{T}_{2\pi} \setminus [-\gamma + \varepsilon, -\gamma)$, so t_0 can be in at most two of the ten sets P_k, N_k.

Now by assumption, $|\langle u_1, u_3\rangle| = |\langle u_1, p_0\rangle| < \alpha$. Also $|\langle u_1, u_4\rangle| < \alpha$ and $|\langle u_1, u_5\rangle| < \alpha$ where $u_4, u_5 \in \{p_1, p_2, p_3, p_4\}$. Thus (5.36) implies t_0 lies in three of the ten intervals represented in Figure 5.5. A contradiction. Thus $\tilde{U}$ cannot be equiangular.

□

Finally, using Lemma 5.16 we have,

Theorem 5.17. *If $U \subset S^2 \subset \mathbb{R}^3$ is $(5, 3)$-Grassmannian, then $\mathcal{M}_\infty(U) = \frac{1}{\sqrt{5}}$.*

Proof. Let $\alpha = \mathcal{M}_\infty(U)$, and consider the graph whose vertices are $u_1, \dots, u_5$, and whose edges are defined as follows: for any pair of points $u_k, u_j \in U$ with $k \neq j$, an edge connects u_k and u_j if and only if $|\langle u_k, u_j\rangle| = \alpha$. We call the number of edges emanating from a vertex u_k, the *degree* of u_k, denoted $\deg(u_k)$. Then Lemma 5.16 implies that

$$\sum_{k=1}^{5} \deg(u_k) \ge \sum_{k=1}^{5} 3 = 15$$

but since each edge connects two vertices, the sum of the degrees must be an even number. Thus at least one vertex u_j must have degree 4, i.e., there is a

$j \in \{1, \dots, 5\}$, such that $|\langle u_j, u_{j_i} \rangle| = \alpha$ for $i = 1, \dots, 4$, where $\{j_1, j_2, j_3, j_4\} = \{1, 2, 3, 4, 5\} \setminus \{j\}$. By relabeling if necessary we may assume

$$|\langle u_1, u_k \rangle| = \alpha \text{ for } k = 2, 3, 4, 5,$$
$$|\langle u_2, u_k \rangle| = \alpha \text{ for } k = 3, 4.$$

Furthermore, we can reduce to the general position used in Theorem 5.10, i.e., assume

$$\begin{aligned} u_1 &= (0, 0, 1)^T \\ u_2 &= (\sqrt{1 - \alpha^2}, 0, \alpha)^T \\ u_3 &= (x_3, y_3, \alpha)^T \\ u_4 &= (x_4, y_4, \alpha)^T \\ u_5 &= (x_5, y_5, \alpha)^T. \end{aligned}$$

Now we have 2 cases:

Case 1. $|\langle u_3, u_4 \rangle| = \alpha$.

Then the subset $\tilde{U} = \{u_1, u_2, u_3, u_4\}$ is equiangular, hence Theorem 5.10 implies $\alpha = \frac{1}{3}$ or $\frac{1}{\sqrt{5}}$. But just as in Lemma 5.16, $\alpha = \frac{1}{3}$ implies that

$$\frac{1}{3} = \mathcal{M}_\infty(U) \leq \frac{1}{\sqrt{6}}.$$

So $\alpha = \frac{1}{\sqrt{5}}$.

Case 2. $|\langle u_3, u_4 \rangle| < \alpha$. Then since each vertex must be of degree 3, we have that $|\langle u_3, u_5 \rangle|$ and $|\langle u_4, u_5 \rangle|$ equal α. Thus if we remove the absolute values, we have the following four equations

$$\langle u_2, u_3 \rangle = \pm\alpha, \quad \langle u_2, u_4 \rangle = \pm\alpha, \quad \langle u_3, u_5 \rangle = \pm\alpha, \quad \langle u_4, u_5 \rangle = \pm\alpha.$$

This gives $2^4 = 16$ possible cases. Of these 16 cases, 7 lead to contradictions and the remaining 9 fall into 5 types but each implies that u_3, u_4, u_5 are three of the four points

$$\left(\alpha\sqrt{\frac{1-\alpha}{1+\alpha}}, \pm\sqrt{\frac{(1+2\alpha)(1-\alpha)}{1+\alpha}}, \alpha\right)^T,$$
$$\left(-\alpha\sqrt{\frac{1+\alpha}{1-\alpha}}, \pm\sqrt{\frac{(1-2\alpha)(1+\alpha)}{1-\alpha}}, \alpha\right)^T,$$

which are the positive endpoints on the remaining 4 diagonals of an icosahedron. Hence in each case, $\alpha = 1/\sqrt{5}$.

