ABSTRACT

Title of dissertation: QUANTUM DETECTION AND FINITE FRAMES

Andrew Kei Kebo, Doctor of Philosophy, 2005

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

In quantum mechanics, the definition of a Von Neumann measurement can be generalized using positive-operator-valued measures. This modified definition of a quantum measurement allows one to better distinguish between a set of nonorthogonal quantum states. In this thesis we examine a quantum detection problem, where we have a physical system whose state is limited to be in only one of a finite number of possibilities. These possible states are not necessarily orthogonal. We want to find the best method of measuring the system in order to distinguish which state the system is in. Mathematically, we want to find a positive-operator-valued measure that minimizes the probability of a detection error.

It is shown that all tight-frames with frame constant 1 correspond to positive-operator-valued measures. We reformulate the problem in terms of tight-frames that minimize the detection error. In the finite dimensional case, the problem of finding the tight-frame that minimizes the error can be converted into a Hamiltonian system on the group $SO(N)$. The minimum energy solutions of this Hamiltonian system correspond exactly to the tight-frames that minimize the detection error. In this

setting, several numerical methods can be applied to give numerical constructions of the desired tight-frames.

QUANTUM DETECTION AND FINITE FRAMES

by

Andrew Kei Kebo

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
2005

Advisory Committee:

Professor John J. Benedetto, Chair/Advisor
Professor William Goldman
Professor Raymond L. Johnson
Professor Stephen Wallace
Professor Robert C. Warner

This dissertation is dedicated to my parents and my best friend Pil.

ACKNOWLEDGMENTS

I would like to thank my parents who have always given me support for all of my endeavors. They have consistently shown an interest in my work and on many occasions have been eager listeners to explanations of my research. Ironically, they still have no idea what I am doing.

I would also like to thank my close friends Pil Dela Cruz and Momo Otani for their encouragement, moral support, and numerous stimulating conversations of our respective studies. I owe my sanity to my wonderful housemates Andrew Dykstra, Aaron Lott, and Yue Xiao who have created a warm and friendly environment away from campus. In particular, I want to thank Andrew for the many evening mathematical discussions. He has the ability to present mathematics in a clear, elegant and beautiful way, which has influenced my own mathematical expositions. I thank Pol Tangboondouangjit for forcing me to really learn real analysis my first year of graduate school and for his help with my LaTeX needs, and Ben Howard for giving me a broad perspective of mathematics. I am grateful for the many frame-theoretic discussions with my colleagues Abdelkrim Bourouihiya, Joe Kolesar, Onur Oktay, and Ming Wei Ong. These are the rare people who know that I am doing something beyond addition and subtraction. I am also indebted to Anita Dahms, Haydeé Hidalgo, Alverda McCoy, and Millie Stengel who have always been helpful and made my interactions with the department extremely pleasant.

And last but not least, I would like to thank my advisor, Dr. John J. Benedetto, for his kindness, encouragement, extreme patience, and his broad knowledge of mathematics. His encouragement kept me in the department during a time when I was considering leaving with a masters. His broad knowledge made it possible to find an intersection in our mathematical interests and to find a problem that both of us could agree to work on. His insights allowed us to approach problems at many different angles and resulted in this thesis.

TABLE OF CONTENTS

List of Figures viii

1 Introduction and outline of thesis 1

2 Frame theory, linear algebra, and the spectral theorem 4

2.1 Frame theory 4

2.2 Finite frames 8

2.2.1 Naimark's theorem 13

2.3 Spectral theorem 17

2.4 The special orthogonal group 20

3 Quantum physics 22

3.1 Motivations of quantum mechanics 22

3.1.1 Light as waves 22

3.1.2 Light as particles 24

3.1.3 Matter has wave-like properties 25

3.1.4 Dynamical quantities as operators 26

3.1.5 Another look at momentum 28

3.1.6 General statistical interpretation 29

3.2 Example: a particle in a box 31

3.3 Generalization using Hilbert theory 33

3.4 Generalization to POMs 35

3.4.1 Example 1 36

3.5 Relationship between POMs and tight-frames 37

3.6 Why finite frames? 38

4 Quantum detection problem 40

4.1 Quantum communication 40

4.2 A closer look at the detection error 41

4.3 Using tight-frames to construct the POM 42

4.4 P_e for tight-frames and orthonormal sets 45

5 Classical mechanical interpretation 49

5.1 Newtonian mechanics of 1 particle 50

5.2 Lagrangian mechanics of N particles 51

5.3 Central force 54

5.4 Frame force 55

5.5 Physical interpretation of the frame problem 56

5.6 Hamiltonian system on $O(N)$ 58

5.7 Friction 64

5.8 Parameterization on $SO(N)$ 65

5.9 Examples 68

5.9.1 $N = 2$ 68

6 Least-squares error 72

6.1 Non-weight case 73

6.2 Weighted case 78

6.2.1 Linearly dependent case 80

6.3 Examples of computing the least-squares solution 96

6.3.1 Explicit example in $\mathbb{R}^2$ 97

6.3.2 Example of ϵ-modified vectors 99

6.4 Geometrically uniform frames 103

6.4.1 Examples of GU vector sets 104
6.4.2 Properties of GU frames and a second solution 105
6.4.3 Preliminaries 109
6.4.4 Change of notation 110
6.4.5 Fourier transform of functions on Q 111
6.5 Minimizers of P_e 113

A Penrose-Moore pseudo inverse 121

B Hilbert space definitions 122
B.1 Examples of Hilbert spaces 124

Bibliography 126

LIST OF FIGURES

3.1 The setup of Young's experiment, indicating that light is a wave. . . . 23

3.2 A diffraction pattern created by the constructive and destructive interference of light from two point sources. 23

3.3 A diagram of two infinite potential barriers creating a box. 31

3.4 A plot of the three lowest energy solutions of the Schrödinger equation of an electron in a box with impenetrable walls. 33

6.1 A GU set consisting of 2 vectors. 104

6.2 A GU set consisting of 3 vectors. 105

Chapter 1

Introduction and outline of thesis

In this thesis we study a quantum detection problem, which entails finding a quantum measurement that is optimal at distinguishing between a set of given nonorthogonal states. This can be reduced to the optimization problem of finding a 1-tight frame that minimizes a term that, in the context of quantum physics, represents the probability of a detection error. Mathematically, we have a d-dimensional Hilbert space H and a set $\{\psi_i\}_{i=1}^N \subset H$, $N \geq d$, with positive weights $\{\rho_i\}_{i=1}^N \subset \mathbb{R}$ that sum to 1. Our goal is to find a 1-tight frame $\{e_i\}_{i=1}^N$ for H that minimizes the term

$$P_e = 1 - \sum_{i=1}^{N} \rho_i |\langle \psi_i, e_i \rangle|^2.$$

We begin in chapter 2 with the mathematical background of the subject. Here, finite frame theory is introduced, followed by the spectral theorem, which is used to define quantum measurements in chapter 3. We end the chapter with a small discussion of the special orthogonal group, which is used in chapter 5 as a means of parameterizing tight frames and converting the quantum detection problem into a classical mechanics problem.

Chapter 3 is an introduction to quantum theory. The theory of quantum measurements is introduced via the spectral theorem, followed by a generalization of the theory with the use of positive-operator-valued measures. We prove that certain classes of positive-operator-valued measures correspond to tight frames, which makes

the application of frame theory to the theory of quantum measurement quite natural. This relation has been mentioned in [13] but was not proven.

Chapter 4 introduces the quantum detection problem. We use the relationship between positive-operator-valued measures and tight frames given in chapter 3 to restate the problem in terms of frames, followed by a proof that solutions to the quantum detection problem exists. We then use Naimark's theorem to further simplify and restate the problem in terms of orthonormal sets.

Chapter 5 begins with some background of classical mechanics, then proceeds with my main results. We give a method of parameterizing tight frames using the group $SO(N)$. A Hamiltonian on $SO(N)$ is introduced, and using Lagrangian mechanics we obtain corresponding equations of motion. We prove that the minimum energy solutions to the equations of motion exist and correspond to the tight frames that solve the quantum detection problem. Furthermore, friction terms can be added to the equations of motion so that solutions tend towards a minimum energy solution. This reformulation of the problem opens the possibility of using numerical methods to approximate the tight frames that solve the quantum detection problem.

Finally, in chapter 6 we examine a least-squares error term. Mathematically, we want to find a tight frame $\{e_i\}_{i=1}^N$ that minimizes the least-squares error

$$E = \sum_{i=1}^{N} \rho_i \|\psi_i - e_i\|^2.$$

When the given set $\{\psi_i\}_{i=1}^N$ are linearly independent, explicit constructions for $\{e_i\}_{I=1}^N$ are given that minimize the least-squares error [13, 12]. We expand this work by examining the case when $\{\psi_i\}_{i=1}^N$ are linearly dependent and produce con-

structive methods of producing tight frames with a small error E and provide bounds for E. We end the chapter with a presentation of the work done in [12]. They show that if the weights $\{\rho_i\}_{i=1}^N$ are all equal and if the given set $\{\psi_i\}_{i=1}^N$ are geometrically uniform, then the frame that minimizes the least-squares error E also minimizes the detection error P_e of the quantum detection problem.

Chapter 2

Frame theory, linear algebra, and the spectral theorem

This chapter presents the mathematical background for studying quantum detection. We start with a short introduction to frame theory [6, 8, 22] followed by Naimark's theorem [7, 13] which relates frames with orthonormal sets. We then present the spectral theorem [30, 27] which plays a central role in the theory of quantum measurement in chapter 3. We end the chapter with a discussion of the group $SO(N)$ and its parameterization with $N(N-1)/2$ real variables. This parameterization is used in chapter 5 to create a Hamiltonian system on $SO(N)$ that corresponds to the quantum detection problem.

2.1 Frame theory

A frame is a generalization of an orthonormal basis. Let H be a separable Hilbert space, $K \subseteq \mathbb{Z}$, and $\{e_i\}_{i\in K}$ an orthonormal basis for H. An orthonormal basis has the property that for all $x \in H$

$$\|x\|^2 = \sum_{i\in K} |\langle x, e_i\rangle|^2.$$

We use this property to motivate the definition of a frame.

Definition 2.1. Let H be a separable Hilbert space and $K \subset \mathbb{Z}$. A set $\{e_i\}_{i\in K} \subset H$

is a *frame* with *frame bounds* $A, B \in \mathbb{R}$ if $0 < A < B$ and for all $x \in H$ we have

$$A\|x\|^2 \leq \sum_{i \in K} |\langle x, e_i \rangle|^2 \leq B\|x\|^2.$$

A frame $\{e_i\}_{i \in K}$ is a *tight frame* if $A = B$, and in this case the constant A is called the *frame constant*. We refer to a tight frame with frame constant A as an A-tight frame.

Let H be a separable Hilbert space and $\{e_i\}_{i \in K}$ a frame for H. Define the *Bessel map* $L : H \to l^2(K)$ defined for all $x \in H$ by

$$Lx = \{\langle x, e_i \rangle\}_{i \in K}.$$

By the definition of a frame, it is easy to see that L is continuous. Consider the adjoint $L^* : l^2(K) \to H$. Let $x \in H$ and $\{c_i\}_{i \in K} \in l^2(K)$. Then

$$\begin{aligned} \langle x, L^*(\{c_i\}_{i \in K}) \rangle &= \langle Lx, \{c_i\}_{i \in K} \rangle = \langle \{\langle x, e_i \rangle\}_{i \in K}, \{c_i\}_{i \in K} \rangle \\ &= \sum_{i \in K} \langle x, e_i \rangle c_i = \left\langle x, \sum_{i \in K} c_i e_i \right\rangle. \end{aligned}$$

Since this is true for all $x \in H$, given any $\{c_i\}_{i \in K} \in l^2(K)$,

$$L^*(\{c_i\}_{i \in K}) = \sum_{i \in K} c_i e_i \in H.$$

Define the *frame operator* $S = L^*L$. It is not hard to show that for any $x \in H$,

$$S(x) = \sum_{i \in K} \langle x, e_i \rangle e_i.$$

Like an orthonormal basis, all elements of H can be written as a linear combination of the frame elements. It can be shown that for all $x \in H$,

$$x = \sum_{i \in K} \langle x, e_i \rangle S^{-1} e_i = \sum_{i \in K} \langle x, S^{-1} e_i \rangle e_i.$$

In fact, if $\{e_i\}_{i\in K}$ is a tight frame with frame constant A, then the reconstruction formula is much simpler. We show that for all $x \in H$ we have

$$x = \frac{1}{A}Sx = \frac{1}{A}\sum_{i\in K}\langle x, e_i\rangle e_i.$$

Theorem 2.1. *Let H be a separable Hilbert space and $\{e_i\}_{i\in K} \subset H$. $\{e_i\}_{i\in K}$ is a tight frame with frame constant A if and only if the frame operator satisfies*

$$S = AI$$

where I is the identity operator on H.

Proof. First, assume that $S = L^*L = AI$. Then given any $x \in H$ we have

$$\begin{aligned} A\|x\|^2 &= A\langle x, x\rangle = \langle Ax, x\rangle = \langle Sx, x\rangle \\ &= \langle L^*Lx, x\rangle = \langle Lx, Lx\rangle \\ &= \|Ly\|^2_{l^2(K)} \\ &= \sum_{i\in K}|\langle x, e_i\rangle|^2, \end{aligned}$$

hence we see that $\{e_i\}_{i\in K}$ is a tight frame for H with frame constant A.

Now assume that $\{e_i\}_{i\in K}$ is a tight frame for H with frame constant A. Then for all $x \in H$,

$$A\langle x, x\rangle = A\|x\|^2 = \sum_{i\in K}|\langle x, e_i\rangle|^2 = \sum_{i\in K}\langle x, e_i\rangle\langle e_i, x\rangle = \left\langle \sum_{i\in K}\langle x, e_i\rangle e_i, x \right\rangle = \langle Sx, x\rangle$$

hence for all $x \in H$

$$\langle (S - AI)x, x\rangle = 0.$$

Note that the operator $S - AI$ is self-adjoint and positive semidefinite. By Theorem 19 from [30], given any $x, y \in H$ we have

$$|\langle (S - AI)x, y\rangle| \leq \sqrt{\langle (S - AI)x, x\rangle \langle (S - AI)y, y\rangle} = 0$$

so it follows that $(S - AI) = 0$ and the result follows. □

Under certain conditions, tight frames coincide with orthonormal bases. For example, if $\{e_i\}_{i\in K}$ is a tight frame consisting of unit normed vectors and has frame constant 1, then it must be an orthonormal basis.

Theorem 2.2. *Let H be a separable Hilbert space. Let $K \subset \mathbb{Z}$ and $\{e_i\}_{i\in K} \subset H$ be a normalized set of vectors. Then $\{e_i\}_{i\in K}$ is a tight frame for H with frame constant 1 if and only if $\{e_i\}_{i\in K}$ is an orthonormal basis for H.*

Proof. Assume that $\{e_i\}_{i\in K}$ is a tight frame for H with frame constant 1. Then given any $y \in H$ we have

$$\|y\|^2 = \sum_{i\in K} |\langle y, e_i\rangle|^2.$$

Since $\{e_i\}_{i\in K}$ are normalized, for each $i \in K$ we have

$$\begin{aligned} 1 &= \|e_i\|^2 = \sum_{k\in K} |\langle e_i, e_k\rangle|^2 \\ &= \|\langle e_i, e_i\rangle|^2 + \sum_{k\in K, k\neq i} |\langle e_i, e_k\rangle|^2 \\ &= 1 + \sum_{k\in K, k\neq i} |\langle e_i, e_k\rangle|^2, \end{aligned}$$

hence we must have

$$\sum_{k\neq i} |\langle e_i, e_k\rangle|^2 = 0.$$

So for $i \neq k$, $\langle e_i, e_k \rangle = 0$, hence $\{e_i\}_{i \in K}$ is an orthonormal set for H. By the previous theorem, the frame operator $S = I$ so for any $y \in H$ we have

$$y = Sy = \sum_{i \in K} \langle y, e_i \rangle e_i$$

so $\{e_i\}_{i \in K}$ is complete, and $\{e_i\}_{i \in K}$ is an orthonormal basis for H.

Conversely, if $\{e_i\}_{i \in K}$ is an orthnormal basis, it is clear that it is a normalized tight frame with frame constant 1. □

2.2 Finite frames

In this section, we consider the case where $H = \mathbb{K}^d$ where $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$ and frames consisting of a finite set of elements of H, that is frames of the form $\{e_i\}_{i=1}^N \subset H$ where N is an integer with $N > d$.

Many of the operators associated with finite frames have a matrix representation. Denote by $\mathcal{M}(m \times n)$ as the set of all $m \times n$ matrices. For example, the Bessel map can be written as a matrix. Let $\{e_i\}_{i=1}^N \subset H$ be a frame for H and $\{b_i\}_{i=1}^d$ an orthonormal basis. The Bessel map $L : H \to l^2(\mathbb{Z}_N)$ can be written as a matrix $L \in \mathcal{M}(N \times d)$ with components $L_{ij} = \langle b_j, e_i \rangle$, i.e,

$$L = \begin{pmatrix} — & e_1^* & — \\ & \vdots & \\ — & e_N^* & — \end{pmatrix}$$

with respect to the basis $\{b_i\}_{i=1}^d$ of H, and where $*$ denotes the conjugate transpose.

To verify this matrix corresponds to the Bessel map, given any $x \in H$ we can write

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix}$$

with respect to the basis $\{b_i\}_{i=1}^d$ and hence

$$L(x) = \begin{pmatrix} — & e_1^* & — \\ & \vdots & \\ — & e_N^* & — \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix} = \begin{pmatrix} \langle x, e_1 \rangle \\ \vdots \\ \langle x, e_N \rangle \end{pmatrix} \in l^2(\mathbb{Z}_N).$$

The conjugate Bessel map $L^* : l^2(\mathbb{Z}_N) \to H$ can also be written as a matrix $L^* \in \mathcal{M}(d \times N)$ with components $L^*_{ij} = \overline{\langle b_i, e_j \rangle}$, i.e. it is of the form

$$L^* = \begin{pmatrix} | & & | \\ e_1 & \dots & e_N \\ | & & | \end{pmatrix}.$$

Let $x \in H$ and denote the ith component of e_k by $e_k(i)$. Then the frame operator $S : H \to H$, defined by $S = L^*L$, can be written as

$$\begin{aligned} Sx &= L^*Lx = \begin{pmatrix} | & & | \\ e_1 & \dots & e_N \\ | & & | \end{pmatrix} \begin{pmatrix} \langle x, e_1 \rangle \\ \vdots \\ \langle x, e_N \rangle \end{pmatrix} \\ &= \begin{pmatrix} \sum_{i=1}^N \langle x, e_i \rangle e_i(1) \\ \vdots \\ \sum_{i=1}^N \langle x, e_i \rangle e_i(d) \end{pmatrix} = \sum_{i=1}^N \langle x, e_i \rangle e_i \end{aligned}$$

Which is the expression for S we obtained earlier.

In the finite dimensional case, it can be shown that any spanning set of vectors is a frame. We first prove a lemma involving the operator S for a given spanning set.

Lemma 2.1. *Let H be a d-dimensional Hilbert space and $\{e_i\}_{i=1}^N \subset H$ such that $span\{e_i\}_{i=1}^N = H$. Then the operator $S : H \to H$ defined for all $x \in H$ by*

$$Sx = \sum_{i=1}^{N} \langle x, e_i \rangle e_i$$

is positive-definite.

Proof. Let $x \in H$ such that $x \neq 0$. Then by definition

$$Sx = \sum_{i=1}^{N} \langle x, e_i \rangle e_i.$$

Taking the inner product with x on both sides gives us

$$\langle Sx, x \rangle = \sum_{i=1}^{N} \langle x, e_i \rangle \langle e_i, x \rangle = \sum_{i=1}^{N} |\langle x, e_i \rangle|^2.$$

Since $x \in H = span\{e_i\}_{i=1}^N$ it follows that $\langle x, e_i \rangle \neq 0$ for some i. Suppose not, that is suppose $\langle x, e_i \rangle = 0$ for all $1 \leq i \leq N$. Since $H = span\{e_i\}_{i=1}^N$, for any basis $\{b_i\}_{i=1}^d$ for H we can write

$$b_i = \sum_{j=1}^{N} A_j^{(i)} e_j$$

for some constants $\{A_j^{(i)} : 1 \leq i \leq d, 1 \leq j \leq N\}$. Hence for all $1 \leq i \leq d$,

$$\langle x, b_i \rangle = \left\langle x, \sum_{j=1}^{N} A_j^{(i)} e_j \right\rangle = \sum_{j=1}^{N} A_j^{(i)} \langle x, e_j \rangle = 0$$

so it follows that $x = 0$, which is a contradiction since we assumed that $x \neq 0$.

So,

$$\langle Sx, x\rangle = \sum_{i=1}^{N} |\langle x, e_i\rangle|^2 > 0$$

and S is positive definite. □

Theorem 2.3. *Let H be a d-dimensional Hilbert space and $\{e_i\}_{i=1}^N \subset H$. $\{e_i\}_{i=1}^N$ is a frame for H if and only if $H = span\{e_i\}_{i=1}^N$.*

Proof. Assume that $span\{e_i\}_{i=1}^N = H$. By Lemma 2.1, we know that the operator $S = L^*L$ is positive-definite and Hermitian, so it follows that S has d positive real eigenvalues $\{\lambda_i\}_{i=1}^d$. So, for all $x \in H$ we have

$$A\|x\|^2 \leq \langle Sx, x\rangle \leq B\|x\|^2$$

where $A = \min\{\lambda_i : 1 \leq i \leq d\}$ and $B = \max_i\{\lambda_i : 1 \leq i \leq d\}$, hence

$$A\|x\|^2 \leq \sum_{i=1}^{N} |\langle x, e_i\rangle|^2 \leq B\|x\|^2.$$

By definition, $\{e_i\}_{i=1}^N$ is a frame for H.

Conversely, assume that $\{e_i\}_{i=1}^N$ is a frame for H. Suppose that $x \in H$ and $x \in \left(span\{e_i\}_{i=1}^N\right)^\perp$. Then

$$A\|x\|^2 \leq \sum_{i} |\langle x, e_i\rangle|^2 = 0$$

so $\|x\| = 0$ hence $x = 0$. So

$$\left(span\{e_i\}_{i=1}^N\right)^\perp = \{0\}$$

and we have

$$H = span\{e_i\}_{i=1}^N + \left(span\{e_i\}_{i=1}^N\right)^\perp = span\{e_i\}_{i=1}^N.$$

□

Here, we give a condition where tight frames coincide with orthogonal bases.

Theorem 2.4. *Let H be a d-dimensional Hilbert space and assume that $\{e_i\}_{i=1}^d$ is a tight frame for H with frame constant A. Then $\{e_i\}_{i=1}^d$ is a $\sqrt{A}$-normed orthogonal set.*

Proof. First, since $\{e_i\}_{i=1}^d$ is a frame in an d-dimensional space, $\{e_i\}_{i=1}^d$ must be linearly independent. Suppose not, that is suppose $\{e_i\}_{i=1}^d$ is linearly dependent. Then there exists constants $\{c_i\}_{i=1}^d \in \mathbb{C}$ not all zero such that

$$\sum_{i=1}^{d} c_i e_i = 0.$$

In particular, there is some $1 \leq j \leq d$ such that $c_j \neq 0$. Then

$$e_j = -\frac{1}{c_j} \sum_{i \neq j} c_i e_i.$$

Hence, for any $x \in H$ we can write,

$$x = \frac{1}{A} \sum_{i=1}^{d} \langle x, e_i \rangle e_i = \frac{1}{A} \sum_{i \neq j} \left(\langle x, e_i \rangle - \frac{c_i}{c_j} \right) e_i$$

so $H = span\{e_i\}_{i \neq j}$ so $dim(H) \leq d - 1$ which contradicts the fact that H is d-dimensional.