□

The $(5,3)$-Grassmannian frame is the first example of a non-optimal Grasmmannian frame since $\frac{1}{\sqrt{5}} > \frac{1}{\sqrt{6}}$. Hence, by Theorem 5.4, the $(5,3)$-Grassmannian frame is the first three dimensional example of a Grasmannian frame which is not tight.

5.5 $(6,3)$-Grassmannian frames

The $(6,3)$-Grassmannian bound can be calculated as a consequence of Theorem 5.4.

Corollary 5.18. *If $U = \{u_1, \dots, u_6\} \subset S^2$ is $(6,3)$-Grassmannian, then*

$$\mathcal{M}_\infty(U) = 1/\sqrt{5}.$$

Proof. Set $\alpha = \frac{1}{\sqrt{5}}$, and consider the set W with vertices

$$\begin{aligned}
w_1 &= (0,0,1)^T, \\
w_2 &= \left(\sqrt{1-\alpha^2}, 0, \alpha\right)^T, \\
w_3 &= \left(\alpha\sqrt{\frac{1-\alpha}{1+\alpha}}, \sqrt{\frac{(1+2\alpha)(1-\alpha)}{1+\alpha}}, \alpha\right)^T, \\
w_4 &= \left(\alpha\sqrt{\frac{1-\alpha}{1+\alpha}}, -\sqrt{\frac{(1+2\alpha)(1-\alpha)}{1+\alpha}}, \alpha\right)^T, \\
w_5 &= \left(-\alpha\sqrt{\frac{1+\alpha}{1-\alpha}}, \sqrt{\frac{(1-2\alpha)(1+\alpha)}{1-\alpha}}, \alpha\right)^T, \\
w_6 &= \left(-\alpha\sqrt{\frac{1+\alpha}{1-\alpha}}, -\sqrt{\frac{(1-2\alpha)(1+\alpha)}{1-\alpha}}, \alpha\right)^T.
\end{aligned}$$

Note that $\pm W$ are the twelve verticies of an icosahedron. Now for $k \neq l$, we compute that $|\langle w_k, w_l \rangle| = \frac{1}{\sqrt{5}}$. Furthermore, by Theorem 5.4, if U is a 6 element subset of S^2, then

$$\mathcal{M}_\infty(U) \geq \sqrt{\frac{6-3}{3(6-1)}} = \frac{1}{\sqrt{5}} = \mathcal{M}_\infty(W)$$

Thus W is a $(6,3)$-Grassmannian frame. □

Notice the $(6,3)$ Grassmanian arrangement is so good that when you remove a vector from it, it remains Grassmanian, and when we remove two vectors from it, it is still a local minimum of $\mathcal{M}_\infty$. Conway has found that there are other instances of this in higher dimensions, particularly when the symmetry group of the frame has a large number of elements.

5.6 Applications of Grassmanian frames to communication theory

Frames have found many applications in communication theory because of the natural redudancy and numerical stability of the frame reconstruction algorithm, [SH03, GKK01, CK03]. Grassmannian frames have the potential of reducing the losses associated with packet-based communication systems such as the internet. By packet-based, we mean a communication system which transmits packets of information of a certain length with the following error controling mechanism: if the packet contains errors, then it is not delivered, i.e., the packet is erased. This type of communication channel is called an erasure channel. For example, if $y \in \mathbb{R}^d$ represents the information to be transmitted, and if $X = \{x_k\}_{k=1}^N$ is a frame for $\mathbb{R}^d$. Then we send the coefficents $\{\langle y, x_k \rangle\}_{k=1}^N$ over the erasure channel. The erasures can then be modeled as erased frame coefficients, or erased frame elements. Thus we desire a frame with the property that if m elements are deleted, the remaining elements still form a frame for $\mathbb{R}^d$.

Using our classification of $(N, 2)$-Grassmannian frames, we now give a brief example to motivate why Grassmannian frames are amenable to erasure channel applications.