Let $1 \leq k \leq d$. Then,

$$e_k = \frac{1}{A} \sum_{i=1}^{d} \langle e_k, e_i \rangle e_i = \frac{1}{A} \|e_k\|^2 e_k + \frac{1}{A} \sum_{i \neq k} \langle e_k, e_i \rangle e_i$$

hence subtracting e_k on both sides gives us,

$$\left[\frac{1}{A} \|e_k\|^2 - 1 \right] e_k + \frac{1}{A} \sum_{i \neq k} \langle e_k, e_i \rangle e_i = 0.$$

Since the $\{e_i\}_{i=1}^d$ are linearly independent, we must have,

$$\|e_k\| = \sqrt{A} \text{ and } \langle e_k, e_i \rangle = 0 \; i \neq k$$

and the result follows. □

2.2.1 Naimark's theorem

In this section we present Naimark's theorem. As a preliminary needed to prove Naimark's theorem, we present the singular value theorem without proof. This theorem generalizes the process of diagonalizing matrices in cases where the matrix is not necessarily square. This decomposition is valid for arbitrary matrices. This theorem is also used in chapter 6 where a least-squares quantum detection problem is examined.

Theorem 2.5. *(Singular Value Decomposition)*
Given any $A \in \mathcal{M}(m \times n)$ there exists matrices $U \in \mathcal{M}(m \times m)$, $V \in \mathcal{M}(n \times n)$ and $\Sigma \in \mathcal{M}(m \times n)$ where U, V are unitary and the diagonal components $\sigma_i = \Sigma_{ii}$ are positive and real, and $\Sigma_{ij} = 0$ if $i \neq j$, such that $A = U\Sigma V^*$. The components σ_i are called the singular values of A.

The proof of this theorem is constructive. U is the unitary matrix whose columns are the orthonormal eigenvectors of the self-adjoint matrix AA^*, V is the unitary matrix whose columns are the orthonormal eigenvectors of the self-adjoint matrix A^*A, and the singular values σ_i are the positive square root of the nonnegative eigenvalues of AA^*.

We now state some observations that shall be useful in the proof of Naimark's theorem, as well as in chapters 4 and 6.

1. We show another form of the singular value decomposition that becomes useful in the proof of Naimark's theorem. If $A \in \mathcal{M}(m \times n)$ is of rank r, we can reorder the indices i in the singular value decomposition such that the first r singular values $\{\sigma_i\}_{i=1}^{r}$ are nonzero. Then with this reordering, it is not hard to show that its singular value decomposition can be written as

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^*$$

where u_i is the ith column of U, v_i is the ith column of V, and σ_i are the nonzero diagonal elements of Σ.

2. Here we give explicit expressions for orthogonal projections. Let H be a d-dimensional Hilbert space and let W be a finite N-dimensional subspace of $\mathcal{H}$ where $N \leq d$. Let $\{w_i\}_{i=1}^{N}$ be an orthonormal basis for W. Then we can express the orthogonal projection $P : H \to W$ as

$$P = \sum_{i=1}^{n} w_i w_i^*,$$

which is not hard to see since for any $x \in H$ we have

$$P(x) = \sum_{i=1}^{n} w_i w_i^* x = \sum_{i=1}^{n} \langle x, w_i \rangle w_i \in W.$$

3. Let $\{e_i\}_{i=1}^{N}$ be a tight frame for H with frame constant A. Then for all $x \in H$ we have

$$Sx = \sum_{i=1}^{N} \langle x, e_i \rangle e_i$$

hence we can write

$$S = \sum_{i=1}^{N} e_i e_i^*.$$

Also, by Theorem 2.1 it follows that $\{e_i\}_{i=1}^N$ is a tight frame if and only if

$$\sum_{i=1}^{N} e_i e_i^* = AI.$$

Naimark's theorem [13] relates tight frames with equal-normed orthogonal sets. All tight frames can be considered as projections of an equal-normed orthogonal set where the orthogonal set exists in a larger Hilbert space. This theorem is crucial for the construction of the Hamiltonian system on $SO(N)$ given in chapter 2.

Theorem 2.6. *(Naimark's Theorem) Let H be a d-dimensional Hilbert space. Let $\{e_i\}_{i=1}^N$ be a tight frame for H with frame constant A. Then there exists an orthogonal A-normed set $\{\widetilde{e}_i\}_{i=1}^N \subset \widetilde{H}$, where $\widetilde{H}$ is a N-dimensional Hilbert space such that H is a linear subspace of $\widetilde{H}$, and*

$$P_H \widetilde{e}_i = e_i$$

where P_H is the orthogonal projection onto H.

Proof. Let $\{b_i\}_{i=1}^d$ be an orthonormal basis for H. Define $M \in \mathcal{M}(N \times d)$ as the Bessel map matrix corresponding to the vector set $\{e_i\}_{i=1}^N$ with respect to the basis $\{b_i\}_{i=1}^d$, that is M is of the form

$$M = \begin{pmatrix} — & e_1^* & — \\ & \vdots & \\ — & e_N^* & — \end{pmatrix}.$$

It suffices to show that there exists a matrix $\widetilde{M} \in \mathcal{M}(N \times N)$ such that

1. $\widetilde{M}\widetilde{M}^* = A^2 I_N$

2. $P_H \widetilde{M}^* = M^*$.

$\widetilde{M}\widetilde{M}^*$ is the matrix whose entries are the inner products of the columns of $\widetilde{M}^*$. Hence, 1. shows that the columns of $\widetilde{M}^*$ are orthogonal and have norm A. 2. shows that the projection of the columns of $\widetilde{M}^*$ onto H gives the original tight frame $\{e_i\}_{i=1}^N$. So the A-normed orthogonal set we are looking for are just the columns of $\widetilde{M}^*$.

We take the singular value decomposition of M^* to get

$$M^* = U\Sigma V^* = \sum_{i=1}^{d} \sigma_i u_i v_i^*.$$

We will show that the singular values $\{\sigma_i\}_{i=1}^d$ are all equal. Since each vector u_i is d dimensional, and $\{u_i\}_{i=1}^d$ are orthonormal, since they are the columns of a unitary matrix, we have $span\{u_i\}_{i=1}^d = H$. Since M is the Bessel map for a tight frame, by Theorem 2.1 and observation 2 we have for the corresponding frame operator,

$$S = M^*M = A^2 I_H = A^2 \sum_{i=1}^{d} u_i u_i^*.$$

We can also write, using the singular value decomposition of M^* and the fact that $\{v_i\}_{i=1}^N$ is an orthonormal set,

$$\begin{aligned} M^*M &= \sum_{i=1}^{d} \sigma_i u_i v_i^* \sum_{k=1}^{d} \sigma_k v_k u_k^* \\ &= \sum_{i,k=1}^{d} \sigma_i \sigma_k u_i \langle v_i, v_k \rangle u_k^* = \sum_{i=1}^{d} \sigma_i^2 u_i u_i^* \end{aligned}$$

hence we see that for all $i = 1, \ldots, d$, we must have $\sigma_i = A$. So

$$M^* = A \sum_{i=1}^{d} u_i v_i^*.$$

Now consider an enlarged N-dimensional Hilbert space $\widetilde{H}$ such that $H \subset \widetilde{H}$. We can find a set of $N-d$ vectors $\{u_i\}_{i=d+1}^N \subset \widetilde{H}$ such that $\{u_i\}_{i=1}^N$ is an orthonormal basis for $\widetilde{H}$. Define

$$\widetilde{M^*} = A \sum_{i=1}^N u_i v_i^*.$$

Then $\widetilde{M}$ has the desired properties we want. We have,

$$\begin{aligned} \widetilde{M}\widetilde{M^*} &= A^2 \sum_{i=1}^N v_i u_i^* \sum_{k=1}^N u_k v_k^* = A^2 \sum_{k,i=1}^N v_i \langle u_i, u_k \rangle v_k^* \\ &= A^2 \sum_{i=1}^N v_i v_i^* = A^2 I_N. \end{aligned}$$

Since $\{u_i\}_{i=1}^d$ is an orthonormal basis for H, we can write $P_H = \sum_{j=1}^d u_j u_j^*$ hence we have,

$$\begin{aligned} P_H \widetilde{M^*} &= \sum_{j=1}^d u_j u_j^* A \sum_{i=1}^N u_i v_i^* = A \sum_{j=1}^d \sum_{i=1}^N u_j \langle u_j, u_i \rangle v_i^* \\ &= A \sum_{j=1}^d u_j v_j^* = M^*. \end{aligned}$$

□

For the case where H is infinite dimensional, see [7].

2.3 Spectral theorem

All self-adjoint $N \times N$ matrices can be diagonalized. If A is an $N \times N$ self-adjoint matrix, then A has N real eigenvalues $\{\lambda_i\}_{i=1}^N$, counting multiplicities, and N corresponding orthonormal eigenvectors $\{v_i\}_{i=1}^N$ such that

$$A = \sum_{i=1}^N \lambda_i v_i v_i^*.$$

Each $v_i v_i^*$ can be considered as an orthogonal projection onto the 1-dimensional space spanned by v_i. The spectral theorem is a generalization of this idea when A

is a self-adjoint operator on a separable Hilbert space H. We replace the sum of the projections by a resolution of the identity.

Definition 2.2. Let $\mathcal{B}$ be a σ-algebra of sets of X and H a Hilbert space. Denote by $\mathcal{L}(H)$ as the collection of all bounded linear operators on H. A mapping $E : \mathcal{B} \to \mathcal{L}(H)$ is a *resolution of the identity* if:

1. $E(\emptyset) = 0$, $E(X) = I$.
2. For all $w \in \mathcal{B}$, $E(w)$ is a orthogonal projection.
3. For all $w, w' \in \mathcal{B}$, $E(w \cap w') = E(w)E(w')$.
4. If $w \cap w' = \emptyset$ then $W(w \cup w') = E(w) + E(w')$.
5. for all $x, y \in H$, $\langle E(\cdot)x, y\rangle$ is a complex measure on $\mathcal{B}$.

With this definition, we can present the spectral theorem [27].

Theorem 2.7. *Let T be a bounded normal operator on a separable Hilbert space H. Then there exists a unique resolution of the identity such that for all $x, y \in H$,*

$$\langle Tx, y\rangle = \int \lambda d\langle E(\lambda)x, y\rangle.$$

As an abuse of notation, we sometimes write

$$T = \int \lambda dE(\lambda).$$

For those familiar with quantum mechanics, many times we consider the Hilbert space $H = L^2(\mathbb{R})$ and the self-adjoint operators O_x and O_p defined on a dense subset of H for some $f \in H$ by

$$O_x f(x) = xf(x),\ O_p f(x) = \frac{\hbar}{i}\frac{df}{dx}(x).$$

Note that these operators are not bounded, nor are they defined on all of H. However, the spectral theorem can be modified to apply to all self-adjoint operators [27, 30].

Definition 2.3. Let H be a separable Hilbert space and A a self-adjoint operator defined on a dense subset of H. The *Cayley transform* U of A is the operator that satisfies for all $f \in Dom(A)$

$$U(Af + if) = Af - if.$$

If A is self-adjoint, then it is shown by [27] that the domain and range of U satisfies

$$\text{Dom}(U) = \text{Range}(U) = H$$

and U is a unitary operator on H.

Theorem 2.8. *Let T be a self-adjoint operator defined on a dense subset of a separable Hilbert space H. Then there exists a unique resolution of the identity such that*

$$T = \int \lambda dE(\lambda).$$

We present a sketch of the proof. Let U be the Cayley transform of T. Since

$$\text{Dom}(U) = \text{Range}(U) = H$$

one can show that U is a unitary operator on H. By the Spectral theorem, there exists a unique resolution of the identity E such that

$$U = \int \sigma dE(\sigma).$$

Using the change of variables,

$$\lambda = -\cot \pi\sigma$$

one can show that

$$T = \int \lambda dE \left(-\frac{1}{\pi} \cot^{-1} \lambda \right)$$

gives a unique resolution of the identity for T.

2.4 The special orthogonal group

Finally, we present some facts of the orthogonal and special orthogonal group [26]. We later use the fact that $O(N)$ is a smooth manifold in chapter 5 to develop a method of parameterizing orthonormal sets in an N-dimensional Hilbert space.

Let N be a positive integer. The orthogonal group is defined by

$$O(N) = \{A \in \mathcal{M}(N \times N) : A^\tau A = I\}$$

where τ denotes the transpose of the matrix A, and I is the $N \times N$ identity matrix. We will mainly be considering the special orthogonal group given by

$$SO(N) = \{A \in O(N) : \det(A) = 1\}.$$

It can be shown that $SO(N)$ is a $N(N-1)/2$-dimensional manifold. To show this, we construct a smooth map from $\mathbb{R}^{N(N-1)/2}$ into $SO(N)$.

Define the exponential map $\exp : \mathcal{M}(N \times N) \to \mathcal{M}(N \times N)$ for all $X \in \mathcal{M}(N \times N)$ by

$$\exp(X) = \sum_{n=0}^{\infty} \frac{1}{n!} X^n.$$

where we define $X^0 = I$. It is not hard to show that if $A, B \in \mathcal{M}(N \times N)$ commute, then

$$\exp(A)\exp(B) = \exp(A + B).$$

Consider the space of all $N \times N$ antisymmetric matrices

$$\mathcal{A}(N) = \{A \in \mathcal{M}(N \times N) : A^\tau = -A\}.$$

$\mathcal{A}(N)$ is a $N(N-1)/2$-dimensional real linear space under matrix addition. Note that for any $A \in \mathcal{A}(N)$,

$$\exp(A)^\tau = \sum_{n=1}^{\infty} \frac{1}{n!}(A^n)^\tau = \sum_{n=0}^{\infty} \frac{1}{n!}(A^\tau)^n = \sum_{n=0}^{\infty} \frac{1}{n!}(-A)^n = \exp(-A)$$

hence

$$\exp(A)^\tau \exp(A) = \exp(-A + A) = \exp(0) = I$$

so $\exp(A) \in SO(N)$.

Let $\{A_1, \ldots, A_{N(N-1)/2}\} \subset \mathcal{A}(N)$ be a basis for $\mathcal{A}(N)$. Define the map $f : \mathbb{R}^{N(N-1)/2} \to SO(N)$ for all $(q_1, \ldots, q_{N(N-1)/2}) \in \mathbb{R}^{N(N-1)/2}$ by

$$f(q_1, \ldots, q_{N(N-1)/2}) = \exp\left(\sum_{i=1}^{N(N-1)/2} q_i A_i\right) \in SO(N).$$

It can be shown that f is a diffeomorphism taking a neighborhood of 0 in $\mathbb{R}^{N(N-1)/2}$ into a neighborhood of I in $SO(N)$. Given any $B \in SO(N)$, one can further show that $Bf : \mathbb{R}^{N(N-2)/2} \to SO(N)$ is a diffeomorphism taking a neighborhood of 0 in $\mathbb{R}^{N(N-1)/2}$ into a neighborhood of B in $SO(N)$. Hence, $SO(N)$ is a smooth $N(N-1)/2$-dimensional manifold.

$O(N)$ is also a smooth $N(N-1)/2$-dimensional manifold with two components, the component with positive determinant corresponds to $SO(N)$.

Chapter 3

Quantum physics

This chapter presents some of the history and motivations of quantum theory. The historical facts in sections 3.1.1, 3.1.2, and 3.1.3 were taken from [16] and the basic theory and examples in sections 3.1.4, 3.1.5, 3.1.6, and 3.2 were taken from [15]. The general quantum theory in terms of Hilbert spaces, section 3.3, was borrowed from [30] and the introduction of POM measurements in section 3.4 can be found in [13, 12, 31, 18] with physical realizations of POM measurements given by [5, 4]. The relationship between POMs and tight frames given in section 3.5 has been done by the author.

3.1 Motivations of quantum mechanics

3.1.1 Light as waves

Ever since the 1600's, there was a debate whether light is a wave or a particle. In 1801, Thomas Young performed the double-slit experiment. This consisted of a monochromatic light source and a screen separated by two barriers, the one on the left with one slit and the one on the right with two slits. See the figure below for an illustration.

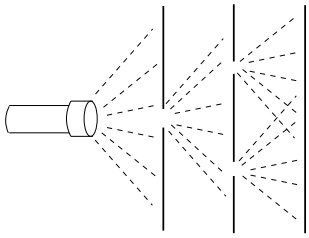

Figure 3.1: The setup of Young's experiment, indicating that light is a wave.

The light spreads by diffraction after passing through the first barrier. The light emanating from the single slit acts as a point source of light. The light then passes through the second barrier and through the two slits, creating two point sources of light. On the screen, on the far right in figure 3.1, the light from the two point sources create a pattern of bright and dark patches as in figure 3.2 below.

Figure 3.2: A diffraction pattern created by the constructive and destructive interference of light from two point sources.

This was strong evidence that light behaves like a wave, and the pattern on the screen was due to constructive and destructive interference of the light waves emanating from the two point sources. In fact, in the mid 1800's it was discovered that light is an electromagnetic wave, an oscillating electric and magnetic field governed by Maxwell's equations. In empty space, the waves are governed by the following wave equations

$$\nabla^2 E = \frac{1}{c^2}\frac{\partial^2 E}{\partial t^2}$$

$$\nabla^2 B = \frac{1}{c^2}\frac{\partial^2 B}{\partial t^2}$$

where c is the speed of light, t corresponds to time, and E and B are the electric and magnetic vectors fields respectively, both functions of space and time. The intensity of the light on the screen in Young's experiment can be shown to be proportional to $|E|^2$.

3.1.2 Light as particles

It has been found that if light of a certain wavelength λ is directed on a piece of metal, it knocks off electrons from the metal. It is as if light, something with no mass, has a momentum and can physically push a particle with mass. Einstein was able to explain this effect by modeling light as a stream of particles, called photons, which have a momentum given by the expression

$$p = \frac{h}{\lambda}$$

where h is Plank's constant. In fact, in 1921 Einstein won the Nobel prize for his analysis of the photoelectric effect. With this view that light is nothing but a stream of particles, the intensity $|E|^2$ of the light on the screen in Young's experiment can be viewed as a probability distribution of where the photons are likely to hit the screen. The bright spots on the screen represent areas where the photon is likely to hit.

3.1.3 Matter has wave-like properties

If light has both wave-like and particle-like properties then perhaps matter does as well, that is perhaps matter also has wave-like properties. Momentum was attributed to a photon using the expression $p = h/\lambda$. In 1924, Louis de Broglie suggested that this same expression might be used to assign particles of mass a wavelength

$$\lambda = \frac{h}{p}.$$

In fact, in 1989 it was shown that using a beam of electrons instead of light in the double-slit experiment also produced interference patterns on the screen, which was consistent with de Broglie's expression for the wavelength of matter [29].

Since the wave-like properties of light are governed by Maxwell's wave equations, then it seems natural to have a wave equation that models the wave-like properties of matter. In 1926, Schrödinger produced the Schrödinger equation to do just that,

$$i\hbar\frac{\partial\Psi}{\partial t} = -\frac{\hbar^2}{2m}\nabla^2\Psi + V\Psi$$

where $\hbar$ is Plank's constant, m is the mass of the particle being modelled, V is a real scalar field, and Ψ is a complex field, sometimes referred to as the *state of the system*. Like the photon interpretation of Young's experiment, $|\Psi|^2$ is interpreted as the probability distribution of the location of the massive particle in question. Since $|\Psi|^2$ is a probability distribution, we require the conditions $\Psi \in L^2$ and

$$\int |\Psi|^2 = 1.$$

3.1.4 Dynamical quantities as operators

Suppose we have a particle in a 1-dimensional space, that is $\Psi(x,t)$ is a function of one space variable x and time t. Since $|\Psi(x,t)|^2$ is the probability distribution of the location of the particle at time t, the expectation value of the position is given by

$$\mathcal{E}(x) = \int_{\mathbb{R}} x|\Psi(x,t)|^2 dx = \langle \Psi(x,t), x\Psi(x,t)\rangle$$

where the above bracket denotes the usual $L^2(\mathbb{R})$ inner-product with respect to the variable x. Since momentum is defined by $p = mv$, it is quite natural to define the expectation value for momentum as

$$\mathcal{E}(p) := m\frac{d}{dt}\mathcal{E}(x).$$

We obtain,

$$\begin{aligned}\frac{d}{dt}\mathcal{E}(x) &= \frac{d}{dt}\int_{\mathbb{R}} x|\Psi(x,t)|^2 dx = \int_{\mathbb{R}} x\frac{\partial}{\partial t}|\Psi(x,t)|^2 dx \\ &= \int_{\mathbb{R}} x\left(\Psi^*(x,t)\frac{\partial}{\partial t}\Psi(x,t) + \Psi(x,t)\frac{\partial}{\partial t}\Psi^*(x,t)\right)dx.\end{aligned}$$

Since Ψ must satisfy the Schrödinger equation, we have

$$\begin{aligned}\frac{\partial \Psi}{\partial t} &= \frac{i\hbar}{2m}\frac{\partial^2\Psi}{\partial x^2} - \frac{i}{\hbar}V\Psi \\ \frac{\partial \Psi^*}{\partial t} &= -\frac{i\hbar}{2m}\frac{\partial^2\Psi^*}{\partial x^2} + \frac{i}{\hbar}V\Psi^*.\end{aligned}$$

Plugging this back into the equation above gives us

$$\begin{aligned}\frac{d}{dt}\mathcal{E}(x) &= \int_{\mathbb{R}} x\left(\Psi^*(x,t)\left[\frac{i\hbar}{2m}\frac{\partial^2\Psi}{\partial x^2} - \frac{i}{\hbar}V\Psi\right] + \Psi(x,t)\left[-\frac{i\hbar}{2m}\frac{\partial^2\Psi^*}{\partial x^2} + \frac{i}{\hbar}V\Psi^*\right]\right)dx \\ &= \frac{i\hbar}{2m}\int_{\mathbb{R}} x\left(\Psi^*(x,t)\frac{\partial^2\Psi}{\partial x^2} - \Psi(x,t)\frac{\partial^2\Psi^*}{\partial x^2}\right)dx\end{aligned}$$

$$
\begin{aligned}
&= \frac{i\hbar}{2m}\int_{\mathbb{R}} x\frac{\partial}{\partial x}\left(\Psi^*(x,t)\frac{\partial\Psi}{\partial x} - \Psi(x,t)\frac{\partial\Psi^*}{\partial x}\right)dx \\
&= -\frac{i\hbar}{2m}\int_{\mathbb{R}} \left(\Psi^*(x,t)\frac{\partial\Psi}{\partial x} - \Psi(x,t)\frac{\partial\Psi^*}{\partial x}\right)dx \text{ integration by parts} \\
&= -\frac{i\hbar}{2m}\int_{\mathbb{R}} \left(\Psi^*(x,t)\frac{\partial\Psi}{\partial x} + \Psi^*(x,t)\frac{\partial\Psi}{\partial x}\right)dx \text{ integration by parts} \\
&= -\frac{i\hbar}{m}\int_{\mathbb{R}} \Psi^*(x,t)\frac{\partial\Psi}{\partial x}dx = \left\langle \Psi(x,t), \frac{\hbar}{im}\frac{\partial}{\partial x}\Psi(x,t)\right\rangle.
\end{aligned}
$$

So,

$$\mathcal{E}(p) = \left\langle \Psi, \frac{\hbar}{i}\frac{\partial}{\partial x}\Psi\right\rangle.$$

It seems natural to define self-adjoint position and momentum operators defined on a dense subset of $L^2(\mathbb{R})$ by

$$
\begin{aligned}
O_x f(x) &= xf(x) \\
O_p f(x) &= \frac{\hbar}{i}\frac{\partial}{\partial x}f(x).
\end{aligned}
$$

Then we can write the expectation values of position and momentum as

$$
\begin{aligned}
\mathcal{E}(x) &= \langle \Psi, O_x\Psi\rangle \\
\mathcal{E}(p) &= \langle \Psi, O_p\Psi\rangle.
\end{aligned}
$$

Since any dynamical variable w can be written in terms of position and momentum $w(x,p)$, we can get a corresponding self-adjoint operator O_w by substituting, in the expression for $w(x,p)$, the momentum p with the operator O_p and substituting the position x with the operator O_x, that is $O_w = w(O_x, O_p)$. If we have a particle in the state $\Psi \in H$, the expectation value for the quantity w is given by

$$\mathcal{E}(w) = \langle \Psi, O_w\Psi\rangle.$$

For example, energy is the sum of kinetic and potential energy

$$E = \frac{1}{2}mv^2 + V(x) = \frac{1}{2m}p^2 + V(x)$$

so its corresponding operator, known as the Hamiltonian, is

$$O_E = \frac{1}{2m}O_p^2 + V(O_x) = -\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2} + V(x).$$

Note that this is the right hand side of the Schrödinger equation, so we can write

$$i\hbar\frac{\partial \Psi}{\partial t} = O_E\Psi.$$

The expectation value of the energy is given by

$$\mathcal{E}(E) = \langle \Psi, O_E\Psi \rangle.$$

3.1.5 Another look at momentum

Suppose $\Psi(x,t)$ is a unit-normed solution of Schrödinger's equation, that is for each fixed $t \in \mathbb{R}$, $\int_{\mathbb{R}} |\Psi(x,t)|^2 dx = 1$. We want to find the average spatial frequency $\overline{\gamma}$ of Ψ. Since the Fourier transform of a function gives information about its frequencies, a natural definition of the average frequency of Ψ for a given t would be

$$\overline{\gamma} = \int_{\mathbb{R}} \gamma |\hat{\Psi}(\gamma,t)|^2 d\gamma$$

where $\hat{\Psi}(\gamma,t)$ is the Fourier transform of $\Psi(x,t)$ with respect to the spatial variable, that is

$$\hat{\Psi}(\gamma,t) = \int_{-\infty}^{\infty} \Psi(x,t)e^{-2\pi i x\gamma} dx.$$

By the de Broglie relation,

$$p = \frac{h}{\lambda}$$

and the relation between wavelength λ and frequency γ

$$\gamma = \frac{2\pi}{\lambda}$$

we have

$$\gamma = \frac{2\pi}{h}p = \frac{1}{\hbar}p,$$

where $\hbar = h/2\pi$. So the momentum p is proportional to the frequency γ and $\mathcal{E}(p)$ should be proportional to $\overline{\gamma}$ with proportionality constant $\hbar$. We have

$$\begin{aligned}
\mathcal{E}(p) &= \hbar\overline{\gamma} = \hbar \int_{\mathbb{R}} \gamma |\hat{\Psi}(\gamma, t)|^2 d\gamma \\
&= \langle \hat{\Psi}(\gamma, t), \hbar\gamma\hat{\Psi}(\gamma, t) \rangle \\
&= \left\langle \Psi(x,t), \frac{\hbar}{i}\frac{d}{dx}\Psi(x,t) \right\rangle \text{ by Parseval.}
\end{aligned}$$

So by the de Broglie relation, it again seems natural to define the momentum operator as

$$O_p = \frac{\hbar}{i}\frac{d}{dx}.$$

3.1.6 General statistical interpretation

Self-adjoint operators can always be decomposed by the Spectral Theorem. If H is a separable Hilbert space, and A a self-adjoint operator, then there exists a resolution of the identity E such that

$$A = \int_{\mathbb{R}} \lambda dE(\lambda).$$

Using this, we can define what it means to take functions of self-adjoint operators. Given a dE measurable function $f : \mathbb{R} \mapsto \mathbb{R}$ we define the operator $f(A)$ by

$$f(A) := \int_{\mathbb{R}} f(\lambda) dE(\lambda).$$

It can be shown that $f(A)$ is also a self-adjoint operator on H, see pages 143-145 of [30].

Suppose we have an electron in a 1-dimensional space with corresponding solution to the Schrödinger equation $\Psi(x,t)$ and we want to know when the values of its momentum lie in the interval $I=[a,b]$. The corresponding dynamical variable can be written as a function of momentum p given by

$$f(p)=\begin{cases} 1 & \text{if } p\in I \\ \\ 0 & \text{if } p\notin I. \end{cases}$$

Suppose O_p is the corresponding operator for momentum and E its corresponding resolution of the identity.

Note that the probability that the momentum of the electron lies in the region I is the same as the expectation value of $f(p)$. So we have

$$\begin{aligned} P(I) &= \mathcal{E}(f(O_p))=\langle\Psi,f(O_p)\Psi\rangle=\int_{\mathbb{R}} f(\lambda)d\langle\Psi,E(\lambda)\Psi\rangle \\ &= \int_a^b d\langle\Psi,E(\lambda)\Psi\rangle=\langle\Psi,E(I)\Psi\rangle. \end{aligned}$$

This probabilistic expression is general and works for other dynamical quantities, not just momentum. Suppose the state of the system is $\Psi\in H$. Given any dynamical quantity w with its corresponding self-adjoint operator O_w and resolution of the identity E, the probability that after measuring w, the outcome lines in a region $I=[a,b]$ is given by

$$P(I)=\langle\Psi,E(I)\Psi\rangle.$$

3.2 Example: a particle in a box

Suppose we have an electron in a 1-dimensional box of length $a > 0$. Suppose further that the walls of the box are inpenetrable. The physicists describe the potential as

$$V(x) = \begin{cases} 0 & \text{if } x \in [0, a] \\ \infty & \text{if } x \notin [0, a] \end{cases}$$

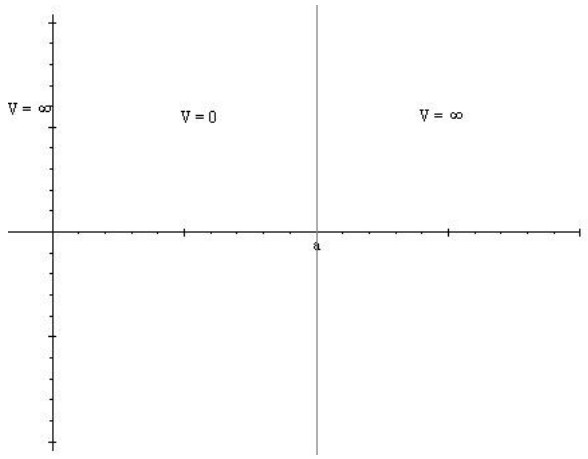

Figure 3.3: A diagram of two infinite potential barriers creating a box.

We consider the closed interval $I = [0, a]$ and solve the Schrödinger equation on I with the boundary conditions $\Psi(0, t) = \Psi(a, t) = 0$. Inside the box, the potential $V = 0$ so we solve

$$i\hbar \frac{\partial \Psi}{\partial t} = -\frac{\hbar^2}{2m} \frac{\partial^2 \Psi}{\partial x^2}.$$

To simplify the Schrödinger equation, we use separation of variables and assume we can decompose the solution as $\Psi(x, t) = \psi(x)\theta(t)$. Plugging this into the Schrödinger equation we obtain

$$i\hbar \psi(x) \frac{\partial \theta}{\partial t}(t) = -\theta(t) \frac{\hbar^2}{2m} \frac{\partial^2 \psi}{\partial x^2}(x)$$

and for all x and t such that $\psi(x) \neq 0$ and $\theta(t) \neq 0$ we can write

$$\frac{i\hbar}{\theta(t)}\frac{\partial\theta}{\partial t}(t) = -\frac{1}{\psi(x)}\frac{\hbar^2}{2m}\frac{\partial^2\psi}{\partial x^2}.$$

So, there must be a constant E such that

$$\frac{i\hbar}{\theta(t)}\frac{\partial\theta}{\partial t}(t) = E \quad \Rightarrow \quad i\hbar\frac{\partial\theta}{\partial t}(t) = E\theta(t) \tag{3.1}$$

$$-\frac{1}{\psi(x)}\frac{\hbar^2}{2m}\frac{\partial^2\psi}{\partial x^2} = E \quad \Rightarrow \quad -\frac{\hbar^2}{2m}\frac{\partial^2\psi}{\partial x^2} = E\psi(x). \tag{3.2}$$

Solving (3.1) gives us

$$\theta(t) = e^{-i\frac{E}{\hbar}t}.$$

The general solution of (3.2) is

$$\psi(x) = A\cos(\kappa x) + B\sin(\kappa x) \text{ where } \kappa = \sqrt{\frac{2mE}{\hbar^2}}.$$

Imposing the boundary conditions gives us

$$\psi(0) = 0 \quad \Rightarrow \quad A = 0$$

$$\psi(a) = 0 \quad \Rightarrow \quad \kappa = \frac{\pi n}{a},\ n \in \mathbb{Z}.$$

The second boundary condition tells us the possible values of the constant E

$$E = \frac{n^2\pi^2\hbar^2}{2ma^2}.$$

Since Ψ is a probability function we require

$$\begin{aligned}\int_0^a |\Psi(x,t)|^2 dx &= \int_0^a |\theta(t)\psi(x)|^2 dx = \int_0^a B^2\sin^2(\kappa x)dx \\ &= \frac{B^2}{2}\left[\frac{-\cos(\kappa x)\sin(\kappa x) + \kappa x}{\kappa}\right]_0^a = B^2\frac{a}{2} = 1\end{aligned}$$

so

$$B = \sqrt{\frac{2}{a}}.$$

So our solution to the Schrödinger equation is

$$\psi_n(x) = \sqrt{\frac{2}{a}} \sin\left(\frac{n\pi}{a}x\right), \; E_n = \frac{n^2\pi^2\hbar^2}{2ma^2}, \; n \in \mathbb{Z}. \tag{3}$$

The first three solutions are plotted below.

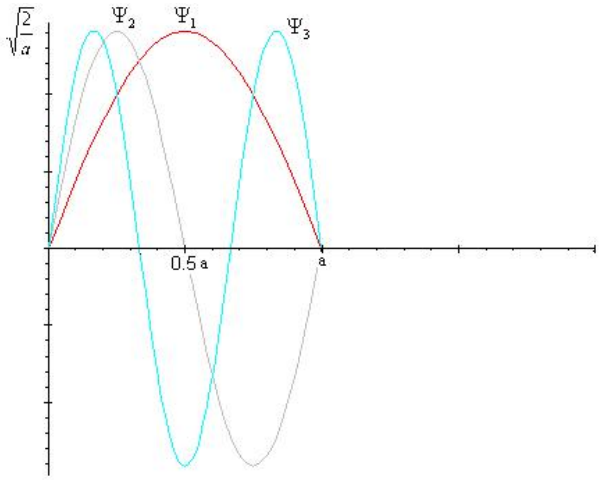

Figure 3.4: A plot of the three lowest energy solutions of the Schrödinger equation of an electron in a box with impenetrable walls.

Since equation (3.2) can be written in terms of the energy operator

$$O_E\psi(x) = E\psi(x)$$

we see that the constants E_n are the eigenvalues of O_E, hence (3.3) gives the different wave-functions corresponding to different energy levels.

3.3 Generalization using Hilbert theory

Since for a fixed t, $|\Psi(x,t)|^2$ is a probability distribution of a particle, if there are no boundaries imposed of where the particle can go then $\Psi \in L^2(\mathbb{R})$, and if the particle

is in a box then $\Psi \in L^2([0,a])$. For each physical system considered, we sometimes use a different Hilbert space H. So it seems natural to generalize the theory to a theory involving general separable Hilbert spaces.

Another motivation for this generalization is that in 1925, before Schrödinger's wave mechanics, Heisenberg developed a matrix theory of quantum mechanics. According to Von Neumann [30], the two theories can be shown to be equivalent. Their apparent difference is due to the fact that they use different representations of essentially the same Hilbert space. So the essence of the theory involves Hilbert spaces.

The theory is as follows:

1. A physical system is modeled using a separable Hilbert space H and a self-adjoint linear operator $\mathcal{H}$. $\mathcal{H}$ is called the Hamiltonian and is related to the total energy of the system.

2. The state of the system is given by a unit-normed element $\Psi \in H$. Ψ evolves in time by the Schrödinger equation

$$i\hbar \frac{d}{dt}\Psi = \mathcal{H}\Psi.$$

3. All dynamical variables w are represented by self-adjoint linear operators O_w on H. Suppose E is the corresponding resolution of the identity of O_w. If the state of the system is given by $\Psi \in H$, then the expectation value of w is given by

$$\mathcal{E}(w) = \langle \Psi, O_w \Psi \rangle$$

and the probability distribution of w is given by the measure

$$\langle \Psi, E(\cdot)\Psi \rangle.$$

With this generalization, physicists have modeled other physical systems with Hilbert spaces other than those that naturally appear when using Schrödinger's original wave mechanics. An example is the Stern-Gerlach experiment, where a beam of electrons are subjected to a inhomogeneous magnetic field [15]. In this situation the Hilbert space $H = \mathbb{C}^2$ is used.

3.4 Generalization to POMs

We now generalize the definition of a quantum measurement. As before, every possible measurable quantity corresponds to a self-adjoint operator on H. However, if instead we only require the corresponding resolution of the identity E to be a family of positive self-adjoint linear operators and not necessarily orthogonal projections, then E is said to be a *positive-operator-valued measure*, or POM for short. To denote the distinction between a POM and a resolution of the identity, we use the symbol Π instead of E. The formal definition of a POM is as follows.

Definition 3.1. Let $\mathcal{B}$ be a σ-algebra of sets of X. A *positive operator-valued measure* (POM) is a function $\Pi : \mathcal{B} \to \mathcal{L}(H)$ such that:

1. $\forall U \in \mathcal{B},\ \Pi(U)$ is a positive self-adjoint operator,
2. $\Pi(\emptyset) = 0$ (zero operator),
3. $\forall$ disjoint $\{U_i\}_{i=1}^{\infty} \subset \mathcal{B},\ x, y \in H \Rightarrow \left\langle \Pi\left(\bigcup_{i=1}^{\infty} U_i\right) x, y \right\rangle = \sum_{i=1}^{\infty} \langle \Pi(U_i)x, y \rangle,$

4. $\Pi(X) = I$ (identity operator).

We think of X as the space of all possible outcomes. X might be countable or uncountable. For example, suppose we wanted to measure the energy of a hydrogen atom. The energy levels of a hydrogen atom are discrete and X would consist of all the possible discrete energy levels, hence X is countable. On the other hand, if we were measuring the position of the electron orbiting the nucleus, then X would be the space of all possible spatial locations of the electron, i.e. $X = \mathbb{R}^3$ which is uncountable.

Every dynamical quantity in quantum mechanics corresponds to a space of outcomes X and a POM Π. If the state of the system is given by $\psi \in H$ with $\|\psi\| = 1$, then the probability that the measured outcome lies in a region $U \subset X$ is given by

$$P(U) = \langle \psi, \Pi(U)\psi \rangle.$$

3.4.1 Example 1

Consider the Hilbert space $H = L^2(\mathbb{R}^3)$ and suppose the state of a particle is given by $\psi \in L^2(\mathbb{R}^3)$ with $\|\psi\| = 1$. Suppose we are interested in measuring position. Then the space of outcomes is just $X = \mathbb{R}^3$. Given a set $U \in \mathcal{B}$ the position POM is given by $\Pi(U) = \mathbb{1}_U$, i.e. point-wise multiplication by $\mathbb{1}_U$ where $\mathbb{1}_U$ is the characteristic function of U. So the probability that the particle is found in a region $U \subset \mathbb{R}^3$ is given by

$$P(U) = \langle \psi, \Pi(U)\psi \rangle = \int_U |\psi|^2.$$

Using a POM in place of a resolution of the identity enriches the subject of quantum communications. There are many reasons for its use. To name a few, in some situations using a POM measurement decreases the likelihood of making a measurement error [24]. Also, the foundation of quantum encryption where messages cannot be intercepted by an eavesdropper is based on the theory of POM measurements [2].

3.5 Relationship between POMs and tight-frames

The theory of tight-frames can be used to construct POMs. Let H be a separable Hilbert space and $K \subset \mathbb{Z}$. Assume $\{e_i\}_{i\in K} \subset H$ is a 1-tight-frame for H. Define a family of self-adjoint positive operators for all $w \subset K$ and $x \in H$ by

$$\Pi(w)x = \sum_{i\in w} \langle x, e_i\rangle e_i.$$

It is clear that this family of operators satisfy conditions 1-3 of the definition of a POM. Since $\{e_i\}_{i\in K}$ is a 1-tight-frame, we also have for all $x \in H$,

$$\Pi(K)x = \sum_{i\in K} \langle x, e_i\rangle e_i = x$$

so condition 4 is satisfied and Π, constructed in this manner, is a POM.

Conversely, we have the following theorem.

Theorem 3.1. *Let H be a d-dimensional Hilbert space. Given a POM Π with a countable set X, there exists a subset $K \subset \mathbb{Z}$, a 1-tight-frame $\{e_i\}_{i\in K}$ for H, and a disjoint partition $\{B_i\}_{i\in X} \subset \mathcal{B}$ of K such that for all $i \in X$ and $x \in H$,*

$$\Pi(i)x = \sum_{j\in B_i} \langle x, e_j\rangle e_j.$$

Proof. For each $i \in X$, $\Pi(i)$ is self-adjoint and positive by definition, so by the spectral theorem there exists an orthonormal set $\{v_j\}_{j\in B_i}$ and positive numbers $\{\lambda_j\}_{j\in B_i}$ such that for all $x \in H$,

$$\Pi(i)x = \sum_{j\in B_i} \lambda_j \langle x, v_j\rangle v_j = \sum_{j\in B_i} \langle x, e_j\rangle e_j$$

where for all $j \in B_i$,

$$e_j = \sqrt{\lambda_j} v_j.$$

Since $\Pi(X) = I$ we have for all $x \in H$,

$$x = \Pi(X)x = \sum_{j\in \cup_i B_i} \langle x, e_j\rangle e_j.$$

It follows that $\{e_j\}_{j\in K}$ is a 1-tight-frame for H. □

So if our Hilbert space H is finite-dimensional, analyzing quantum measurements with a discrete set of outcomes X reduces to analyzing tight-frames.

3.6 Why finite frames?

In chapters 5 and 6, we focus on analyzing the quantum detection problem using finite frames. In the theory of quantum computing and quantum encryption, finite-dimensional Hilbert spaces are used. For example, quantum computers store information using qubits, which correspond to a finite-dimensional complex Hilbert space. Some of the physical realizations of these qubits are the spin directions of an electron, or the polarization directions of photons [25].

The application of the quantum detection problem is to be able to transmit and receive information encoded through a quantum channel. Some justifications of

using finite frames for the quantum detection problem are:

1. We only need a finite alphabet to transfer information. An infinite alphabet is not necessary.

2. If quantum detection is applied to the areas of quantum computing or quantum encryption, then finite-dimensional Hilbert spaces are used hence finite frames are sufficient.

Chapter 4

Quantum detection problem

In this chapter, sections 4.1 and 4.2, we present a quantum detection problem and in section 4.3 we reformulate it as a frame-theoretic optimization problem as discussed in [13, 12]. Using a compactness argument, we show that solutions exist. Section 4.4 simplifies the problem by showing that we need only consider orthonormal sets rather than tight-frames. This last observation is made by the author.

4.1 Quantum communication

Suppose we have a separable Hilbert space H corresponding to a physical system, but we cannot determine beforehand what state the physical system is in. However, we do know that the state of the system must be in one of a countable number of possible unit normed states $\{\psi_i\}_{i\in K} \subset H$, where $K \subset \mathbb{Z}$, with corresponding probabilities $\{\rho_i\}_{i\in K}$ that sum to 1. Our job is to determine what state the system is in, and the only way to do so is to perform a measurement. Hence, our job is to construct a POM Π with outcomes $X = K$ with the property that if the state of the system is ψ_i for some $i \in K$, our measurement tells us the system is in the ith state with high probability

$$P(j) = \langle \psi_i, \Pi(j)\psi_i \rangle \approx \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}.$$

If the state of the system is ψ_i, then $\langle \psi_i, \Pi(j)\psi_i \rangle$ is the probability that our measurement device outputs j. So, $\langle \psi_i, \Pi(i)\psi_i \rangle$ is the probability of a correct measurement. Since each ψ_j occurs with probability ρ_j, the average probability of a successful measurement is

$$\mathcal{E}(\text{success}) = \mathcal{E}(\{\langle \psi_i, \Pi(i)\psi_i \rangle\}_{i \in K}) = \sum_{i \in K} \rho_i \langle \psi_i, \Pi(i)\psi_i \rangle.$$

Quite naturally the probability of a detection error, that is the average probability that our measurement is incorrect, is given by

$$P_e = 1 - \sum_{i \in K} \rho_i \langle \psi_i, \Pi(i)\psi_i \rangle.$$

So we want to construct a POM Π that minimizes P_e.

4.2 A closer look at the detection error

Here we show that the above expression for P_e is the average of the probabilities of incorrect measurements. If the state of the system is ψ_i for some $i \in K$ and if $i \neq j$, then $\langle \psi_i, \Pi(j)\psi_i \rangle$ is the probability that we incorrectly measure the system to be ψ_j, an incorrect measurement. So, the average probability of an incorrect measurement is given by

$$\mathcal{E}(\text{incorrect}) = \mathcal{E}(\{\langle \psi_i, \Pi(j)\psi_i \rangle\}_{i \neq j}) = \sum_{i \neq j} \rho_i \langle \psi_i, \Pi(j)\psi_i \rangle.$$

We want to show that $P_e = \mathcal{E}(\text{incorrect})$. To show this, note that

$$\sum_{i \neq j} \rho_i \langle \psi_i, \Pi(j)\psi_i \rangle + \sum_{i \in K} \rho_i \langle \psi_i, \Pi(i)\psi_i \rangle \quad = \quad \sum_{i,j \in K} \rho_i \langle \psi_i, \Pi(j)\psi_i \rangle$$

$$
\begin{aligned}
&= \sum_{i\in K} \rho_i \left\langle \psi_i, \sum_{j\in K} \Pi(j)\psi_i \right\rangle \\
&= \sum_{i\in K} \rho_i \langle \psi_i, I\psi_i \rangle \\
&= \sum_{i\in K} \rho_i = 1
\end{aligned}
$$

hence

$$P_e = 1 - \sum_{i\in K} \rho_i \langle \psi_i, \Pi(i)\psi_i \rangle = \sum_{i\neq j} \rho_i \langle \psi_i, \Pi(j)\psi_i \rangle = \mathcal{E}(\text{incorrect}).$$

4.3 Using tight-frames to construct the POM

Suppose we use a 1-tight-frame $\{e_i\}_{i\in K} \subset H$ to construct our POM. Then for $i \in K$ and $x \in H$,

$$\Pi(i)x = \langle x, e_i \rangle e_i$$

and the detection error becomes,

$$
\begin{aligned}
P_e &= 1 - \sum_{i\in K} \rho_i \langle \psi_i, \Pi(i)\psi_i \rangle \\
&= 1 - \sum_{i\in K} \rho_i \langle \psi_i, \langle e_i, \psi_i \rangle e_i \rangle \\
&= 1 - \sum_{i\in K} \rho_i |\langle \psi_i, e_i \rangle|^2.
\end{aligned}
$$

So our problem reduces to finding a 1-tight-frame that minimizes P_e. Suppose $H = \mathbb{K}^d$, where $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$, and $K = \mathbb{Z}_N$. We shall show that in this case, such a tight-frame exists using a compactness argument. We start with a lemma.