Consider the following two frames for $\mathbb{R}^2$,

$$X = \{(\pm 1, 0), (0, \pm 1)\}$$

and

$$Y = \left\{ \big(\cos(\pi k/N), \sin(\pi k/N) \big) : k = 0, 1, 2, 3 \right\},$$

see Figure 5.6. If exactly one of the elements of either X or Y is removed at random, then both X and Y remain a frame for $\mathbb{R}^2$, in this case we say that both

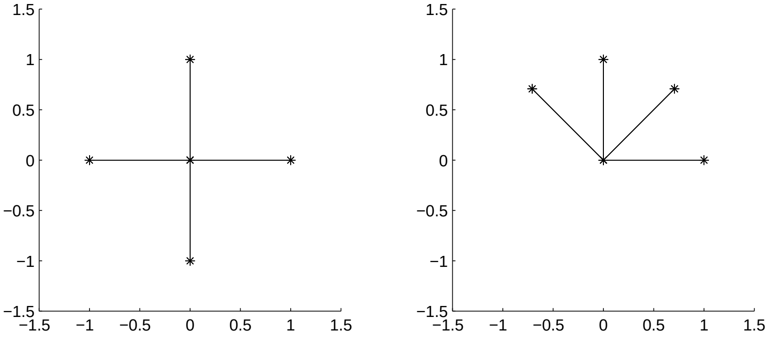


Figure 5.6: The frames X (left), and Y (right). To model an erasure channel remove two elements at random from X and Y. With X, the remaining elements may not span $\mathbb{R}^2$, but with Y, the remaining elements will still span $\mathbb{R}^2$.

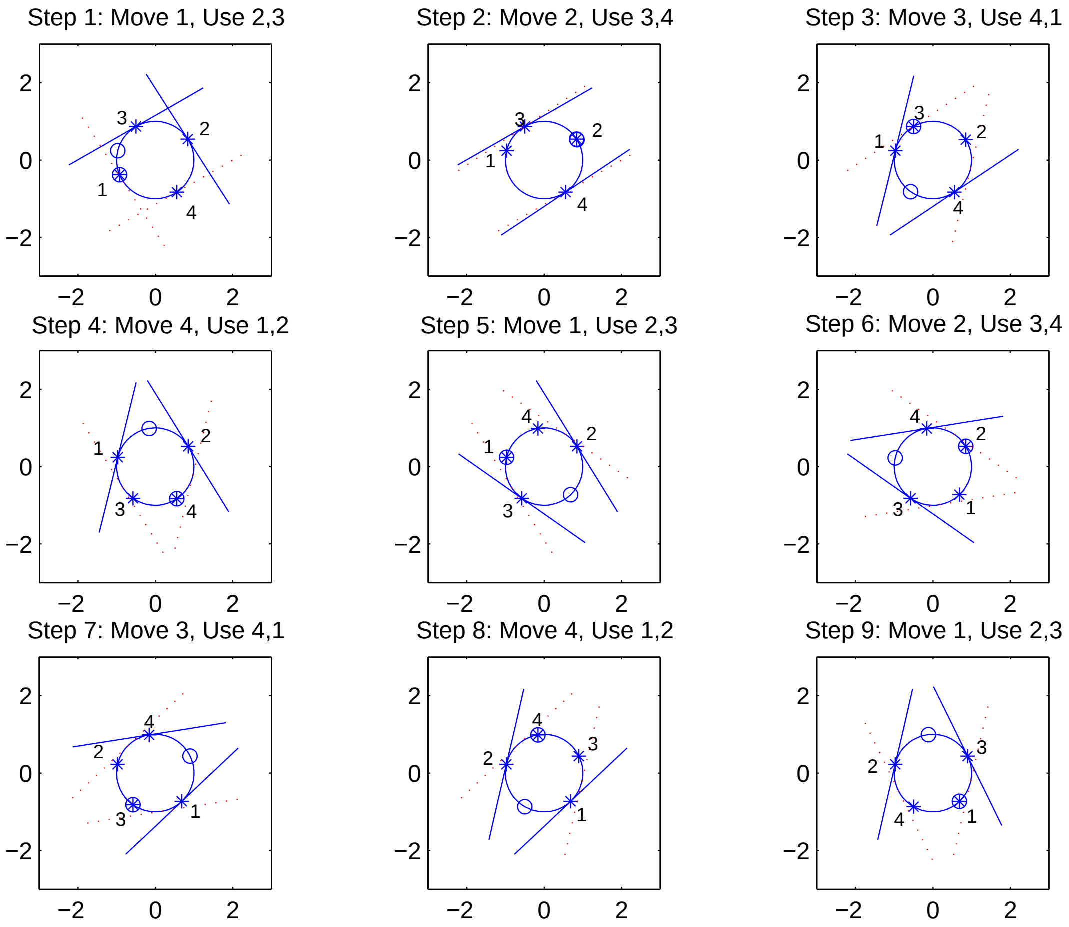
X and Y are robust to 1 erasure. Now if 2 elements are erased from Y, then we see that Y still remains a frames for $\mathbb{R}^2$ since the remaining vectors are not scalar multiples of each other. On the other hand, if both $(\pm 1, 0)$ is erased form X, then the remaining vectors only span the y-axis. Hence Y is robust to 2 erasures, whereas X is not.