Lemma 4.1. *Assume that $\{e_i\}_{i=1}^N$ is an A-tight frame for a d-dimensional Hilbert space H. Then,*

$$\|e_i\| \leq \sqrt{A}.$$

Proof. Note that for any $1 \leq k \leq N$ we have

$$\begin{aligned} A\|e_k\|^2 &= \sum_{i=1}^{N} |\langle e_k, e_i \rangle|^2 \\ &= \|e_k\|^4 + \sum_{i \neq k} |\langle e_k, e_i \rangle|^2 \end{aligned}$$

hence,

$$\|e_k\|^4 - A\|e_k\|^2 = -\textstyle\sum_{i \neq k} |\langle e_k, e_i \rangle|^2 \leq 0$$

$$\Rightarrow \|e_k\|^2 - A \leq 0$$

$$\Rightarrow \|e_k\| \leq \sqrt{A}. \qquad \square$$

Theorem 4.1. *Suppose H is a d-dimensional Hilbert space and $\{\psi_i\}_{i=1}^N \subset H$ are given. Suppose we are also given a set of positive numbers $\{\rho_i\}_{i=1}^N \subset \mathbb{R}$ such that*

$$\sum_{i=1}^{N} \rho_i = 1.$$

Then there exists a 1-tight frame $\{e_i\}_{i=1}^N \subset H$ that minimizes the error

$$P_e = 1 - \sum_{i=1}^{N} \rho_i |\langle \psi_i, e_i \rangle|^2$$

over all 1-tight frames of N elements.

Proof. Let F be the set of all N element 1-tight frames. By observation 3 of section 2.2.1, we can write this set as

$$F = \left\{ \{v_i\}_{i=1}^N \subset H : \sum_{i=1}^{N} v_i v_i^* = I \right\}.$$

First note that F is closed. Given any $\{u_i\}_{i=1}^N \subset H$ define the norm

$$\|\{u_i\}_{i=1}^N\| = \sum_{k=1}^{N} \|u_k\|_H$$

where $\|\cdot\|_H$ is the norm on H and define the operator norm for any $d \times d$ matrix A as

$$\|A\| = \sup_{\|v\|_H=1} \|Av\|_H.$$

Suppose we have a sequence $\{\{u_i^k\}_{i=1}^N\}_{k=1}^\infty \subset F$ such that

$$\lim_{k\to\infty} \|\{u_i^k\}_{i=1}^N - \{u_i\}_{i=1}^N\| = 0$$

for some set $\{u_i\}_{i=1}^N \subset H$. Then given any $\epsilon > 0$ there exists a $k > 0$ such that $\|\{u_i^k\}_{i=1}^N - \{u_i\}_{i=1}^N\| < \epsilon$. Then,

$$\begin{aligned}
\left\|\sum_{i=1}^N u_i u_i^* - I\right\| &= \left\|\sum_{i=1}^N u_i u_i^* - \sum_{i=1}^N u_i^k (u_i^k)^*\right\| + \left\|\sum_{i=1}^N u_i^k (u_i^k)^* - I\right\| \\
&= \left\|\sum_{i=1}^N u_i u_i^* - \sum_{i=1}^N u_i^k (u_i^k)^*\right\| \\
&= \sup_{\|v\|_H=1} \left\|\sum_{i=1}^N \langle v, u_i^k\rangle u_i^k - \langle v, u_i\rangle u_i\right\|_H \\
&\leq \sup_{\|v\|_H=1} \sum_{i=1}^N \|\langle v, u_i^k\rangle u_i^k - \langle v, u_i\rangle u_i\|_H \\
&\leq \sup_{\|v\|_H=1} \sum_{i=1}^N \left(\|\langle v, u_i^k\rangle u_i^k - \langle v, u_i^k\rangle u_i\|_H + \|\langle v, u_i^k\rangle u_i - \langle v, u_i\rangle u_i\|_H\right) \\
&= \sup_{\|v\|_H} \sum_{i=1}^N \left(|\langle v, u_i^k\rangle| \|u_i^k - u_i\|_H + |\langle v, u_i^k - u_i\rangle| \|u_i^k\|_H\right) \\
&\leq \sup_{\|v\|_H=1} \left(\|\{u_i^k\}_{i=1}^N - \{u_i\}_{i=1}^N\| + \|\{u_i^k\}_{i=1}^N - \{u_i\}_{i=1}^N\|\right) \\
&\leq 2\epsilon.
\end{aligned}$$

Since $\epsilon > 0$ was arbitrary, it follows that

$$\sum_{i=1}^N u_i u_i^* = I$$

hence $\{u_i\}_{I=1}^N \in F$, so F is closed.

F is also bounded since given any $\{u_i\}_{i=1}^N \in F$, by Lemma 4.1 we know that

$$\|\{u_i\}_{i=1}^N\| = \sum_{i=1}^N \|u_i\|_H \leq N.$$

Now consider the function $f_{\{\psi_i\}_{i=1}^N} : F \to \mathbb{R}$ defined for all $\{e_i\}_{i=1}^N \in F$ by

$$f_{\{\psi_i\}_{i=1}^N}(\{e_i\}_{i=1}^N) = 1 - \sum_{i=1}^N \rho_i |\langle \psi_i, e_i \rangle|^2.$$

Given any $\{v_i\}_{i=1}^N, \{u_i\}_{i=1}^N \in F$ we have

$$\begin{aligned}
|f_{\{\psi_i\}_{i=1}^N}(\{v_i\}_{i=1}^N) - f_{\{\psi_i\}_{i=1}^N}(\{u_i\}_{i=1}^N)| &= \left| \sum_{i=1}^N \rho_i |\langle \psi_i, u_i \rangle|^2 - \sum_{i=1}^N \rho_i |\langle \psi_i, e_i \rangle|^2 \right| \\
&\leq \sum_{i=1}^N \rho_i \left| |\langle \psi_i, u_i \rangle|^2 - |\langle \psi_i, e_i \rangle|^2 \right| \\
&= \sum_{i=1}^N \rho_i (|\langle \psi_i, u_i \rangle| - |\langle \psi_i, v_i|)(|\langle \psi_i, u_i \rangle| + |\langle \psi_i, v_i \rangle|) \\
&\leq 2 \sum_{i=1}^N |\langle \psi_i, u_i \rangle - \langle \psi_i, v_i \rangle| \\
&= 2 \sum_{i=1}^N |\langle \psi_i, u_i - v_i \rangle| \\
&\leq 2 \sum_{i=1}^N \|\psi_i\|_2 \|u_i - v_i\|_H = 2 \sum_{i=1}^N \|u_i - v_i\|_H \\
&= 2 \|\{u_i\} - \{v_i\}\|
\end{aligned}$$

so $f_{\{\psi_i\}_{i=1}^N}$ is continuous on F. Since F is compact, it follows that there exists $\{e_i\}_{i=1}^N \in F$ that minimizes $f_{\{\psi_i\}_{i=1}^N}$. □

4.4 P_e for tight-frames and orthonormal sets

Here, we simplify the quantum detection problem by showing that we only need to consider orthonormal sets rather than 1-tight frames. Let H be a d-dimensional

Hilbert space and let $N \in \mathbb{N}$ such that $N \geq d$. Let $\{\psi_i\}_{i=1}^N \subset H$ and $\{\rho_i\}_{i=1}^N \subset \mathbb{R}^+$ be given. For any vector set $\{e_i\}_{i=1}^N$ denote the probability error by

$$P(\{e_i\}_{i=1}^N) = 1 - \sum_{i=1}^N \rho_i |\langle \psi_i, e_i \rangle|^2.$$

Lemma 4.2. *Assume that $\widetilde{H}$ is an N-dimensional Hilbert space and $\{\widetilde{e}_i\}_{i=1}^N$ an orthonormal set for $\widetilde{H}$. Then for any subspace $U \subset \widetilde{H}$, $\{P_U \widetilde{e}_i\}_{i=1}^N$ is a 1-tight frame for U, where P_U denotes the orthogonal projection onto U.*

Proof. For any $x \in U$, note that $P_U x = x$. Since $\{\widetilde{e}_i\}_{i=1}^N$ is an orthonormal basis for $\widetilde{H}$ we can write

$$\begin{aligned} \|x\|^2 &= \sum_{i=1}^N |\langle \widetilde{e}_i, x \rangle|^2 \\ &= \sum_{i=1}^N |\langle \widetilde{e}_i, P_U x \rangle|^2 \\ &= \sum_{i=1}^N |\langle P_U \widetilde{e}_i, x \rangle|^2. \end{aligned}$$

Since this is true for all $x \in U$, it follows that $\{P_U \widetilde{e}_i\}_{i=1}^N$ is a 1-tight frame for U. $\square$

Theorem 4.2. *Let H be a d-dimensional Hilbert space and let the set of unit normed vectors $\{\psi_i\}_{i=1}^N \subset H$ be given with weights $\{\rho_i\}_{i=1}^N$. Let $\widetilde{H}$ be a N-dimensional Hilbert space such that H is a subspace of $\widetilde{H}$. Let $\{e_i\}_{i=1}^N$ be the closest 1-tight frame for H that minimizes P_e over all N element 1-tight frames for H, that is*

$$P(\{e_i\}_{i=1}^N) = \inf \left\{ P(\{\xi_i\}_{i=1}^N) : \{\xi_i\}_{i=1}^N \text{ a 1-tight frame for } H \right\}.$$

Let $\{\widetilde{e}_i\}_{i=1}^N$ *be the closest orthonormal set in* $\widetilde{H}$ *that minimizes* P_e *over all other orthonormal sets in* $\widetilde{H}$*, that is*

$$P(\{\widetilde{e}_i\}_{i=1}^N) = \inf\left\{P(\{\varphi_i\}_{i=1}^N) : \{\varphi_i\}_{i=1}^N \text{ a orthonormal set in } \widetilde{H}\right\}.$$

Then,

$$P(\{e_i\}_{i=1}^N) = P(\{\widetilde{e}_i\}_{i=1}^N) = P(\{P_H\widetilde{e}_i\}_{i=1}^N)$$

where P_H *is the orthogonal projection onto* H*.*

Proof. Since each $\psi_i \in H$, note that $P_H\psi_i = \psi_i$, so we have

$$\begin{aligned} P(\{\widetilde{e}_i\}_{i=1}^N) &= 1 - \sum_{i=1}^N \rho_i |\langle \psi_i, \widetilde{e}_i\rangle|^2 \\ &= 1 - \sum_{i=1}^N \rho_i |\langle P_H\psi_i, \widetilde{e}_i\rangle|^2 \\ &= 1 - \sum_{i=1}^N \rho_i |\langle \psi_i, P_H\widetilde{e}_i\rangle|^2 \text{ since } P_H \text{ is self-adjoint,} \\ &= P(\{P_H\widetilde{e}_i\}_{i=1}^N) \end{aligned}$$

so it remains to show that $P(\{e_i\}_{i=1}^N) = P(\{\widetilde{e}_i\}_{i=1}^N)$. By Lemma 4.2 $\{P_H\widetilde{e}_i\}_{i=1}^N$ is a 1-tight frame for H, so by the definition of the set $\{e_i\}_{i=1}^N \subset H$ it follows that

$$P(\{\widetilde{e}_i\}_{i=1}^N) = P(\{P_H\widetilde{e}_i\}_{i=1}^N) \geq P(\{e_i\}_{i=1}^N).$$

Now, by Naimark's theorem, there exists an orthonormal set $\{\theta_i\}_{i=1}^N \subset \widetilde{H}$ such that

$$\{P_H\theta_i\}_{i=1}^N = \{e_i\}_{i=1}^N.$$

Hence we have,

$$P(\{e_i\}_{i=1}^N) = 1 - \sum_{i=1}^N \rho_i |\langle \psi_i, e_i\rangle|^2$$

$$
\begin{aligned}
&= 1 - \sum_{i=1}^{N} \rho_i |\langle \psi_i, P_H \theta_i \rangle|^2 \\
&= 1 - \sum_{i=1}^{N} \rho_i |\langle P_H \psi_i, \theta_i \rangle|^2 \\
&= 1 - \sum_{i=1}^{N} \rho_i |\langle \psi_i, \theta_i \rangle|^2 \\
&= P(\{\theta_i\}_{i=1}^N) \\
&\geq P(\{\widetilde{e}_i\}_{i=1}^N)
\end{aligned}
$$

where the last inequality follows from the definition of the set $\{\widetilde{e}_i\}_{i=1}^N \subset \widetilde{H}$. The result now follows. □

In conclusion, finding the N element 1-tight frame $\{e_i\}_{i=1}^N$ for H that minimizes P_e over all N element 1-tight frames is equivalent to finding the N element orthonormal set $\{\widetilde{e}_i\}_{i=1}^N$ in $\widetilde{H}$ that minimizes P_e over all N element orthonormal sets in $\widetilde{H}$. Once we find $\{\widetilde{e}_i\}_{i=1}^N$, we project back onto H and $\{P_H \widetilde{e}_i\}_{i=1}^N$ is a 1-tight frame for H that minimizes P_e over all N element 1-tight frames.

So the quantum detection problem becomes: Let H be a d-dimensional Hilbert space and let $\{\psi_i\}_{i=1}^N \subset H$ be a normalized set with positive weights $\{\rho_i\}_{i=1}^N \subset \mathbb{R}$ where $N \geq d$. Let $\widetilde{H}$ be a N-dimensional Hilbert space such that $H \subset \widetilde{H}$. We want to find an orthonormal set $\{\widetilde{e}_i\}_{i=1}^N \subset \widetilde{H}$ that minimizes P_e over all N-element orthonormal sets in $\widetilde{H}$.

Chapter 5

Classical mechanical interpretation

In this chapter, we present a classical mechanical interpretation of the quantum detection problem. The background on Newtonian and Lagrangian mechanics in sections 5.1 and 5.2 was borrowed from [21] and the presentation of central forces and the frame force in sections 5.3 and 5.4 are from [1]. The remainder of the chapter is the contribution of the author.

In section 5.5 we give a classical mechanical interpretation of the quantum detection problem by treating the error P_e as a potential. In section 5.6, we give a method of parameterizing orthonormal sets using the group $O(N)$ and use Lagrangian mechanics to get a corresponding set of differential equations. We prove that the minimum energy solutions correspond to the tight frames that solve the quantum detection problem. In section 5.7, we add a friction term to the differential equations and show that the energies of solutions decrease. In section 5.8, we prove that it suffices to parameterize orthonormal sets using only $SO(N)$ when working with the quantum detection problem. We end the chapter with a closed form of a solution of the quantum detection problem when given two vectors.

5.1 Newtonian mechanics of 1 particle

Suppose we have a function $x : \mathbb{R} \to \mathbb{R}^d$ which is twice differentiable. For $t \in \mathbb{R}$, we denote the derivative of $x(t)$ as $\dot{x}(t)$ and the second derivative as $\ddot{x}(t)$. $x(t)$ is interpreted as the position of a particle in $\mathbb{R}^d$ at time $t \in \mathbb{R}$. A force acting on x is a vector field $F : \mathbb{R}^d \to \mathbb{R}^d$ and determines the dynamics of x by Newton's equation

$$\ddot{x}(t) = F(x(t)).$$

The force is said to be a *conservative force* if there exists a differentiable function $V : \mathbb{R}^n \to \mathbb{R}$ such that

$$F = -\nabla V$$

where ∇ is the d-dimensional gradient. V is called the *potential* of the force F.

Theorem 5.1. *If $x(t)$ is a solution to Newton's equation and the force is conservative, then it can be shown that the total energy defined by*

$$E(t) = \frac{1}{2}[\dot{x}(t)]^2 + V(x(t))$$

is constant with respect to the variable t.

Proof. Assume that $x(t)$ is a solution to Newton's equation. Multiplying Newton's equation by $\dot{x}(t)$, we obtain,

$$\dot{x}(t) \cdot \ddot{x}(t) = \dot{x}(t) \cdot F(x(t)).$$

Since F is conservative, there exists a function $V : \mathbb{R}^d \to \mathbb{R}$ such that $F = -\nabla V$. So we have

$$\frac{d}{dt}\left[\frac{1}{2}[\dot{x}(t)]^2\right] = \dot{x}(t) \cdot \ddot{x}(t) = \dot{x}(t) \cdot F(x(t)) = -\nabla V(x(t)) \cdot \dot{x}(t) = -\frac{d}{dt}V(x(t))$$

so

$$\frac{d}{dt}E(t) = \frac{d}{dt}\left[\frac{1}{2}[\dot{x}(t)]^2 + V(x(t))\right] = 0.$$

Since $E(t)$ is clearly continuous, the result follows. □

5.2 Lagrangian mechanics of N particles

Suppose we have N particles in $\mathbb{R}^d$ whose positions are modeled by N twice differentiable functions, for $i = 1, \ldots, N$, $\widetilde{e}_i : \mathbb{R}^K \to \mathbb{R}^d$ where K is not necessarily 1. Suppose each particle has a corresponding force F_i acting on it with a corresponding potential V_i. Denote by $C^2(\mathbb{R})$ the space of all real valued functions that are twice differentiable. Define the Lagrangian function $L : (C^2(\mathbb{R}))^K \to C^1(\mathbb{R})$ for all $\{q_i(t)\}_{i=1}^K \subset C^2(\mathbb{R})$ by

$$L = \sum_{i=1}^{N}\left[\frac{1}{2}\dot{\widetilde{e}}_i(q_1(t), \ldots, q_K(t)) \cdot \dot{\widetilde{e}}_i(q_1(t), \ldots, q_K(t)) - V_i(\widetilde{e}_i(q_1(t), \ldots, q_K(t)))\right]$$

where for each $1 \leq i \leq N$,

$$\dot{\widetilde{e}}_i(q_1(t), \ldots, q_K(t)) = \frac{d}{dt}\widetilde{e}_i(q_1(t), \ldots, q_K(t)).$$

Then the Euler-Lagrange equations of motion are the K differential equations for each $1 \leq i \leq K$ given by

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_i}\right) - \frac{\partial L}{\partial q_i} = 0$$

that determine the dynamics of $\{q_i(t)\}_{i=1}^K$.

We define the *total energy* of the system by

$$E = \frac{1}{2}\sum_{i=1}^{N}\dot{\widetilde{e}}_i \cdot \dot{\widetilde{e}}_i + \sum_{i=1}^{N} V_i(\{\widetilde{e}_i\}_{i=1}^N).$$

Theorem 5.2. *We can write the kinetic energy in terms of the variables* $\{q_i\}_{i=1}^K$ *as*

$$T = \frac{1}{2}\sum_{i=1}^{N} \dot{\widetilde{e}}_i \cdot \dot{\widetilde{e}}_i = \frac{1}{2}\sum_{i=1}^{K} \dot{q}_i \frac{dT}{d\dot{q}_i}.$$

Proof. Denote the kth component of $\widetilde{e}_i$ by $\widetilde{e}_{i,k}$. Then

$$\dot{\widetilde{e}}_{i,l} = \sum_{k=1}^{K} \frac{\partial \widetilde{e}_{i,l}}{\partial q_k}\dot{q}_k$$

and

$$\dot{\widetilde{e}}_{i,l}^2 = \sum_{k=1}^{K}\sum_{m=1}^{K} \frac{\partial \widetilde{e}_{i,l}}{\partial q_k}\frac{\partial \widetilde{e}_{i,l}}{\partial q_m}\dot{q}_k\dot{q}_m.$$

So,

$$\begin{aligned}
\dot{\widetilde{e}}_i \cdot \dot{\widetilde{e}}_i &= \sum_{l=1}^{d} \dot{\widetilde{e}}_{i,l}^2 \\
&= \sum_{l=1}^{d}\sum_{k=1}^{K}\sum_{m=1}^{K} \frac{\partial \widetilde{e}_{i,l}}{\partial q_k}\frac{\partial \widetilde{e}_{i,l}}{\partial q_m}\dot{q}_k\dot{q}_m.
\end{aligned}$$

Hence,

$$\begin{aligned}
T &= \frac{1}{2}\sum_{i=1}^{N} \dot{\widetilde{e}}_i \cdot \dot{\widetilde{e}}_i \\
&= \frac{1}{2}\sum_{i=1}^{N}\sum_{l=1}^{d}\sum_{k=1}^{K}\sum_{m=1}^{K} \frac{\partial \widetilde{e}_{i,l}}{\partial q_k}\frac{\partial \widetilde{e}_{i,l}}{\partial q_m}\dot{q}_k\dot{q}_m \\
\frac{\partial T}{\partial \dot{q}_p} &= \frac{1}{2}\sum_{i=1}^{N}\sum_{l=1}^{d}\sum_{k=1}^{K} \frac{\partial \widetilde{e}_{i,l}}{\partial q_k}\frac{\partial \widetilde{e}_{i,l}}{\partial q_p}\dot{q}_k + \frac{1}{2}\sum_{i=1}^{N}\sum_{l=1}^{d}\sum_{m=1}^{K} \frac{\partial \widetilde{e}_{i,l}}{\partial q_p}\frac{\partial \widetilde{e}_{i,l}}{\partial q_m}\dot{q}_m \\
\sum_{p=1}^{K}\dot{q}_p\frac{\partial T}{\partial \dot{q}_p} &= \frac{1}{2}\sum_{p=1}^{K}\sum_{i=1}^{N}\sum_{l=1}^{d}\sum_{k=1}^{K} \frac{\partial \widetilde{e}_{i,l}}{\partial q_k}\frac{\partial \widetilde{e}_{i,l}}{\partial q_p}\dot{q}_p\dot{q}_k + \frac{1}{2}\sum_{p=1}^{K}\sum_{i=1}^{N}\sum_{l=1}^{d}\sum_{m=1}^{K} \frac{\partial \widetilde{e}_{i,l}}{\partial q_p}\frac{\partial \widetilde{e}_{i,l}}{\partial q_m}\dot{q}_p\dot{q}_m \\
&= \sum_{p=1}^{K}\sum_{i=1}^{N}\sum_{l=1}^{d}\sum_{m=1}^{K} \frac{\partial \widetilde{e}_{i,l}}{\partial q_k}\frac{\partial \widetilde{e}_{i,l}}{\partial q_p}\dot{q}_p\dot{q}_m = 2T.
\end{aligned}$$

Solving for T gives us

$$T = \frac{1}{2}\sum_{p=1}^{K} \dot{q}_p \frac{\partial T}{\partial \dot{q}_p}.$$

□

Theorem 5.3. *If $\{q_i(t)\}_{i=1}^K$ satisfies the Euler-Lagrange equations of motion and the potential V_i is independent of the variables $\{\dot{q}_i\}_{i=1}^K$, then E is a constant in time.*

Proof. We first take the time derivative of the Lagrangian and get

$$\frac{dL}{dt} = \sum_{j=1}^{K} \frac{\partial L}{\partial q_j}\dot{q}_j + \sum_{j=1}^{K} \frac{\partial L}{\partial \dot{q}_j}\ddot{q}_j.$$

Since the $\{q_i\}_{i=1}^K$ satisfy the Euler-Lagrange equations,

$$\frac{\partial L}{\partial q_j} = \frac{d}{dt}\frac{\partial L}{\partial \dot{q}_j}$$

hence plugging this into our derivative of L gives us,

$$\frac{dL}{dt} = \sum_{j=1}^{K} \left[\dot{q}_j \frac{d}{dt}\frac{\partial L}{\partial \dot{q}_j} + \frac{\partial L}{\partial \dot{q}_j}\ddot{q}_j\right] = \sum_{j=1}^{K} \frac{d}{dt}\left[\dot{q}_j \frac{\partial L}{\partial \dot{q}_j}\right]$$

hence

$$\frac{d}{dt}\left[\sum_{j=1}^{K} \dot{q}_j \frac{\partial L}{\partial \dot{q}_j} - L\right] = 0.$$

Now since each V_i is independent of the variables $\{\dot{q}_i\}_{i=1}^K$, we have that

$$\frac{\partial L}{\partial \dot{q}_j} = \frac{\partial T}{\partial \dot{q}_j}.$$

Using this relation and the previous theorem gives us

$$\sum_{j=1}^{K} \dot{q}_j \frac{\partial L}{\partial \dot{q}_j} - L = \sum_{j=1}^{K} \dot{q}_j \frac{\partial T}{\partial \dot{q}_j} - L = 2T - L = T + V = E$$

hence

$$\frac{dE}{dt} = 0.$$

□

5.3 Central force

Suppose we have an ensemble of particles in $\mathbb{R}^d$ that interact with one another by a conservative force $F : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$. Given two particles $\vec{a}, \vec{b} \in \mathbb{R}^d$, $\vec{a}$ feels the force from $\vec{b}$ given by $F(\vec{a}, \vec{b})$. If the force is conservative, then there exists a potential function $P : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ such that

$$F(\vec{a}, \vec{b}) = -\nabla_{\vec{a}-\vec{b}} P(\vec{a}, \vec{b})$$

where $\nabla_{\vec{a}-\vec{b}}$ is the gradient taken by keeping $\vec{b}$ fixed and differentiating with respect to $\vec{a}$. Denote by $\mathbb{R}^+$ as the set of all positive real numbers. The force F is a *central force* if its magnitude depends only on the distance $\|\vec{a} - \vec{b}\|$, that is there exists a function $f : \mathbb{R}^+ \to \mathbb{R}$ such that for all $\vec{a}, \vec{b} \in \mathbb{R}^d$,

$$F(\vec{a}, \vec{b}) = f(\|\vec{a} - \vec{b}\|)[\vec{a} - \vec{b}].$$

In this case, the same can be said of the potential, that is if the force is conservative and central, then there is a function $p : \mathbb{R}^+ \to \mathbb{R}$ such that

$$P(\vec{a}, \vec{b}) = p(\|\vec{a} - \vec{b}\|).$$

Computing the potential for conservative central forces is simple. For any $\vec{a}, \vec{b} \in \mathbb{R}^d$ the condition

$$F(\vec{a}, \vec{b}) = -\nabla_{(\vec{a}-\vec{b})} P(\vec{a}, \vec{b})$$

implies that for all $x \in \mathbb{R}^+$,

$$p'(x) = -xf(x).$$

To show this, note that for some $\vec{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$,

$$\nabla\|\vec{x}\| = \nabla\sqrt{x_1^2 + \ldots x_d^2} = \begin{bmatrix} \frac{x_1}{\sqrt{x_1^2+\ldots+x_d^2}} \\ \vdots \\ \frac{x_d}{\sqrt{x_1^2+\ldots+x_d^2}} \end{bmatrix} = \frac{\vec{x}}{\|\vec{x}\|}$$

So, setting $\vec{x} = \vec{a} - \vec{b} \in \mathbb{R}^d$,

$$-\nabla P(\vec{a}, \vec{b}) = -\nabla p(\|\vec{x}\|)\|\vec{x}\| = -p'(\|\vec{x}\|)\nabla\|\vec{x}\| = -p'(\|\vec{x}\|)\vec{x}/\|\vec{x}\|.$$

Setting this equal to $F(\vec{a}, \vec{b}) = f(\|\vec{x}\|)\vec{x}$ gives us

$$p'(\|\vec{x}\|) = -\|\vec{x}\|f(\|\vec{x}\|)$$

which is want we wanted.

5.4 Frame force

Two electrons with charge e and positions given by $x, y \in \mathbb{R}^3$ feel a repulsive force given by Coulomb's law. Particle x feels the force exerted on it by particle y given by

$$F = K\frac{e^2}{\|x - y\|^2}(x - y)$$

where K is a constant. Suppose we have a metallic sphere where a number of electrons move freely and interact with each other by the Coulomb force. An unsolved problem in physics is to determine the equilibrium positions of the electrons, that is an arrangement of the electrons where all the interaction forces cancel so there is no motion.

Benedetto-Fickus [1] used a similar idea to characterize all finite unit-normed tight frames. The goal was to find a force such that the equilibrium positions on the sphere would correspond to finite unit-normed tight-frames. Given two points $x, y \in \mathbb{R}^d$, particle x feels the force exerted on it by particle y given by the frame force

$$FF(x, y) = \langle x, y \rangle (x - y).$$

It can be shown that this is a central force with the frame potential given by

$$FP = \frac{1}{2} |\langle x, y \rangle|^2.$$

Given a collection of unit-normed points $\{x_i\}_{i=1}^N \subset \mathbb{R}^d$ the total frame potential is given by

$$TFP\left(\{x_i\}_{i=1}^N\right) = \sum_{m=1}^{N} \sum_{n=1}^{N} |\langle x_m, x_n \rangle|^2.$$

Theorem 5.4. *Let $N \leq d$. The minimum value of the total frame potential for the frame force and N variables, is N; and the minimizers are the orthonormal sets of N elements in $\mathbb{R}^N$.*

Theorem 5.5. *Let $N \geq d$. The minimum value of the total frame potential, for the frame force and N variables, is N^2/d; and the minimizers are the finite-unit-normed tight frames of N elements for $\mathbb{R}^d$.*

5.5 Physical interpretation of the frame problem

Inspired by the Benedetto-Fickus frame force [1], the quantum detection problem, as stated in section 4.4, can be given another physical interpretation in the case where

$H = \mathbb{R}^d$. Let $H \subset \widetilde{H} = \mathbb{R}^N$. We want to find the orthonormal set $\{\widetilde{e}_i\}_{i=1}^N \subset \widetilde{H}$ that minimizes P_e over all N element orthonormal sets in $\widetilde{H}$. We consider the error P_e as a potential

$$V = P_e = \sum_{i=1}^{N} \rho_i (1 - |\langle \psi_i, \widetilde{e}_i \rangle|^2) = \sum_{i=1}^{N} V_i$$

where each

$$\begin{aligned} V_i &= \rho_i (1 - \langle \psi_i, \widetilde{e}_i \rangle^2) = \rho_i \left(1 - \left(1 - \frac{1}{2} \|\psi_i - \widetilde{e}_i\|^2 \right)^2 \right) \\ &= \rho_i \left(1 - \left(1 - \frac{1}{2} \|\psi_i - \widetilde{e}_i\|^2 \right)^2 \right) \end{aligned}$$

where we have used the fact that $\|\psi_i\| = \|\widetilde{e}_i\| = 1$ and the relation

$$\|\psi_i - \widetilde{e}_i\|^2 = \langle \psi_i - \widetilde{e}_i, \psi_i - \widetilde{e}_i \rangle = \|\psi_i\|^2 - 2\langle \psi_i, \widetilde{e}_i \rangle + \|\widetilde{e}_i\|^2 = 2 - 2\langle \psi_i, \widetilde{e}_i \rangle.$$

Since each V_i is a function of the distance $\|\psi_i - \widetilde{e}_i\|$, V_i corresponds to a conservative central force between the points ψ_i and $\widetilde{e}_i$ given by $F_i = -\nabla_i V_i$ where ∇_i is an N-dimensional gradient taken by keeping ψ_i fixed and differentiating with respect to the variable $\widetilde{e}_i$. Setting $x = \|\psi_i - \widetilde{e}_i\|$ we can write

$$V_i(\widetilde{e}_i, \psi_i) = v_i(\|\widetilde{e}_i - \psi_i\|) = \rho_i \left[1 - \left(1 - \frac{1}{2} x^2 \right)^2 \right].$$

Taking the derivative gives us

$$v_i'(x) = -2\rho_i \left(1 - \frac{1}{2} x^2 \right) (-x) = 2\rho_i \left(1 - \frac{1}{2} x^2 \right) x = -x f_i(x)$$

so

$$f_i(x) = -2\rho_i \left(1 - \frac{1}{2} x^2 \right)$$

and the corresponding central force can be written as

$$F_i(\psi_i, \widetilde{e}_i) = f_i(\|\psi_i - \widetilde{e}_i\|)(\psi_i - \widetilde{e}_i) = -2\rho_i \left(1 - \frac{1}{2} \|\psi_i - \widetilde{e}_i\|^2 \right) (\psi_i - \widetilde{e}_i) = -2\rho_i \langle \psi_i, \widetilde{e}_i \rangle (\psi_i - \widetilde{e}_i).$$

Hence this can be viewed as a physical system where the given vectors $\{\psi_i\}_{i=1}^N$ are fixed points on a sphere in $\widetilde{H}$, and we have a "rigid" N element orthonormal set $\{\widetilde{e}_i\}_{i=1}^N$ which moves according to the interactions between each $\widetilde{e}_i$ and ψ_i according to the force f_i. We want to find the equilibrium points $\{\overline{\widetilde{e}}_i\}_{i=1}^n$. These are the points where all the forces f_i balance and produce no net motion. In this situation, the potential V is minimized.

5.6 Hamiltonian system on $O(N)$

We now need to take into consideration the constraint that the set $\{\widetilde{e}_i\}_{i=1}^N$ is an orthonormal basis. In this process, we get a Hamiltonian system on $O(N)$ where $O(N)$ is the orthogonal group.

Let $\{b_i\}_{i=1}^N$ be a fixed orthonormal basis for $\widetilde{H}$. Since $O(N)$ is a smooth compact $N(N-1)/2$ dimensional manifold, there exists a finite number of open sets $\{U_k\}_{k=1}^M$ in $\mathbb{R}^{N(N-1)/2}$ and smooth mappings $\Theta_k : U_k \to O(N)$ such that

$$\bigcup_{k=1}^{M} \Theta(U_k) = O(N).$$

Since any two orthonormal sets are related by an orthogonal transformation, for each $k = 1, \ldots, M$, we can smoothly parameterize our orthonormal set in terms of $N(N-1)/2$ variables in U_k by

$$\{\widetilde{e}_i(q_1, \ldots, q_{N(N-1)/2})\}_{i=1}^N = \{\Theta_k(q_1, \ldots, q_{N(N-1)/2})b_i\}_{i=1}^N.$$

As k runs from 1 to M, we get all possible orthonormal sets. We now use Lagrangian mechanics to convert the frame forces f_i acting on the tight-frame $\{\widetilde{e}_i\}_{i=1}^N$ into a set

of differential equations that determine the dynamics of functions $\{q_i(t)\}_{i=1}^{N(N-1)/2} \subset C^2(\mathbb{R})$. Since the $N(N-1)/2$ variables can be considered as local coordinates of $O(N)$, we get a Hamiltonian system on $O(N)$ with trajectories given by $(q_1(t), \ldots, q_{N(N-1)/2}(t))$.

Define the Lagrangian function for each $\{q_i(t)\}_{i=1}^{N(N-1)/2} \subset C^2(\mathbb{R})$ by

$$\begin{aligned} L &= \frac{1}{2}\sum_{i=1}^{N}\left\|\frac{d}{dt}\widetilde{e}_i(q_1(t),\ldots,q_{N(N-1)/2}(t))\right\|^2 - P_e(q_1(t),\ldots,q_{N(N-1)/2}(t)) \\ &= T(q_1(t),\ldots,q_{N(N-1)/2}(t)) - P_e(q_1(t),\ldots,q_{N(N-1)/2}(t)) \end{aligned}$$

where

$$T(q_1(t),\ldots,q_{N(N-1)/2}(t)) = \frac{1}{2}\sum_{i=1}^{N}\|\dot{\widetilde{e}}_i(q_1(t),\ldots,q_{N(N-1)/2}(t))\|^2$$

and the dot denotes the total derivative with respect to t. Then the equations of motion of the functions $\{q_i\}_{i=1}^{N(N-1)/2}$ is given by the $N(N-1)/2$ Euler-Lagrange equations

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_j}\right) - \frac{\partial L}{\partial q_j} = 0$$

where $1 \leq j \leq N(N-1)/2$. We omit writing the variables

$$T = T(q_1(t),\ldots,q_{N(N-1)/2}(t)),\ P_e = P_e(q_1(t),\ldots,q_{N(N-1)/2}(t))$$

to simplify the notation. We can write

$$\begin{aligned} \frac{\partial L}{\partial q_j} &= -\frac{\partial V}{\partial q_j} = -\frac{\partial}{\partial q_j}\sum_{i=1}^{N} V_i \\ &= -\sum_{i=1}^{N}\nabla V_i \cdot \frac{\partial \widetilde{e}_i}{\partial q_j} = 2\sum_{i=1}^{N}\rho_i\langle\psi_i,\widetilde{e}_i\rangle(\widetilde{e}_i - \psi_i)\cdot\frac{\partial\widetilde{e}_i}{\partial q_j} \\ &= 2\sum_{i=1}^{N}\rho_i\langle\psi_i,\widetilde{e}_i\rangle\left\langle\widetilde{e}_i,\frac{\partial\widetilde{e}_i}{\partial q_j}\right\rangle - 2\sum_{i=1}^{N}\rho_i\langle\psi_i,\widetilde{e}_i\rangle\left\langle\psi_i,\frac{\partial\widetilde{e}_i}{\partial q_j}\right\rangle. \end{aligned}$$

Since

$$\langle\widetilde{e}_i,\widetilde{e}_i\rangle = 1$$

taking the derivative with respect to q_j gives us

$$\left\langle \frac{\partial}{\partial q_j}\widetilde{e}_i, \widetilde{e}_i \right\rangle + \left\langle \widetilde{e}_i, \frac{\partial}{\partial q_j}\widetilde{e}_i \right\rangle = 0$$

so,

$$\left\langle \frac{\partial}{\partial q_j}\widetilde{e}_i, \widetilde{e}_i \right\rangle = 0$$

and we have

$$\frac{\partial L}{\partial q_j} = -2\sum_{i=1}^{N} \rho_i \langle \psi_i, \widetilde{e}_i \rangle \left\langle \psi_i, \frac{\partial \widetilde{e}_i}{\partial q_j} \right\rangle.$$

Also

$$\frac{\partial L}{\partial \dot{q}_j} = \frac{\partial}{\partial \dot{q}_j}(T - P_e) = \frac{\partial T}{\partial \dot{q}_j}$$

since P_e is independent of $\dot{q}_j$. So the Euler-Lagrange equations become

$$\frac{d}{dt}\left(\frac{\partial T}{\partial \dot{q}_j}\right) = -2\sum_{i=1}^{N} \rho_i \langle \psi_i, \widetilde{e}_i \rangle \left\langle \psi_i, \frac{\partial \widetilde{e}_i}{\partial q_j} \right\rangle.$$

By Theorem 5.3, it can be shown that if $(q_1(t), \ldots, q_{N(N-1)/2}(t))$ is a solution to the Euler-Lagrange equations of motion, then the energy

$$E(t) = \frac{1}{2}\sum_{i=1}^{N} \|\dot{\widetilde{e}}_i(q_1(t), \ldots, q_{N(N-1)/2}(t))\|^2 + P_e(q_1(t), \ldots, q_{N(N-1)/2}(t))$$

is a constant in time t.

Lemma 5.1. *Let $\{\psi_i\}_{i=1}^N \subset H$ be given with corresponding positive weights $\{\rho_i\}_{i=1}^N$. Let $\{\overline{\widetilde{e}}_i\}_{i=1}^N$ be the orthonormal set that minimizes P_e. Let $\Theta_k(\tilde{q}_1, \ldots, \tilde{q}_{N(N-1)/2})$ be a point in $O(N)$ such that for each $i = 1, \ldots, N$,*

$$\widetilde{e}_i(\tilde{q}_1, \ldots, \tilde{q}_{N(N-1)/2}) = \overline{\widetilde{e}}_i.$$

Then the constant function defined by

$$(q_1(t), \ldots, q_{N(N-1)/2}(t)) = (\tilde{q}_1, \ldots, \tilde{q}_{N(N-1)/2})$$

is a solution of the Euler-Lagrange equations of motion in $O(N)$ that minimizes the energy E and

$$\sum_{i=1}^{N} \rho_i \langle \psi_i, \widetilde{e}_i(\tilde{q}_1, \ldots, \tilde{q}_{N(N-1)/2}) \rangle \left\langle \psi_i, \frac{\partial \widetilde{e}_i}{\partial q_j}(\tilde{q}_1, \ldots, \tilde{q}_{N(N-1)/2}) \right\rangle = 0.$$

Proof. First, since $\{\widetilde{e}_i\}_{i=1}^N$ minimizes P_e at the point $(\tilde{q}_1, \ldots, \tilde{q}_{N(N-1)/2})$, we must have for all $j = 1, \ldots, N(N-1)/2$,

$$\frac{\partial P_e}{\partial q_j}(\tilde{q}_1, \ldots, \tilde{q}_{N(N-1)/2}) = 0.$$

Since

$$\frac{\partial P_e}{\partial q_j} = \sum_{i=1}^{N} \rho_i \langle \psi_i, \widetilde{e}_i \rangle \left\langle \psi_i, \frac{\partial \widetilde{e}_i}{\partial q_j} \right\rangle$$

we have one of our assertions.

Second, we show that this is a solution to the Euler-Lagrange equations. Each $\widetilde{e}_i(q_1, \ldots, q_{N(N-1)/2})$ is constant with respect to t, hence

$$\begin{aligned} \frac{d}{dt}\frac{\partial T}{\partial \dot{q}_j}(q_1, \ldots, q_{N(N-1)/2}) &= 0 = -2\frac{\partial P_e}{\partial q_j}(q_1, \ldots, q_{N(N-1)/2})) \\ &= -2\sum_{i=1}^{N} \rho_i \langle \psi_i, \widetilde{e}_i(q_1, \ldots, q_{N(N-1)/2}) \rangle \left\langle \psi_i, \frac{\partial \widetilde{e}_i}{\partial q_j}(q_1, \ldots, q_{N(N-1)/2}) \right\rangle \end{aligned}$$

so $(q_1, \ldots, q_{N(N-1)/2})$ is a solution to the Euler-Lagrange equations.

Furthermore, since for each $i = 1, \ldots, N$

$$\dot{\widetilde{e}}_i(\tilde{q}_1, \ldots, \tilde{q}_{N(N-1)/2}) = 0$$

the energy becomes

$$E = P_e$$

and since $\widetilde{e}_i(\tilde{q}_1, \ldots, \tilde{q}_{N(N-1)/2})$ minimizes P_e, it follows that the energy is minimized.

□

The utility of this lemma is that it opens the problem to numerical approximations, for example a multidimensional Newton iteration could be used to approximate the $(q_1, \ldots, q_{N(N-1)/2})$ that satisfy the above expression. Furthermore, the error P_e can now be considered as a smooth function of the variables $(q_1, \ldots, q_{N(N-1)/2})$, hence other numerical methods become available. For example, the conjugate gradient method may be used to approximate a 1-tight-frame that minimizes P_e.

The following lemma and theorem relates the Hamiltonian system with the original quantum detection problem.

Lemma 5.2. *Assume that $(q_1(t), \ldots, q_{N(N-1)/2}(t))$ is a solution to the equations of motion that is not a constant solution. Denote the domain of $(q_1(t), \ldots, q_{N(N-1)/2}(t))$ by $Dom(\vec{q})$. Then there exists a $t_0 \in Dom(\vec{q})$ such that*

$$T(q_1(t_0), \ldots, q_{N(N-1)/2}(t_0)) \neq 0.$$

Proof. Suppose not. Then for all $t \in Dom(\vec{q})$,

$$T(q_1(t), \ldots, q_{N(N-1)/2}(t)) = 0.$$

By the definition of T it follows that for all $i = 1, \ldots, N$ and $t \in Dom(\vec{q})$,

$$\dot{\widetilde{e}}_i(q_1(t), \ldots, q_{N(N-1)/2}(t)) = 0$$

hence for all $i = 1, \ldots, N$ and some $k = 1, \ldots, M$,

$$\widetilde{e}_i(q_1(t), \ldots, q_{N(N-1)/2}(t)) = \Theta_k(q_1(t), \ldots, q_{N(N-1)/2}(t))b_i$$

is constant with respect to t. Since $\{b_i\}_{i=1}^N$ is an orthonormal basis, it follows that

$$\Theta_k(q_1(t), \ldots, q_{N(N-1)/2}(t))$$

is constant, with respect to t. Since it was assumed that $(q_1(t), \ldots, q_{N(N-1)/2}(t))$ is not constant with respect to t, this contradicts the fact that $\Theta_k : U_k \to O(N)$ is a diffeomorphism since it would not be one-to-one. $\square$

Theorem 5.6. *Let $(q_1(t), \ldots, q_{N(N-1)/2}(t))$ be the solution of the Euler-Lagrange equations of motion that minimizes the energy E. Then $(q_1(t), \ldots, q_{N(N-1)/2}(t))$ is a constant solution, that is for all $i = 1, \ldots, N(N-1)/2$,*

$$\frac{dq_i}{dt}(t) = 0$$

and

$$\{P_H \widetilde{e}_i(q_1(t), \ldots, q_{N(N-1)/2}(t))\}_{i=1}^N \subset H$$

is the 1-tight frame for H that minimizes P_e.