5.7 Future research

Let $d \geq 2$, $N > d+1$, and $X = \{x_1, \ldots, x_N\} \subset S^{d-1}$. For $k = 1, \ldots, N$, set

$$Q_k = \left\{ v \in \mathbb{R}^d : |\langle v, y_l \rangle| \leq 1, \quad \text{for } l \in \{1, \ldots, N\} \setminus \{k\} \right\}$$

and let C_k be the set of extreme points. Also, let c_k be an element of C_k of maximal norm. Now, we consider the following replacement algorithm to reduce the ∞-correlation: as k cyclically ranges through the numbers $1, \ldots, N$, compute the new Q_k, C_k, and c_k, and set $y_k = c_k$. Hence after many interations of this replacement algorithm, since we are reducing the correlation at each step, we expect our algorithm to converge to a local minimum of the function $\mathcal{M}_\infty$. Determining the speed of convergence and the value of $\mathcal{M}_\infty$ on the limit set is a topic for future research. See Figure 5.7 for an example of this algorithm when $N = 4$ and $d = 2$.


Step 1: Move 1, Use 2,3

Step 2: Move 2, Use 3,4

Step 3: Move 3, Use 4,1

Step 4: Move 4, Use 1,2

Step 5: Move 1, Use 2,3

Step 6: Move 2, Use 3,4

Step 7: Move 3, Use 4,1

Step 8: Move 4, Use 1,2

Step 9: Move 1, Use 2,3

Figure 5.7: Applying the replacement algorithm cyclically to more than 3 points on S^1. The empty circle is the point to which we are moving the circled star.

BIBLIOGRAPHY

[ASVDS96] P. Aziz, H. Sorensen, and J. Van Der Spiegel, *An overview of sigma delta converters*, IEEE Sig. Proc. Mag. **13** (Jan. 1996), no. 1, 61–84.

[Ben97] J. J. Benedetto, *Harmonic Analysis and Applications*, CRC Press, Boca Raton, FL, 1997.

[BF94] John J. Benedetto and Michael W. Frazier (eds.), *Wavelets: Mathematics and Applications*, CRC Press, Boca Raton, FL, 1994. MR 94f:42048

[BF01] John J. Benedetto and Paulo J. S. G. Ferreira (eds.), *Modern Sampling Theory: Mathematics and Applications*, Birkhäuser Boston, Boston, MA, 2001.

[BF03] John J. Benedetto and Matthew Fickus, *Finite normalized tight frames*, Adv. Comput. Math. **18** (2003), no. 2–4, 357–385, Frames. MR 2004c:42059

[BPY] J.J. Benedetto, A. Powell, and Ö. Yılmaz, *Sigma-delta quantization and finite frames*, preprint.

[Chr02] Ole Christensen, *An Introduction to Frames and Riesz Bases*, Birkhäuser, Boston, MA, 2002.

[CHS96] John H. Conway, Ronald H. Hardin, and Neil J. A. Sloane, *Packing lines, planes, etc.: packings in Grassmannian spaces*, Experiment. Math. **5** (1996), no. 2, 139–159. MR 98a:52029

[CK03] Peter G. Casazza and Jelena Kovačević, *Equal-norm tight frames with erasures*, Adv. Comput. Math. **18** (2003), no. 2-4, 387–430, Frames. MR 1 968 127

[Dau92] I. Daubechies, *Ten Lectures on Wavelets*, SIAM, Philadelphia, PA, 1992.