Proof. Suppose $(q_1(t), \ldots, q_{N(N-1)/2}(t))$ is a solution of the Euler-Lagrange equations of motion that minimizes the energy E. Assume that $(q_1(t), \ldots, q_{N(N-1)/2}(t))$ is not a constant solution. Denote by $(\tilde{q}_1, \ldots, \tilde{q}_{N(N-1)/2})$ the point from Lemma 5.1 such that

$$\{\widetilde{e}_i(\tilde{q}_1, \ldots, \tilde{q}_{N(N-1)/2})\}_{i=1}^N$$

is the orthonormal set that minimizes P_e. Then by Lemma 5.2 there exists a $t_0 \in \mathbb{R}$ such that

$$T = \frac{1}{2} \sum_{i=1}^{N} \|\dot{\widetilde{e}}_i(q_1(t_0), \ldots, q_{N(N-1)/2}(t_0))\|^2 \neq 0$$

and by Theorem 5.3 the energy is constant, so for all t we have

$$\begin{aligned} E(q_1(t), \ldots, q_{N(N-1)/2}(t)) &= T(q_1(t_0), \ldots, q_{N(N-1)/2}(t_0)) + P_e(q_1(t_0), \ldots, q_{N(N-1)/2}(t_0)) \\ &> P_e(\tilde{q}_1, \ldots, \tilde{q}_{N(N-1)/2}) = E(\tilde{q}_1, \ldots, \tilde{q}_{N(N-1)/2}) \end{aligned}$$

which contradicts the assumption that $(q_1(t), \ldots, q_{N(N-1)/2}(t))$ is the solution that minimizes the energy E. It follows that $(q_1(t), \ldots, q_{N(N-1)/2}(t))$ must be a constant solution, hence $T = 0$, so it minimizes $E = P_e$. By Theorem 4.2 it follows that

$$\{P_H \widetilde{e}_i(q_1(t), \ldots, q_{N(N-1)/2}(t))\}_{i=1}^N \subset H$$

is the 1-tight frame for H that minimizes P_e. □

5.7 Friction

Intuitively, since the equations of motion is a conservative system, it is possible that solutions may oscillate around the optimum value. However, if we add a friction term in the equations of motion, it is possible that solutions will converge to the optimal value of $(q_1, \ldots, q_{N(N-1)/2})$ that minimize the potential V.

Now consider adding a friction term. The idea is that with friction, solutions will tend to the minimum energy solutions. The modified equations of motion with friction are for each $j = 1, \ldots, N(N-1)/2$

$$\frac{d}{dt}\frac{\partial T}{\partial \dot{q}_j} + \frac{\partial P_e}{\partial q_j} = -\dot{q}_j.$$

Theorem 5.7. *Assume that $(q_1(t), \ldots, q_{N(N-1)/2}(t))$ is a solution to the modified equations of motion given by*

$$\frac{d}{dt}\frac{\partial T}{\partial \dot{q}_j} + \frac{\partial P_e}{\partial q_j} = -\dot{q}_j.$$

Then the energy satisfies

$$\frac{d}{dt}E(t) = -\sum_{i=1}^{N(N-1)/2} \dot{q}_i^2.$$

Proof. Multiplying the equations of motion by $\dot{q}_j$ and summing over j gives us

$$\sum_{j=1}^{N(N-1)/2} \left[\frac{d}{dt}\frac{\partial T}{\partial \dot{q}_j} + \frac{\partial P_e}{\partial q_j} \right] \dot{q}_j \;=\; -\sum_{j=1}^{N(N-1)/2} \dot{q}_j^2.$$

Note that the right term is

$$\sum_{j=1}^{N(N-1)/2} \frac{\partial P_e}{\partial q_j}\dot{q}_j = \frac{dP_e}{dt}.$$

We can write

$$\frac{d}{dt}\sum_{j=1}^{N(N-1)/2} \frac{\partial T}{\partial \dot{q}_j}\dot{q}_j = \sum_{i=1}^{N(N-1)/2}\sum_{j=1}^{N(N-1)/2} \frac{\partial^2 T}{\partial \dot{q}_i \partial \dot{q}_j}\dot{q}_j + \sum_{j=1}^{N(N-1)/2} \frac{\partial T}{\partial \dot{q}_j}\ddot{q}_j$$

hence

$$\begin{aligned}\sum_{j=1}^{N(N-1)/2} \frac{d}{dt}\frac{\partial T}{\partial \dot{q}_j}\dot{q}_j &= \sum_{j=1}^{N(N-1)/2}\sum_{i=1}^{N(N-1)/2} \frac{\partial^2 T}{\partial \dot{q}_i \partial \dot{q}_j}\ddot{q}_i\dot{q}_j \\ &= \frac{d}{dt}\sum_{1}^{N(N-1)/2} \frac{\partial T}{\partial \dot{q}_j}\dot{q}_j - \frac{d}{dt}T = \frac{dT}{dt}\end{aligned}$$

So we have,

$$\frac{d}{dt}(T + P_e) = -\sum_{j=1}^{N(N-1)/2} \dot{q}_j^2.$$

which is what we wanted. □

5.8 Parameterization on $SO(N)$

Let $\{\widetilde{e}_i\}_{i=1}^N$ be an orthonormal basis for H'. We can locally parameterize the elements in $O(N)$ by $N(N-1)/2$ variables so that $\theta(q_1, \ldots, q_{N(N-1)/2}) \in O(N)$. We get a smooth parameterization of our orthonormal set by setting for all $i = 1, \ldots, N$,

$$\widetilde{e}_i(q_1, \ldots, q_{N(N-1)/2}) = \theta(q_1, \ldots, q_{N(N-1)/2})\widetilde{e}_i.$$

Now $O(N)$ has two connected components, $SO(N)$ and $G(N) = O(N) - SO(N)$. So this parameterization depends on the choice of which component $\theta(q_1, \ldots, q_{N(N-1)/2}) \in O(N)$ is in.

Lemma 5.3. *Let $\{\widetilde{e}_i\}_{i=1}^N$ be an orthonormal basis for the Hilbert space H' and denote by ξ the linear transformation defined by*

$$\begin{aligned} \xi(\widetilde{e}_1) &= -\widetilde{e}_1 \\ \xi(\widetilde{e}_i) &= \widetilde{e}_i \forall\ N > i > 1. \end{aligned}$$

Define the function $g : SO(N) \to G(N)$ for all $\theta \in SO(N)$ by

$$g(\theta) = \theta \cdot \xi.$$

Then g is a bijection.

Proof. For all $\theta \in SO(N)$, it is clear that $g(\theta) \in G(N)$ since

$$\det(\theta) = 1 \Rightarrow \det(g(\theta)) = \det(\theta \cdot \xi) = \det(\theta) \cdot \det(\xi) = -1 \Rightarrow g(\theta) \in G(N).$$

With respect to the basis $\{\widetilde{e}_i\}_{i=1}^N$, we can write ξ as

$$\xi = \begin{bmatrix} -1 & 0 & & \\ 0 & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix}.$$

Clearly, ξ is invertible, hence injective, and surjective, so g is a bijection. $\square$

Theorem 5.8. *Let $\{\widetilde{e}_i\}_{i=1}^N$ be a orthonormal basis for a real Hilbert space H', $\{\psi_i\}_{i=1}^N \subset H'$ a fixed set of unit normed vectors, and weights $\{\rho_i\}_{i=1}^N \subset \mathbb{R}^+$. Consider the error function $P : O(N) \to \mathbb{R}$ defined for all $\theta \in O(N)$ by*

$$P(\theta) = 1 - \sum_{i=1}^{N} \rho_i |\langle \psi_i, \theta \widetilde{e}_i \rangle|^2.$$

Since $SO(N)$ is compact and P is continuous, there exists a $\theta' \in SO(N)$ such that for all $\theta \in SO(N)$,

$$P(\theta') \leq P(\theta).$$

Similarly, since $G(N)$ is compact, there exists a $\theta'' \in G(N)$ such that for all $\theta \in G(N)$,

$$P(\theta'') \leq P(\theta).$$

Then,

$$P(\theta') = P(\theta'').$$

Proof. First, note that for any $\theta \in SO(N)$,

$$\begin{aligned}
P(g(\theta)) &= 1 - \sum_{i=1}^{N} \rho_i |\langle \psi_i, g(\theta) \widetilde{e}_i \rangle|^2 \\
&= 1 - \sum_{i=1}^{N} \rho_i |\langle \psi_i, \theta \cdot \xi \widetilde{e}_i \rangle|^2 \\
&= 1 - \rho_1 |\langle \psi_1, \theta(-\widetilde{e}_1) \rangle|^2 - \sum_{i=2}^{N} \rho_i |\langle \psi_i, \theta \widetilde{e}_i \rangle|^2 \\
&= 1 - \rho_1 |\langle \psi_1, \theta(\widetilde{e}_1) \rangle|^2 - \sum_{i=2}^{N} \rho_i |\langle \psi_i, \theta \widetilde{e}_i \rangle|^2 \\
&= P(\theta).
\end{aligned}$$

We complete the proof by contradiction. Suppose that $P(\theta') \neq P(\theta'')$. Consider the case that $P(\theta'') > P(\theta')$. Then $g(\theta') \in G(N)$ has the property that $P(\theta'') > P(g(\theta'))$

which contradictions the definition of $\theta'' \in G(N)$. A similar argument works for the case with $P(\theta'') < P(\theta')$ by considering the function $g^{-1} : G(N) \to SO(N)$. $\square$

By the above theorem, it suffices to do the parameterization over $SO(N)$.

5.9 Examples

5.9.1 $N = 2$

Example: Consider the case where we are given $\{\psi_i\}_{i=1}^2 \subset H = \mathbb{R}^2$ and corresponding nonnegative weights $\{\rho_i\}_{i=1}^2$. We want to find the orthonormal system $\{\widetilde{e}_i\}_{i=1}^2$ that minimizes P_e. $SO(2)$ is a 1-dimensional manifold. A parameterization of $SO(2)$ can be given for all $q \in [0, 2\pi)$,

$$\Theta(q) = \begin{pmatrix} \cos(q) & -\sin(q) \\ \sin(q) & \cos(q) \end{pmatrix}.$$

Let $\{w_i\}_{i=1}^2$ be the standard orthonormal basis for $H = \mathbb{R}^2$. We construct the parameterized orthonormal set by defining

$$\widetilde{e}_1(q) = \Theta(q)w_1 = \begin{pmatrix} \cos(q) \\ \sin(q) \end{pmatrix}, \ \widetilde{e}_2(q) = \Theta(q)w_2 = \begin{pmatrix} -\sin(q) \\ \cos(q) \end{pmatrix}.$$

Now assume $q(t)$ is a function of time. We have

$$\begin{aligned} \dot{\widetilde{e}}_1(q(t)) &= \frac{d}{dt}\begin{pmatrix} \cos(q(t)) \\ \sin(q(t)) \end{pmatrix} = \begin{pmatrix} -\sin(q(t))\dot{q}(t) \\ \cos(q(t))\dot{q}(t) \end{pmatrix} = \widetilde{e}_2(q(t))\dot{q}(t) \\ \dot{\widetilde{e}}_2(q(t)) &= \frac{d}{dt}\begin{pmatrix} -\sin(q(t)) \\ \cos(q(t)) \end{pmatrix} = \begin{pmatrix} -\cos(q(t))\dot{q}(t) \\ -\sin(q(t))\dot{q}(t) \end{pmatrix} = -\widetilde{e}_1(q(t))\dot{q}(t) \end{aligned}$$

$$
\begin{aligned}
T &= \frac{1}{2}\sum_{i=1}^{2}\|\dot{e}_i(q(t))\|^2 = \frac{1}{2}[\dot{q}(t)^2 + \dot{q}(t)^2] = \dot{q}(t)^2 \\
\frac{d}{dt}\frac{d}{d\dot{q}}T &= \frac{d}{dt}2\dot{q}(t) = 2\ddot{q}(t) \\
\frac{d}{dq}\widetilde{e}_1(q(t)) &= \frac{d}{dq}\begin{pmatrix} \cos(q(t)) \\ \sin(q(t)) \end{pmatrix} = \begin{pmatrix} -\sin(q(t)) \\ \cos(q(t)) \end{pmatrix} = \widetilde{e}_2(q(t)) \\
\frac{d}{dq}\widetilde{e}_2(q(t)) &= \frac{d}{dq}\begin{pmatrix} -\sin(q(t)) \\ \cos(q(t)) \end{pmatrix} = \begin{pmatrix} -\cos(q(t)) \\ -\sin(q(t)) \end{pmatrix} = -\widetilde{e}_1(q(t))
\end{aligned}
$$

So our Lagrangian can be written as

$$L = T - P_e = \dot{q}^2 - \sum_{i=1}^{2}\rho_i[1 - \langle\psi_i, \widetilde{e}_i\rangle^2]$$

and our equation of motion is given by

$$\frac{d}{dt}\frac{dT}{d\dot{q}} = -2\sum_{i=1}^{2}\rho_i\langle\psi_i, \widetilde{e}_i(q(t))\rangle\left\langle\psi_i, \frac{d}{dq}\widetilde{e}_i(q(t))\right\rangle.$$

Plugging in the expressions for T and the derivatives of $\widetilde{e}_i$ gives us

$$2\ddot{q} = 2[\rho_2\langle x_2, \widetilde{e}_2(q(t))\rangle\langle x_2, \widetilde{e}_1(q(t))\rangle - \rho_1\langle x_1, \widetilde{e}_1(q(t))\rangle\langle x_1, \widetilde{e}_2(q(t))\rangle]$$

which is a second-order ordinary differential equation.

In $\mathbb{R}^2$, the minimizer can be explicitly found. To simplify the notation, we write

$$\widetilde{e}_i = \widetilde{e}_i(q(t)) \text{ and } q = q(t).$$

We can write our given vectors as

$$\psi_1 = \begin{pmatrix} a \\ b \end{pmatrix},\ \psi_2 = \begin{pmatrix} c \\ d \end{pmatrix}.$$

We get,

$$\begin{aligned}
\sum_{i=1}^{2} \rho_i \langle \widetilde{e}_i, \psi_i \rangle^2 &= \rho_1 (a\cos(q) + b\sin(q))^2 + \rho_2 (-c\sin(q) + d\cos(q))^2 \\
&= (\rho_1 a^2 + \rho_2 d^2)\cos^2(q) + 2(\rho_1 ab - \rho_2 cd)\cos(q)\sin(q) + (\rho_1 b^2 + \rho_2 c^2)\sin^2(q) \\
&= (\rho_1 a^2 + \rho_2 d^2 - \rho_1 b^2 - \rho_2 c^2)\cos^2(q) + 2(\rho_1 ab - \rho_2 cd)\cos(q)\sin(q) + (\rho_1 b^2 + \rho_2 c^2) \\
&= \alpha\cos^2(q) + \beta\cos(q)\sin(q) + \gamma
\end{aligned}$$

where

$$\begin{aligned}
\alpha &= (\rho_1 a^2 + \rho_2 d^2 - \rho_1 b^2 - \rho_2 c^2) \\
\beta &= 2(\rho_1 ab - \rho_2 cd) \\
\gamma &= (\rho_1 b^2 + \rho_2 c^2).
\end{aligned}$$

So we have,

$$\begin{aligned}
\sum_{i=1}^{2} \rho_i \langle \widetilde{e}_i, \psi_i \rangle^2 &= \cos(q)[\alpha\cos(q) + \beta\sin(q)] + \gamma \\
&= \sqrt{\alpha^2 + \beta^2}\cos(q)[\cos(\xi)\cos(q) + \sin(\xi)\sin(q)] + \gamma
\end{aligned}$$

where $\xi \in [0, 2\pi)$ such that

$$\cos(\xi) = \frac{\alpha}{\sqrt{\alpha^2 + \beta^2}},\ \sin(\xi) = \frac{\beta}{\sqrt{\alpha^2 + \beta^2}}.$$

Using the relation

$$\cos(A)\cos(A + B) = \frac{1}{2}[\cos(2A + B) + \cos(B)]$$

we get,

$$\begin{aligned}
\sum_{i=1}^{2} \rho_i \langle \widetilde{e}_i, \psi_i \rangle^2 &= \sqrt{\alpha^2 + \beta^2}\cos(q)[\cos(\xi)\cos(q) + \sin(\xi)\sin(q)] + \gamma \\
&= \sqrt{\alpha^2 + \beta^2}\cos(q)[\cos(q - \xi)] + \gamma \\
&= \frac{\sqrt{\alpha^2 + \beta^2}}{2}[\cos(2q - \xi) + \cos(\xi)] + \gamma.
\end{aligned}$$

So to minimizes the error P_e, we want to maximize $\sum_{i=1}^{2} \rho_i \langle \psi_i, \widetilde{e}_i \rangle^2$ which occurs exactly when $q = \xi/2 + \pi n$ for some integer n. We can write

$$q = \frac{1}{2} \tan^{-1} \left(\frac{2(\rho_1 ab - \rho_2 cd)}{(\rho_1 a^2 + \rho_2 d^2 - \rho_1 b^2 - \rho_2 c^2)} \right) + \pi n$$

for some $n \in \mathbb{N}$.

Chapter 6

Least-squares error

Other authors have solved the problem by considering different types of error. In this chapter, we consider a least-squares error. Given a d-dimensional Hilbert space $H = \mathbb{K}^d$, where $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$, and a set $\{\psi_i\}_{i=1}^N \subset H$ with corresponding positive weights $\{\rho_i\}_{i=1}^N \subset \mathbb{R}$, we want to find a tight-frame $\{e_i\}_{i=1}^N$ for H that minimizes the error

$$E = \sum_{i=1}^{N} \rho_i \|\psi_i - e_i\|^2.$$

We first present a solution in section 6.1 without weights, that is for all $i = 1, \ldots, N$ we set

$$\rho_i = 1.$$

We then present a solution for general weights in section 6.2 when the given vectors $\{\psi_i\}_{i=1}^N$ are linearly independent. These solutions are based on work done in [13]. In section 6.2.1 we present some original work and analyze the case when $\{\psi_i\}_{i=1}^N$ is linearly dependent and develop a method of obtaining a 1-tight frame $\{e_i\}_{I=1}^N$ that has a small weighted least-squares error and provide bounds for the error. In sections 6.3 we construct examples of 1-tight frames that minimize the least-squares error. In section 6.4 we borrow material from [11] and introduce geometrically uniform frames and illustrate some of their properties. Finally in section 6.5, we show the result from [12] which states that if the given vectors $\{\psi_i\}_{i=1}^N$ are geometrically

uniform, then the tight-frame that minimizes the non-weighted least-squares error also minimizes the equal-weighted probability of a detection error P_e.

6.1 Non-weight case

Theorem 6.1. *Let H be a d-dimensional Hilbert space with $\{\psi_i\}_{i=1}^N \subset H$ such that $span\{\psi_i\}_{i=1}^N = H$. Let $A \in \mathbb{R}$ with $A > 0$ be given. Then there exists a unique tight frame $\{e_i\}_{i=1}^N$ for H with frame constant A^2 such that*

$$E = \sum_{i=1}^{N} \|\psi_i - e_i\|^2 = \inf\left\{\sum_{i=1}^{N} \|\psi_i - \xi_i\|^2 : \{\xi_i\}_{i=1}^N \text{ a } A^2\text{-tight frame on } H\right\} = \sum_{i=1}^{d} (\sigma_i - A)^2$$

where $\{\sigma_i\}_{i=1}^d$ are the singular values of the corresponding Bessel map matrix for the sequence $\{\psi_i\}_{i=1}^N$.

The proof of this will be constructive. (1) We will first assume that we have a A^2 tight frame $\{\xi_i\}_{i=1}^N$. We then plug $\{\xi_i\}_{i=1}^N$ into the expression for E, and then minimize E which gives restrictions on $\{\xi_i\}_{i=1}^N$. These restrictions will completely determine $\{\xi_i\}_{i=1}^N$.

Note that minimizing E and trying to determine $\{\xi_i\}_{i=1}^N$ would be much easier if $\{\xi_i\}_{i=1}^N$ were an orthogonal set instead of a tight frame. (2) We change this problem into an equivalent one by replacing $\{\xi_i\}$ by an equal-normed orthogonal set $\{a_i\}$. The error would then become

$$E = \sum_{i=1}^{N} \|\psi_i - \xi_i\|^2 = \sum_{i=1}^{N} \|\psi_i' - a_i\|^2.$$

We have N vectors ξ_i and $dim(H) = d \leq N$. So we cannot replace $\{\xi_i\}_{i=1}^N$ by an equal-normed orthogonal set $\{a_i\}_{i=1}^N$ since we would have more orthogonal

vectors than we have dimensions. So before we do (2), we need to (3) change the sum in the error E from a sum of N terms into a sum of d terms.

$$E = \sum_{i=1}^{d} \|\psi_i' - a_i\|^2.$$

(4) Finally we minimize E which determines $\{a_i\}_{i=1}^d$ and in turn determines $\{\xi_i\}_{i=1}^N$. We now present the proof.

Proof. Let $\{e_i\}_{i=1}^d$ be an orthonormal basis for H. Define $\Psi \in \mathcal{M}(N \times d)$ as the Bessel map matrix of the set $\{\psi_i\}_{i=1}^N$, i.e. Ψ is the matrix whose ith row is ψ_i^*, for $1 \leq i \leq N$ with respect to the basis $\{e_i\}_{i=1}^d$, where $*$ denotes complex conjugation. We can write this as

$$\Psi = \begin{pmatrix} — & \psi_1^* & — \\ & \vdots & \\ — & \psi_N^* & — \end{pmatrix}.$$

(1) Let $\{\xi_i\}_{i=1}^N$ be a A^2-tight frame for H and define $F \in \mathcal{M}(N \times d)$ as the Bessel map matrix corresponding to $\{\xi_i\}_{i=1}^N$, i.e. the matrix whose ith row is ξ_i^* with respect to the basis $\{e_i\}_{i=1}^d$. We can write this as

$$F = \begin{pmatrix} — & \xi_1^* & — \\ & \vdots & \\ — & \xi_N^* & — \end{pmatrix}.$$

As in step (3), we want to change the number of things being summed in E. The error can be written as

$$E = \sum_{i=1}^{N} \langle \psi_i - \xi_i, \psi_i - \xi_i \rangle$$

$$
\begin{aligned}
&= Tr((\Psi - F)(\Psi - F)^*) \\
&= Tr((\Psi - F)^*(\Psi - F)),
\end{aligned}
$$

where $(\Psi - F)^*(\Psi - F) \in \mathcal{M}(d \times d)$, hence the trace now becomes a sum of d terms. We now further simply this expression. We take the singular value decomposition $\Psi^* = U\Sigma V^* = \sum_{i=1}^{d} \sigma_i u_i v_i^*$. Since $U \in \mathcal{M}(d \times d)$ and we know that similar matrices have the same trace, we have

$$
\begin{aligned}
E &= Tr((\Psi - F)^*(\Psi - F)) \\
&= Tr(U^*(\Psi - F)^*(\Psi - F)U) \\
&= \sum_{i=1}^{d} \langle d_i, d_i \rangle
\end{aligned}
$$

where $d_i = (\Psi - F)u_i$. We further simplify d_i. Now,

$$
\Psi u_i = \sum_{k=1}^{d} \sigma_k v_k u_k^* u_i = \sum_{k=1}^{d} \sigma_k v_k \langle u_k, u_i \rangle = \sigma_i v_i
$$

since $\{u_i\}_{i=1}^{d}$ are the columns of a unitary matrix, hence are orthonormal. So,

$$
d_i = \sigma_i v_i - a_i
$$

where $a_i = Fu_i$. (2) We will now show the $\{a_i\}_{i=1}^{d}$ are an equal-normed orthogonal set. Since F is the Bessel map of a tight frame, by Theorem 2.1 we know that the corresponding frame operator satisfies

$$
F^*F = S = A^2 I_H
$$

hence

$$
\begin{aligned}
\langle a_i, a_k \rangle &= \langle Fu_i, Fu_k \rangle = \langle u_i, F^*Fu_k \rangle \\
&= A^2 \langle u_i, I_H u_k \rangle = A^2 \delta_{ik}
\end{aligned}
$$

where

$$\delta_{ik} = \begin{cases} 1 \text{ if } i = k \\ \\ 0 \text{ if } i \neq k \end{cases} .$$

So, $\{a_i\}_{i=1}^d$ are a A-normed orthogonal set.

We now minimize E and determine $\{a_i\}_{i=1}^d$. We expand the error E and get

$$\begin{aligned} E &= \sum_{i=1}^d \langle d_i, d_i \rangle = \sum_{i=1}^d \langle \sigma_i v_i - a_i, \sigma_i v_i - a_i \rangle \\ &= \sum_{i=1}^d \left[\langle \sigma_i v_i, \sigma_i v_i \rangle - \langle \sigma_i v_i, a_i \rangle - \langle a_i, \sigma_i v_i \rangle + \langle a_i, a_i \rangle \right] \\ &= \sum_{i=1}^d \left[\sigma_i^2 - 2\Re\{\langle a_i, \sigma_i v_i \rangle\} + A^2 \right] \end{aligned}$$

where we have used the fact that $\{v_i\}_{i=1}^N$ are columns of a unitary matrix, hence are orthonormal, and $\{a_i\}_{i=1}^d$ are a A-normed orthogonal set. Note that σ_i, v_i and A are all given in the hypothesis of the theorem. Hence we only have control over the A-normed orthogonal set $\{a_i\}_{i=1}^d$. In order to minimize E we need to make $\Re\{\langle a_i, \sigma_i v_i \rangle\}$ as large as possible. Note that we have the upper bound

$$\begin{aligned} \Re\{\langle a_i, \sigma_i v_i \rangle\} &\leq \sigma_i |\langle a_i, v_i \rangle| \\ &\leq \sigma_i \langle a_i, a_i \rangle^{1/2} \langle v_i, v_i \rangle^{1/2} = \sigma_i A \end{aligned}$$

and we have equality if and only if $a_i = Av_i$. So the A-normed orthogonal set that minimizes the error E is just $\{Av_i\}_{i=1}^d$.

Finally, we determine $\{\xi_i\}_{i=1}^N$. The Bessel map matrix F that minimizes the error E must satisfy for all $1 \leq i \leq d$,

$$Fu_i = a_i = Av_i$$

and with a bit of algebra this implies that

$$F = A \sum_{i=1}^{d} v_i u_i^*$$

with corresponding error

$$E = \sum_{i=1}^{d} \left[\sigma_i^2 - 2\sigma_i A + A^2 \right] = \sum_{i=1}^{d} (\sigma_i - A)^2.$$

With a bit of work, we can also write

$$F = [A\Psi^*((\Psi\Psi^*)^{1/2})^\dagger]^* = A[((\Psi^*\Psi)^{1/2})^\dagger \Psi^*]^*$$

where † corresponds to the Penrose-Moore pseudo inverse. See the appendix for the definition of the Penrose-Moore pseudo inverse. The set $\{\xi_i\}_{i=1}^N$ that minimizes E are the columns of the matrix F^*.

We have shown that if $\{\xi_i\}_{i=1}^N$ minimizes E, then the corresponding matrix F must be of the form

$$F = [A\Psi^*((\Psi\Psi^*)^{1/2})^\dagger]^* = A[((\Psi^*\Psi)^{1/2})^\dagger \Psi^*]^*.$$

Since this uniquely determines the matrix F, this shows that the set $\{\xi_i\}_{i=1}^N$ that minimizes E is unique. □

Remark. If we are not constrained to keep the frame constant A fixed, we can further decrease the error E by setting

$$A = \frac{1}{r} \sum_{i=1}^{d} \sigma_i$$

which is not hard to show by using some calculus.

6.2 Weighted case

Lemma 6.1. *Let H be a separable Hilbert space. Let $K \subset \mathbb{Z}$ and $\{\psi_i\}_{i\in K} \subset H$ be a set of normalized vectors with corresponding positive weights $\{\rho_i\}_{i\in K}$. Suppose that $\{e_i\}_{i\in K} \subset H$ is a normalized set of vectors. Then the least-squares error becomes,*

$$E = \sum_{i\in K} \rho_i \|\psi_i - e_i\|^2 = \sum_{i\in K} \|\rho_i\psi_i - e_i\|^2 - \sum_{i\in K}(1-\rho_i)^2.$$

Proof. For a given $i \in K$ we have,

$$\begin{aligned}
\|\rho_i\psi_i - e_i\|^2 + (1-\rho_i)(\rho_i\|\psi_i\|^2 - \|e_i\|^2) &= \rho_i^2\|\psi_i\|^2 - 2\rho_i\Re(\langle\psi_i, e_i\rangle) + \|e_i\|^2 \\
&\quad -\|e_i\|^2 + \rho_i\|e_i\|^2 + \rho_i\|\psi_i\|^2 - \rho_i^2\|\psi_i\|^2 \\
&= \rho_i\|e_i\|^2 - 2\rho_i\Re(\langle\psi_i, e_i\rangle) + \rho_i\|\psi_i\|^2 \\
&= \rho_i\|\psi_i - e_i\|^2.
\end{aligned}$$

So the weighted-error becomes

$$E = \sum_{i\in K} \rho_i \|\psi_i - e_i\|^2 = \sum_{i\in K} \|\rho_i\psi_i - e_i\|^2 - \sum_{i\in K}(1-\rho_i)(\rho_i\|\psi_i\|^2 - \|e_i\|^2).$$

Now using the fact that $\{\psi_i\}_{i\in K}$ and $\{e_i\}_{i\in K}$ are normalized, we have

$$E = \sum_{i\in K} \rho_i \|\psi_i - e_i\|^2 = \sum_{i\in K} \|\rho_i\psi_i - e_i\|^2 - \sum_{i\in K}(1-\rho_i)^2$$

which is what we wanted. □

From this lemma, it appears that the problem of finding a tight frame that minimizes the weighted least squares error E reduces to a non-weighted problem. Given a set of normalized vectors $\{\psi_i\}_{i=1}^N \subset H$ we consider a modified error defined

for all tight frames $\{e_i\}_{i=1}^N \subset H$ by

$$\tilde{E} = \sum_{i=1}^{N} \|\rho_i \psi_i - e_i\|^2.$$

We then apply Theorem 6.1 to find the unique tight frame that minimizes $\tilde{E}$, and by the lemma if this tight frame is normalized then we minimizes E. However, the constructed tight frame of Theorem 6.1 is not necessarily normalized. In the case when $N = d$, the resulting 1-tight frame construction of Theorem 6.1 is normalized, so we have the following theorem.

Theorem 6.2. *Let H be a d-dimensional Hilbert space, and $\{\psi_i\}_{i=1}^d \subset H$ a normalized linearly independent set with corresponding positive weights $\{\rho_i\}_{i=1}^d$. Then there exists a unique normalized tight frame $\{e_i\}_{i=1}^d$ with frame constant 1 that satisfies*

$$E = \sum_{i=1}^{d} \rho_i \|\psi_i - e_i\|^2 = \inf \left\{ \sum_{i=1}^{d} \rho_i \|\psi_i - \xi_i\|^2 : \{\xi_i\}_{i=1}^d \text{ a 1-tight frame} \right\}.$$

Proof. Consider the error defined for all tight frames $\{\xi_i\}_{i=1}^d$ by

$$\tilde{E} = \sum_{i=1}^{d} \|\rho_i \psi_i - \xi_i\|^2.$$

By Theorem 6.1 there exists a unique 1-tight frame $\{e_i\}_{i=1}^d$ that minimizes $\tilde{E}$ over all other tight frames with frame constant 1.

Let $\{e_i\}_{i=1}^d$ be an orthonormal basis for H and let $\Psi \in \mathcal{M}(d \times d)$ be the matrix corresponding to the Bessel map of $\{\rho_i \psi_i\}_{i=1}^d$. Then take the singular value decomposition of $\Psi^* = U \Sigma V^*$. By the proof of Theorem 6.1, we know the Bessel map $F \in \mathcal{M}(d \times d)$ corresponding to the 1-tight frame $\{e_i\}_{i=1}^d$ that minimizes $\tilde{E}$ is given by

$$F = VU^*.$$

Since V and U are $d \times d$ unitary matrices, it follows that F^* is a $d \times d$ unitary matrix. Since the columns of F^* are the components of the tight frame $\{e_i\}_{i=1}^d$ with respect to the basis $\{e_i\}_{i=1}^d$, it follows that $\{e_i\}_{i=1}^d$ are normalized.

We can now apply the previous lemma and write the error as

$$E = \sum_{i \in K} \rho_i \|\psi_i - e_i\|^2 = \tilde{E} - \sum_{i \in K} (1 - \rho_i)^2.$$

Since $\{e_i\}_{i=1}^d$ is the unique tight frame with frame constant 1 that minimizes $\tilde{E}$, we see that by the above expression that $\{e_i\}_{i=1}^d$ is also the unique tight frame with frame constant 1 that minimizes E. □

6.2.1 Linearly dependent case

In this section, we analyze the weighted least-squares problem case when the given set $\{\psi_i\}_{i=1}^N \subset H$ are linearly dependent and $span\{\psi_i\}_{i=1}^N = H$. The idea used here is to perturb the set by ϵ so that $\{\psi_i(\epsilon)\}_{i=1}^N \subset H'$ is a linearly independent set in some enlarged Hilbert space H' for $\epsilon > 0$ and $\{\psi_i(0)\}_{i=1}^N = \{\psi_i\}_{i=1}^N$. We then find the optimal tight frame $\{e_i(\epsilon)\}_{i=1}^N \subset H'$ corresponding to this vector set. We then take the limit $\epsilon \to 0$ and hope $\{e_i\}_{i=1}^N = \lim_{\epsilon \to 0} \{e_i(\epsilon)\}_{i=1}^N$ has properties that we desire.

Suppose we have a set of normalized linearly dependent vectors $\{\psi_i\}_{i=1}^N \subset H$ such that H is d-dimensional. We construct the $d \times N$ matrix,

$$\Psi = \begin{pmatrix} | & & | \\ \psi_1 & \dots & \psi_N \\ | & & | \end{pmatrix}.$$

Consider the $(N+d) \times N$ matrix

$$\Psi(\epsilon) = \begin{pmatrix} | & & | \\ \psi_1 & \dots & \psi_N \\ | & & | \\ \epsilon & & 0 \\ & \ddots & \\ 0 & & \epsilon \end{pmatrix}.$$

The columns of $\Psi(\epsilon)$ can be interpreted as the linearly independent equal normed perturbed vectors $\{\psi_i(\epsilon)\}_{i=1}^N \subset H'$ where H' is an $(N+d)$-dimensional Hilbert space. In fact, for all $i = 1, \dots, N$, $\|\psi_i(\epsilon)\|^2 = 1 + \epsilon^2$. We want to find the 1-tight frame $\{e_i(\epsilon)\}_{i=1}^N$ for the $span\{\psi_i(\epsilon)\}_{i=1}^N$ that minimizes the weighted least-squares error. From Theorem 6.2, we know that $\{e_i(\epsilon)\}_{i=1}^N$ is orthonormal, and using the expression from Lemma 6.1 we write the error as

$$\begin{aligned} E &= \sum_{i=1}^{d} \rho_i \|\psi_i(\epsilon) - e_i(\epsilon)\|^2 \\ &= \sum_{i=1}^{d} \|\rho_i \psi_i(\epsilon) - e_i(\epsilon)\|^2 + \sum_{i=1}^{d} (1-\rho_i)(\rho_i \|\psi_i(\epsilon)\|^2 - \|e_i(\epsilon)\|^2) \\ &= \sum_{i=1}^{d} \|\rho_i \psi_i(\epsilon) - e_i(\epsilon)\|^2 + \sum_{i=1}^{d} (1-\rho_i)(\rho_i(1+\epsilon^2) - 1). \end{aligned}$$

So the problem reduces to finding the closest 1-tight frame $\{e_i(\epsilon)\}$ to the set of

vectors $\{\tilde{\psi}_i(\epsilon)\} = \{\rho_i \psi_i(\epsilon)\}$. So, we consider the $(N+d) \times N$ matrix

$$\tilde{\Psi}(\epsilon) = \begin{pmatrix} | & & | \\ \rho_1 \tilde{\psi}_1(\epsilon) & \dots & \rho_N \tilde{\psi}_N(\epsilon) \\ | & & | \end{pmatrix} = \begin{pmatrix} | & & | \\ \rho_1 \psi_1 & \dots & \rho_N \psi_N \\ | & & | \\ \rho_1 \epsilon & & 0 \\ & \ddots & \\ 0 & & \rho_N \epsilon \end{pmatrix}.$$

To simplify the analysis, we omit the weights ρ_i on the ϵ terms, since we plan to take the limit as $\epsilon \to 0$. We have,

$$\tilde{\Psi}(\epsilon) = \begin{pmatrix} | & & | \\ \rho_1 \psi_1 & \dots & \rho_N \psi_N \\ | & & | \\ \epsilon & & 0 \\ & \ddots & \\ 0 & & \epsilon \end{pmatrix}.$$

Recall that the corresponding tight-frame matrix whose columns are the tight-frame vectors that minimize the least-squares error is

$$M(\epsilon) = \tilde{\Psi}(\epsilon)((\tilde{\Psi}(\epsilon)^* \tilde{\Psi}(\epsilon))^{1/2})^\dagger$$

where † corresponds to the Moore-Penrose pseudo-inverse.

It is easy to check that

$$\tilde{\Psi}(\epsilon)^* \tilde{\Psi}(\epsilon) = \tilde{\Psi}^*(0) \tilde{\Psi}(0) + \epsilon^2 I_N.$$

Since $\tilde{\Psi}^*(0)\tilde{\Psi}(0)$ is a self-adjoint positive $N \times N$ matrix, there exists an $N \times N$ unitary matrix V such that

$$\tilde{\Psi}(\epsilon)^*\tilde{\Psi}(\epsilon) = V \begin{pmatrix} \sigma_1^2 + \epsilon^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_N^2 + \epsilon^2 \end{pmatrix} V^*$$

where the $\{\sigma_i\}$ are the singular values of $\tilde{\Psi}$ (i.e. $\{\sigma_i^2\}$ are the nonnegative eigenvalues of $\Psi^*\Psi$). Note that V is independent of ϵ, since the columns of V consists of the orthonormal eigenvectors of $\Psi(\epsilon)^*\Psi(\epsilon)$, which by the above expression can be seen to be the orthonormal eigenvectors of $\tilde{\Psi}^*(0)\tilde{\Psi}(0)$ which are independent of ϵ.

We now take the "square-root" and pseudo inverse to get

$$((\tilde{\Psi}(\epsilon)^*\tilde{\Psi}(\epsilon))^{1/2})^\dagger = V \begin{pmatrix} \frac{1}{\sqrt{\sigma_1^2+\epsilon^2}} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{\sqrt{\sigma_N^2+\epsilon^2}} \end{pmatrix} V^*.$$

Recall that we want to analyze

$$\lim_{\epsilon \to 0} M(\epsilon) = \lim_{\epsilon \to 0} \tilde{\Psi}(\epsilon)((\tilde{\Psi}(\epsilon)^*\tilde{\Psi}(\epsilon))^{1/2})^\dagger.$$

The singular value decomposition of $\tilde{\Psi}(\epsilon)$ is of the form

$$\tilde{\Psi}(\epsilon) = U(\epsilon) \begin{pmatrix} \sqrt{\sigma_1^2 + \epsilon^2} & & 0 \\ & \ddots & \\ 0 & & \sqrt{\sigma_N^2 + \epsilon^2} \\ 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{pmatrix} V^*$$

where $U \in \mathcal{M}((d+N)\times(d+N))$, $V \in \mathcal{M}(N\times N)$ are unitary and the diagonal matrix has dimensions $(d+N)\times N$. We obtain,

$$
\begin{aligned}
M(\epsilon) &= \tilde{\Psi}(\epsilon)((\tilde{\Psi}(\epsilon)^*\tilde{\Psi}(\epsilon))^{1/2})^\dagger \\
&= U(\epsilon)\begin{pmatrix} \sqrt{\sigma_1^2+\epsilon^2} & & 0 \\ & \ddots & \\ 0 & & \sqrt{\sigma_N^2+\epsilon^2} \\ 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{pmatrix} V^*V \begin{pmatrix} \frac{1}{\sqrt{\sigma_1^2+\epsilon^2}} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{\sqrt{\sigma_N^2+\epsilon^2}} \end{pmatrix} V^* \\
&= U(\epsilon)\begin{pmatrix} \sqrt{\sigma_1^2+\epsilon^2} & & 0 \\ & \ddots & \\ 0 & & \sqrt{\sigma_N^2+\epsilon^2} \\ 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{\sigma_1^2+\epsilon^2}} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{\sqrt{\sigma_N^2+\epsilon^2}} \end{pmatrix} V^* \\
&= U(\epsilon)\begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \\ 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{pmatrix} V^*.
\end{aligned}
$$

We want $\lim_{\epsilon\to 0} M(\epsilon)$ so we need to find $\lim_{\epsilon\to 0} U(\epsilon)$. Proving that this limit exists

is not an easy task. However, it is not hard to show that $U(\epsilon = 0)$ exists. First, $\tilde{\Psi}(\epsilon = 0)\tilde{\Psi}(\epsilon = 0)^*$ is a self-adjoint $N \times N$ matrix, hence there exists a set of N orthonormal eigenvectors. The matrix $U(\epsilon = 0)$ is the matrix whose columns are the orthonormal eigenvectors of $\tilde{\Psi}(\epsilon = 0)\tilde{\Psi}(\epsilon = 0)^*$, hence $U(\epsilon = 0)$ exists.

Note that

$$\begin{aligned} M(\epsilon)^* M(\epsilon) &= V \begin{pmatrix} 1 & \dots & 0 & 0 & \dots & 0 \\ & \ddots & & \vdots & \ddots & \vdots \\ 0 & & 1 & 0 & \dots & 0 \end{pmatrix} U(\epsilon)^* U(\epsilon) \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \\ 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{pmatrix} V^* \\ &= V \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix} V^* = VV^* = I_N. \end{aligned}$$

So the columns of $M(\epsilon)$ are orthonormal for all ϵ, hence $\{e_i(\epsilon = 0)\}_{i=1}^N$ is an orthonormal set. We shall show that $\{e_i(\epsilon = 0)\}_{i=1}^N \subset H'$ minimizes the weighted least-squares error E over all other N element orthonormal sets in H'.

Lemma 6.2. *Let $\{\psi_i\}_{i=1}^N \subset H$, where H is a d-dimensional Hilbert space, and $\{\rho_i\}_{i=1}^N \subset \mathbb{R}$ be given where the weights have the property that*

$$\sum_{i=1}^{N} \rho_i = 1$$

and span$\{\psi_i\}_{i=1}^N = H$. Let H' be a larger Hilbert space such that $H \subset H'$. Then the

orthonormal set $\{e_i\}_{i=1}^N \subset H'$ *minimizes the weighted error*

$$E = \sum_{i=1}^{N} \rho_i \|\psi_i - e_i\|^2$$

over all other N*-element orthnormal sets in* H' *if and only if* $\{P_H e_i\}_{i=1}^N$ *minimizes the error*

$$E' = \sum_{i=1}^{N} \|\rho_i \psi_i - P_H e_i\|^2$$

over all N*-element 1-tight frames for* H*, where* P_H *denotes the orthogonal projection onto* H*. Furthermore, the minimal error for the orthonormal set in* H *must satisfy*

$$E = \sum_{i=1}^{N} \rho_i \|\psi_i - e_i\|^2 = \sum_{i=1}^{d} (\sigma_i - 1)^2 - N + d + \sum_{i=1}^{N} (1 - \rho_i)^2$$

where $\{\sigma_i\}$ *are the singular values of the Bessel map matrix corresponding to the sequence* $\{\rho_i \psi_i\}_{i=1}^N$*.*

Proof. Assume that $\{e_i\}_{i=1}^N$ is an orthonormal set in H'. By lemma 6.1, the error is can be written as,

$$E = \sum_{i=1}^{N} \rho_i \|\psi_i - e_i\|^2 = \sum_{i=1}^{N} \|\rho_i \psi_i - e_i\|^2 + \sum_{i=1}^{N} (1 - \rho_i)^2.$$

So the orthonormal set $\{e_i\}_{i=1}^N$ also minimizes the non-weighted error

$$E' = \sum_{i=1}^{N} \|\rho_i \psi_i - e_i\|^2.$$

We decompose our orthogonal set as $e_i = e_i^H + e_i^\perp$ where $e_i^H = P_H e_i \in H$ and $e_i^\perp \perp \mathcal{U}$. Then the non-weighted error becomes,

$$\begin{aligned} E' &= \sum_{i=1}^{N} \|\rho_i \psi_i - e_i^H - e_i^\perp\|^2 \\ &= \sum_{i=1}^{N} (\|\rho_i \psi_i - e_i^H\|^2 + \|e_i^\perp\|^2) \\ &= \sum_{i=1}^{N} \langle \rho_i \psi_i - e_i^H, \rho_i \psi_i - e_i^H \rangle - \sum_{i=1}^{N} \langle e_i^\perp, e_i^\perp \rangle. \end{aligned}$$

Note by Lemma 4.1, $\{e_i^H\}$ forms a 1-tight frame for H. Denote by S_H as the frame operator for the set $\{e_i^H\}$.

$$\begin{aligned}\sum_{i=1}^{N}\langle e_i^{\perp}, e_i^{\perp}\rangle &= \sum_{i=1}^{N}\langle e_i, e_i\rangle - \sum_{i=1}^{N}\langle e_i^H, e_i^H\rangle \\ &= N - Tr(S_H) = N - dim(H) = N - d.\end{aligned}$$

So the total error becomes,

$$E = \underbrace{\sum_{i=1}^{N}\|\rho_i\psi_i - e_i^H\|^2}_{=\tilde{E}} - N + dim(H) + \sum_{i=1}^{N}(1-\rho_i)^2.$$

So E is minimized if and only if E' is minimized, and the result is clear.

To get the expression for the minimal E, note that by Theorem 6.1, if $\{e_i^H\}_{i=1}^N$ minimizes E', then

$$E' = \sum_{i=1}^{d}(\sigma_i - 1)^2$$

where $\{\sigma_i\}_{i=1}^d$ are the singular values of the matrix whose columns are given by $\{\rho_i\psi_i\}_{i=1}^N$. Plugging E' into the expression for E gives us our result. □

Theorem 6.3. *The set $\{e_i(\epsilon = 0)\}_{i=1}^N \subset H'$ is the closest orthonormal set that minimizes the weighted error E over all other N-element orthonormal sets in H'.*

Proof. Consider the $d \times N$ matrix $\tilde{\Psi}$ defined by

$$\tilde{\Psi} = \begin{pmatrix} | & & | \\ \rho_1\psi_1 & \dots & \rho_N\psi_N \\ | & & | \end{pmatrix}.$$

By the construction, $\{e_i(\epsilon=0)\}_{i=1}^N$ are the columns of the matrix

$$M(\epsilon=0) \;=\; U(\epsilon=0)\begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \\ 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{pmatrix} V^*$$

where V is the $N \times N$ unitary matrix whose columns are the orthonormal eigenvectors of $\tilde{\Psi}^*(\epsilon=0)\tilde{\Psi}(\epsilon=0) = \tilde{\Psi}^*\tilde{\Psi}$, and $U(\epsilon=0)$ is the $(d+N)$ unitary matrix whose columns are the orthonormal eigenvectors of $\tilde{\Psi}(\epsilon=0)\tilde{\Psi}^*(\epsilon=0)$ and the diagonal matrix is a $(N+d)\times N$ matrix consisting of the $N \times N$ identity matrix on top and the rest zeroes.