[DD03] Ingrid Daubechies and Ron DeVore, *Approximating a bandlimited function using very coarsely quantized data: a family of stable sigma-delta modulators of arbitrary order*, Ann. of Math. (2) **158** (2003), no. 2, 679–710. MR 2 018 933

[DS52] R. J. Duffin and A. C. Schaeffer, *A class of nonharmonic Fourier series*, Trans. Amer. Math. Soc. **72** (1952), 341–366. MR 13,839a

[EB03] Yonina C. Eldar and Helmut Bölcskei, *Geometrically uniform frames*, IEEE Trans. Inform. Theory **49** (2003), no. 4, 993–1006. MR 1 984 484

[EFKM03] R. Eschbach, Z. Fan, K. Knox, and G. Marcu, *Threshold modulation and stability in error diffusion*, IEEE Sig. Proc. Mag. (2003), 39–50.

[For91] G. David Forney, Jr., *Geometrically uniform codes*, IEEE Trans. Inform. Theory **37** (1991), no. 5, 1241–1260. MR 92j:94018

[FS76] R. Floyd and L. Steinberg, *An adaptive algorithm for spatial greyscale*, Proc. Soc. Image Display **17** (1976), no. 2, 75–77.

[GB85] L. C. Grove and C. T. Benson, *Finite Reflection Groups, second edition*, Springer-Verlag, New York, NY, 1985.

[GKK01] Vivek K. Goyal, Jelena Kovačević, and Jonathan A. Kelner, *Quantized frame expansions with erasures*, Appl. Comput. Harmon. Anal. **10** (2001), no. 3, 203–233. MR 2002h:94012

[Gra90] Robert M. Gray, *Quantization noise spectra*, IEEE Trans. Inform. Theory **36** (1990), no. 6, 1220–1244.

[GVL83] G. Golub and C. Van Loan, *Matrix Computations*, Johns Hopkins University Press, Baltimore, MD, 1983.

[Kit98] T Kite, *Design and quality assessment of forward and inverse error diffusion halftoning algorithms*, Ph.D. thesis, The Univerity of Texas at Austin, 1998.

[KM40] M. Krein and D. Milman, *On extreme points of regular convex sets*, Studia Math. **9** (1940), 133–138. MR 3,90a

[Lay82] S. Lay, *Convex Sets and their Applications*, John Wiley and Sons, Inc., New York, NY, 1982.

[Lay03] D. Lay, *Linear Algebra and its Applications*, Addison Wesley, Boston, MA, 2003.

[LS73] P. W. H. Lemmens and J. J. Seidel, *Equiangular lines*, J. Algebra **24** (1973), 494–512. MR 46 #7084

[OS99] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*, Pearson Education, Inc., Upper Saddle River, NJ, 1999.

[Pan93] M. Panik, *Fundamentals of Convex Analysis: Duality, Seperation, Representation, and Resolution*, Kluwer Academic Publishers, Dordrecht, The Netherlands; Boston, 1993.

[Ros97] Moshe Rosenfeld, *In praise of the Gram matrix*, The Mathematics of Paul Erdős, II, Algorithms Combin., vol. 14, Springer, Berlin, 1997, pp. 318–323. MR 97i:52020

[SH03] Thomas Strohmer and Robert W. Heath, Jr., *Grassmannian frames with applications to coding and communication*, Appl. Comput. Harmon. Anal. **14** (2003), no. 3, 257–275. MR 2004d:42053

[Sle68] David Slepian, *Group codes for the Gaussian channel*, Bell System Tech. J. **47** (1968), 575–602. MR 38 #6879

[Str88] Gilbert Strang, *Linear Algebra and its Applications*, Harcourt Brace Jovanovich, Orlando, FL, 1988.

[Tot65] L. Fejes Toth, *Distribution of points in the elliptic plane*, Acta Math. Acad. Sci. Hungar. **16** (1965), 437–440.

[Uli87] R. Ulichney, *Digital Halftoning*, MIT Press, Cambridge, MA, 1987.

[VW] R. Vale and S. Waldron, *Tight frames and their symmetries*, preprint, submitted (accepted?) to Constr. Approx.

[Web94] R. Webster, *Convexity*, Oxford University Press, New York, NY, 1994.

[Wol84] J. A. Wolf, *Spaces of Constant Curvature, fifth edition*, Publish or Perish, Inc., Wilmington, DE, 1984.

[Yıl02] Ö. Yılmaz, *Mathematical properties of coarse quantization schemes in signal analysis with new applications*, Ph.D. thesis, Princeton University, 2002.