Note that

$$\begin{aligned} \tilde{\Psi}(\epsilon)\tilde{\Psi}^*(\epsilon) &= \begin{pmatrix} \tilde{\Psi} \\ \\ \epsilon I_N \end{pmatrix}\begin{pmatrix} \tilde{\Psi}^* & \epsilon I_N \end{pmatrix} \\ &= \begin{pmatrix} \tilde{\Psi}\tilde{\Psi}^* & \epsilon\tilde{\Psi} \\ \\ \epsilon\tilde{\Psi}^* & \epsilon^2 I_N \end{pmatrix}. \end{aligned}$$

Since the columns of $U(\epsilon)$ consist of the orthonormal eigenvectors of $\tilde{\Psi}(\epsilon)\tilde{\Psi}^*(\epsilon)$, its not hard to show that

$$U(\epsilon=0) = \begin{pmatrix} U & 0 \\ \\ 0 & I_N \end{pmatrix}$$

where U is the unitary matrix whose columns are the orthonormal eigenvectors of the matrix $\Psi\Psi^*$ and I_N is the $N \times N$ identity matrix. Let P_H denote the orthogonal

projection onto H. Then,

$$\begin{aligned} P_H M(\epsilon = 0) &= P_H \begin{pmatrix} U & 0 \\ 0 & I_N \end{pmatrix} \begin{pmatrix} I_N \\ 0 \end{pmatrix} V^* \\ &= \begin{pmatrix} U & 0 \end{pmatrix} \begin{pmatrix} I_N \\ 0 \end{pmatrix} V^* \\ &= U \Sigma V^* \end{aligned}$$

where Σ is a $d \times N$ matrix with 1s on the diagonal and zeroes elsewhere. By the construction given in Theorem 6.1, it follows that the columns of $P_H M(\epsilon = 0)$ consist of the unique tight-frame that minimizes the error

$$E' = \sum_{i=1}^{N} \|\rho_I \psi_i - e_i\|^2$$

over all other 1-tight frames. It follows from Lemma 6.2 that the columns of $M(\epsilon = 0)$ is an orhogonal set that minimizes E over all other N-element orthogonal sets. $\square$

When the given set of vectors $\{\psi_i\}_{i=1}^N \subset H$ are linearly dependent, we can find the orthonormal set $\{e_i\}_{i=1}^N \subset H'$ that minimizes the error E over all other N-element orthonormal sets in H'. A natural question is whether $\{P_H e_i\}_{i=1}^N$ minimizes E over all other N-element 1-tight frames for H. A partial answer is given in Theorem 6.4 where upper and lower bounds are computed for the weighted least-squares error for the set $\{P_H e_i\}_{i=1}^N$.

We first present a lemma dealing with partial traces of matrices.

Lemma 6.3. *Let W be a self-adjoint operator on an N-dimensional Hilbert space H and $\{s_i\}_{i=1}^d \subset H$ be an orthonormal set where $d \leq N$. Let $\{\lambda_i\}_{i=1}^N \subset \mathbb{R}$ be the*

eigenvalues of W *ordered such that* $\lambda_i \geq \lambda_{i+1}$. *Then*

$$\sum_{i=N-d}^{N} \lambda_i \leq \sum_{i=1}^{d} \langle s_i, W s_i \rangle \leq \sum_{i=1}^{d} \lambda_i.$$

Proof. Set $U = span\{s_i\}_{i=1}^{d}$ and denote by P_U the orthogonal projection onto U.

Since W is self-adjoint, by the spectral theorem we can find an orthonormal basis $\{b_i\}_{i=1}^{N} \subset H$ for H of eigenvectors of W such that for all $x \in \mathbb{K}^N$,

$$Wx = \sum_{i=1}^{N} \lambda_i \langle x, b_i \rangle b_i.$$

We get,

$$\begin{aligned}
\sum_{i=1}^{d} \langle s_i, W s_i \rangle &= \sum_{i=1}^{d} \left\langle s_i, \sum_{j=1}^{N} \lambda_i \langle s_i, b_i \rangle b_i \right\rangle \\
&= \sum_{j=1}^{N} \sum_{i=1}^{d} \lambda_j \langle s_i, b_j \rangle \langle b_j, s_i \rangle \\
&= \sum_{j=1}^{N} \lambda_j \sum_{i=1}^{d} \langle b_j, s_i \rangle \langle s_i, b_j \rangle \\
&= \sum_{j=1}^{N} \lambda_j \left\langle b_j, \sum_{i=1}^{d} \langle b_j, s_i \rangle s_i, \right\rangle \\
&= \sum_{j=1}^{N} \lambda_j \langle b_j, P_U b_j \rangle.
\end{aligned}$$

Note that

$$\sum_{j=1}^{N} \langle b_j P_U b_j \rangle = d.$$

Define the sequence $\{\alpha_i\}_{i=1}^{N+1}$ recursively by

$$\begin{aligned}
\alpha_0 &= 0 \\
\alpha_i &= \alpha_{i-1} + \langle b_i, P_U b_i \rangle.
\end{aligned}$$

Consider the interval $[0, d)$ and partition the interval into N disjoint subintervals $\{I_n\}_{i=1}^N \subset [0, d]$ such that for $n = 1, \ldots, N$,

$$I_n = [\alpha_{n-1}, \alpha_n).$$

Define the step functions f, g, and h for all $x \in [0, d)$ by

$$\begin{aligned} f(x) &= \sum_{i=1}^{d} \lambda_i \mathbb{1}_{I_i}(x) \\ g(x) &= \sum_{i=1}^{N} \lambda_i \mathbb{1}_{[i-1,i]}(x) \\ h(x) &= \sum_{i=N-d}^{N} \lambda_i \mathbb{1}_{[i-1,i]}(x). \end{aligned}$$

With these definitions, its not hard to show that for all $x \in [0, d)$,

$$h(x) \leq f(x) \leq g(x).$$

We show $g \leq f$. Let $x \in [0, d)$. Then there exists an integer n such that $x \in I_n$ and $f(x) = \lambda_n$. Note that for all $i = 1, \ldots, N$,

$$|I_i| = \alpha_i - \alpha_{i-1} = \langle b_i, P_U b_i \rangle \leq 1.$$

Hence,

$$\bigcup_{i=1}^{n} I_i \subset [0, n)$$

so $x \in [0, n)$ and by the definition of the function g,

$$g(x) \geq \lambda_n = f(x).$$

A similar argument shows that $h \leq f$. Integrating the inequality gives us

$$\sum_{i=N-d}^{N} \lambda_i \leq \sum_{i=1}^{N} \lambda_i \langle b_i, P_U b_i \rangle \leq \sum_{i=1}^{d} \lambda_i.$$

But we have shown that

$$\sum_{i=1}^{d} \langle s_i, W s_i \rangle = \sum_{j=1}^{N} \lambda_i \langle b_i, P_U b_i \rangle,$$

hence we have our result. □

Lemma 6.4. *Let $\{e_i\}_{i=1}^N \subset H$ be a frame for a d-dimensional Hilbert space H. Given weights $\{\rho_i\}_{i=1}^N \subset \mathbb{R}$, define the weighted-frame operator $S' : H \mapsto H$ for all $x \in H$ by*

$$S'(x) = \sum_{i=1}^{N} \rho_i \langle e_i, x \rangle e_i.$$

Then,

$$Tr(S') = \sum_{i=1}^{N} \rho_i \|e_i\|^2.$$

Furthermore, if $\{e_i\}_{i=1}^N$ is a 1-tight frame for H and if the weights are ordered such that $\rho_i \geq \rho_{i+1}$, then

$$\sum_{i=N-d}^{N} \rho_i \leq Tr(S') \leq \sum_{i=1}^{d} \rho_i.$$

Proof. Assume $\{e_i\}_{i=1}^N \subset H$ is a frame. Let $\{b_i\}_{i=1}^d$ be a basis for H. Then,

$$\begin{aligned}
Tr(S') &= \sum_{l=1}^{d} \langle S' b_l, b_l \rangle \\
&= \sum_{l=1}^{d} \left\langle \sum_{i=1}^{N} \rho_i \langle b_l, e_i \rangle e_i, b_l \right\rangle \\
&= \sum_{l=1}^{d} \sum_{i=1}^{N} \rho_i \langle b_l, e_i \rangle \langle e_i, b_l \rangle \\
&= \sum_{i=1}^{N} \rho_i \sum_{l=1}^{d} |\langle b_l, e_i \rangle|^2 \\
&= \sum_{i=1}^{N} \rho_i \|e_i\|^2
\end{aligned}$$

which is what we wanted.

Now assume further that $\{e_i\}_{i=1}^N$ is a 1-tight frame for H. Let $\{b_i\}_{i=1}^d$ be an orthonormal basis for H. Consider the corresponding $N \times d$ Bessel map matrix with respect to the basis $\{b_i\}_{i=1}^d$

$$L = \begin{pmatrix} — & e_1^* & — \\ & \vdots & \\ — & e_N^* & — \end{pmatrix}$$

and the $N \times N$ weight matrix W defined as the diagonal matrix with diagonal elements $W_{ii} = \rho_i$,

$$W = \begin{pmatrix} \rho_1 & \dots & 0 \\ & \ddots & \\ 0 & \dots & \rho_N \end{pmatrix}.$$

Then, we can write the weighted frame operator as $S' = L^*WL$. We take the trace

$$\begin{aligned} tr(S') &= \sum_{i=1}^d \langle b_i, S'b_i \rangle \\ &= \sum_{i=1}^d \langle b_i, L^*WLb_i \rangle \\ &= \sum_{i=1}^d \langle Lb_i, WLb_i \rangle. \end{aligned}$$

Note that the set $\{Lb_i\}_{i=1}^d \subset \mathbb{C}^N$ is an orthonormal set since for any intergers $i = 1, \dots, d$ and $j = 1, \dots, d$ we have

$$\langle Lb_i, Lb_j \rangle_{\mathbb{C}^N} = \langle L^*Lb_i, b_j \rangle_H = \langle b_i, b_j \rangle_H = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

since the frame operator satisfies $S = L^*L = I_H$. By Lemma 6.3 we have

$$\sum_{i=N-d}^N \rho_i \leq \sum_{i=1}^d \langle Le_i, WLe_i \rangle \leq \sum_{i=1}^d \rho_i.$$

Since

$$tr(S') = \sum_{i=1}^{d} \langle Lb_i, WLb_i \rangle$$

we have our result. □

Theorem 6.4. *Assume that $\{e_i\}_{i=1}^N$ is the orthonormal set in H' that minimizes the weighted least-squares error E over all other N-element orthonormal sets in H'. Assume further that the weights are ordered so that $\rho_i \geq \rho_{i+1}$. Let $\{e_i^H\}_{i=1}^N = \{P_H e_i\}_{i=1}^N$ be the 1-tight frame for H obtained by projecting the orthonormal set $\{e_i\}_{i=1}^N$ into H. Then,*

$$E + \sum_{i=N-d}^{N} \rho_i - 1 \leq \sum_{i=1}^{N} \rho_i \|\psi_i - e_i^H\|^2 \leq E + \sum_{i=1}^{d} \rho_i - 1$$

where

$$E = \sum_{i=1}^{N} \rho_i \|\psi_i - e_i\|^2 = \sum_{i=1}^{d} (\sigma_i - 1)^2 - N + d + \sum_{i=1}^{N} (1 - \rho_i)^2$$

where $\{\sigma_i\}$ are the singular values of the matrix with columns $\{\rho_i \psi_i\}_{i=1}^N$.

Proof. Assume that $\{e_i\}_{i=1}^N$ is the orthonormal set in H' that minimizes E over all other N-element orthonormal sets in H'. We can write the error E as

$$E = \sum_{i=1}^{N} \rho_i \|\psi_i - e_i^H\|^2 + \sum_{i=1}^{N} \rho_i \|e_i^\perp\|^2$$

where $e_i^H \in H$ is the orthgonal projection into H and $e_i^\perp \perp H$. We can write the second term on the right as

$$\begin{aligned} \sum_{i=1}^{N} \rho_i \langle e_i^\perp, e_i^\perp \rangle &= \sum_{i=1}^{N} \rho_i \langle e_i, e_i \rangle - \sum_{i=1}^{N} \rho_i \langle e_i^H, e_i^H \rangle \\ &= \sum_{i=1}^{N} \rho_i - \sum_{i=1}^{N} \rho_i \langle e_i^H, e_i^H \rangle = 1 - \sum_{i=1}^{N} \rho_i \langle e_i^H, e_i^H \rangle. \end{aligned}$$

Since the projected set $\{e_i^H\}_{i=1}^N$ is a 1-tight frame for H, by Lemma 6.4 we have

$$-\sum_{i=1}^{d}\rho_i \leq -\sum_{i=1}^{N}\rho_i\|e_i^H\|^2 \leq -\sum_{i=N-d}^{N}\rho_i$$

hence,

$$1-\sum_{i=1}^{d}\rho_i\sum_{i=1}^{N}\rho_i\langle e_i^{\perp}, e_i^{\perp}\rangle \leq 1-\sum_{i=N-d}^{N}\rho_i.$$

Since

$$E-\sum_{i=1}^{N}\rho_i\|\psi_i-e_i^H\|^2=\sum_{i=1}^{N}\rho_i\|e_i^{\perp}\|^2$$

plugging this into the above inequality gives us

$$1-\sum_{i=1}^{d}\rho_i \leq E-\sum_{i=1}^{N}\rho_i\|\psi_i-e_i^H\|^2 \leq 1-\sum_{i=N-d}^{N}\rho_i.$$

Subtracting E and multiplying everything by -1 gives us

$$E+\sum_{i=N-d}^{N}\rho_i-1 \leq \sum_{i=1}^{N}\rho_i\|\psi_i-e_i^H\|^2 \leq E+\sum_{i=1}^{d}\rho_i-1$$

which is the inequality we wanted. Also, by Lemma 6.2 we have the corresponding expression for E. □

Note that since

$$\sum_{i=1}^{N}\rho_i=1$$

we always have

$$\sum_{i=1}^{d}\rho_i-1\leq 0.$$

Hence the projected tight-frame has a smaller error than the corresponding orthonormal set – that is, we have

$$\sum_{i=1}^{N}\rho_i\|\psi_i-e_i^H\|^2 \leq E=\sum_{i=1}^{N}\rho_i\|\psi_i-e_i\|^2.$$

6.3 Examples of computing the least-squares solution

Consider the Hilbert space $H = \mathbb{R}^2$ and suppose we have a set of vectors $\{\psi_i\}_{i=1}^3 \subset H$ such that $span\{\psi_i\} = H$. We want to compute the 1-tight frame for H that minimizes the error

$$E = \frac{1}{3}\sum_{i=1}^{3} \|\psi_i - e_i\|^2$$

over all other 3-element 1-tight frames. We construct the 2×3 matrix,

$$\Psi = \begin{pmatrix} | & | & | \\ \psi_1 & \psi_2 & \psi_3 \\ | & | & | \end{pmatrix}.$$

We perform the singular value decomposition of Ψ to get,

$$\Psi = U \underbrace{\begin{pmatrix} \sigma_1 & 0 & 0 \\ 0 & \sigma_2 & 0 \end{pmatrix}}_{\Sigma} V^*$$

where $U \in \mathcal{M}(2 \times 2)$ and $V \in \mathcal{M}(3 \times 3)$ are unitary matrices. The columns of U consist of the orthonormal eigenvectors of the self-adjoint matrix $\Psi\Psi^*$ and the columns of V consist of the orthonormal eigenvectors of the self-adjoint matrix $\Psi^*\Psi$. The eigenvalues of the 2×2 matrix $\Psi\Psi^*$ are the square of the singular values, hence we see that Ψ has at most two singular values. The closest 1-tight frame are the columns of the matrix

$$\begin{aligned} M &= ((\Psi\Psi^*)^{1/2})^\dagger \Psi \\ &= ((U\Sigma V^* V \Sigma^* U^*)^{1/2})^\dagger U\Sigma V^* \\ &= U((\Sigma\Sigma^*)^{1/2})^\dagger U^* U \Sigma V^* \end{aligned}$$

$$
\begin{aligned}
&= U((\Sigma\Sigma^*)^{1/2})^{\dagger}\Sigma V^* \\
&= U\left[\left(\begin{array}{ccc}\sigma_1 & 0 & 0\\ 0 & \sigma_2 & 0\end{array}\right)\left(\begin{array}{cc}\sigma_1 & 0\\ 0 & \sigma_2\\ 0 & 0\end{array}\right)\right]^{\dagger/2}\left(\begin{array}{ccc}\sigma_1 & 0 & 0\\ 0 & \sigma_2 & 0\end{array}\right)V^* \\
&= U\left[\left(\begin{array}{cc}\sigma_1^2 & 0\\ 0 & \sigma_2^2\end{array}\right)\right]^{\dagger/2}\left(\begin{array}{ccc}\sigma_1 & 0 & 0\\ 0 & \sigma_2 & 0\end{array}\right)V^* \\
&= U\left(\begin{array}{cc}\frac{1}{\sigma_1} & 0\\ 0 & \frac{1}{\sigma_2}\end{array}\right)\left(\begin{array}{ccc}\sigma_1 & 0 & 0\\ 0 & \sigma_2 & 0\end{array}\right)V^* \\
&= U\left(\begin{array}{ccc}1 & 0 & 0\\ 0 & 1 & 0\end{array}\right)V^*.
\end{aligned}
$$

In general, to compute the closest 1-tight frame matrix, we simply replace the singular values of Ψ by 1s.

6.3.1 Explicit example in $\mathbb{R}^2$

Consider the Hilbert space $H = \mathbb{R}^2$ and the vectors

$$\psi_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \psi_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \psi_3 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

with weights

$$\rho_1 = \rho_2 = \rho_3 = \frac{1}{3}.$$

We want to find the 1-tight frame $\{e_i\}_{i=1}^3$ that minimizes the least-squares error

$$E = \frac{1}{3}\sum_{i=1}^{3} \|\psi_i - e_i\|^2$$

over all other 3-element 1-tight frames. We construct the matrix

$$\Psi = \begin{pmatrix} | & | & | \\ \psi_1 & \psi_2 & \psi_3 \\ | & | & | \end{pmatrix} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}.$$

We now take the singular decomposition of Ψ. First we look at

$$\Psi\Psi^* = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}.$$

The eigenvalues of $\Psi\Psi^*$ are 1 and 3 with corresponding eigenvectors

$$\frac{1}{\sqrt{2}}\begin{pmatrix} -1 \\ 1 \end{pmatrix}, \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

These eigenvalues are the squares of the singular values of Ψ. We form the matrix $U \in \mathcal{M}(2 \times 2)$ and $\Sigma \in \mathcal{M}(2 \times 3)$ by

$$U = \frac{1}{\sqrt{2}}\begin{pmatrix} -1 & 1 \\ 1 & 1 \end{pmatrix}, \Sigma = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \sqrt{3} & 0 \end{pmatrix}.$$

Finally, we consider the matrix

$$\Psi^*\Psi = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 2 \end{pmatrix}.$$

The eigenvalues are 1, 3, and 0 with corresponding eigenvectors

$$\frac{1}{\sqrt{2}}\begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}, \frac{1}{\sqrt{3}}\begin{pmatrix} -1 \\ -1 \\ 1 \end{pmatrix}, \frac{1}{\sqrt{6}}\begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix}.$$

We form the matrix $V \in \mathcal{M}(3 \times 3)$ by

$$V = \begin{pmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{3}} \\ 0 & \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{3}} \end{pmatrix}.$$

These are exactly the matrices that are used in the singular value decomposition of Ψ, i.e. $\Psi = U\Sigma V^*$. To find the 1-tight-frame that minimizes the least-squares error, we replace the singular values of Ψ with $1s$ and get

$$M = U \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} V^* = \begin{pmatrix} \frac{1}{2} + \frac{\sqrt{3}}{6} & -\frac{1}{2} + \frac{\sqrt{3}}{6} & \frac{\sqrt{3}}{3} \\ -\frac{1}{2} + \frac{\sqrt{3}}{6} & \frac{1}{2} + \frac{\sqrt{3}}{6} & \frac{\sqrt{3}}{3} \end{pmatrix}$$

the columns for the 1-tight frame of $\mathbb{R}^2$. that minimizes the least squares error.

6.3.2 Example of ϵ-modified vectors

The vectors in the previous example were linearly dependent. We now expand the Hilbert space to $H' = \mathbb{R}^5$ and consider the linearly independent vectors

$$\psi_1 = \begin{pmatrix} 1 \\ 0 \\ \epsilon \\ 0 \\ 0 \end{pmatrix}, \psi_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ \epsilon \\ 0 \end{pmatrix}, \psi_3 = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ \epsilon \end{pmatrix}$$

for some $\epsilon > 0$ and we find the orthonormal set in $\{e_i(\epsilon)\}_{i=1}^3 \subset H' = \mathbb{R}^5$ that minimizes the least-squares error

$$E = \frac{1}{3} \sum_{i=1}^{3} \|\psi_i(\epsilon) - e_i(\epsilon)\|^2$$

over all other 3-element orthonormal sets.

Note that it may seem simpler to just go up one dimension, that is consider the Hilbert space $\mathbb{R}^3$ and the perturbed vectors

$$\psi_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \psi_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \psi_3 = \begin{pmatrix} 1 \\ 1 \\ \epsilon \end{pmatrix}.$$

However, if we were to do this, we lose some symmetry which makes things much harder to compute. For example, the corresponding matrix $\Psi\Psi^*$ has eigenvalues

$$1, \frac{\epsilon^2}{2} \pm \frac{3}{2} + \frac{\sqrt{9 + 2\epsilon^2 + \epsilon^4}}{2}$$

and eigenvectors of the form

$$\begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} -\frac{\frac{\epsilon^2}{2} - \frac{3}{2} \pm \frac{\sqrt{9+2\epsilon^2+\epsilon^4}}{2}}{2\epsilon} \\ -\frac{\frac{\epsilon^2}{2} - \frac{3}{2} \pm \frac{\sqrt{9+2\epsilon^2+\epsilon^4}}{2}}{2\epsilon} \\ 1 \end{pmatrix}.$$

Since we need the normalized eigenvectors to form the matrix $V(\epsilon)$, the expression for $V(\epsilon)$ is further complicated by the normalization factors. The resulting matrices $U(\epsilon)$ and $V(\epsilon)$ are very complicated, and the expression for the resulting 1-tight frame matrix is so complicated that it will not even fit onto the page!

We analogously go through the same procedure. We form the matrix

$$\Psi = \begin{pmatrix} | & | & | \\ \psi_1 & \psi_2 & \psi_3 \\ | & | & | \end{pmatrix} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ \epsilon & 0 & 0 \\ 0 & \epsilon & 0 \\ 0 & 0 & \epsilon \end{pmatrix}.$$

We now take the singular decomposition of Ψ. First we look at

$$\Psi\Psi^* = \begin{pmatrix} 2 & 1 & \epsilon & 0 & \epsilon \\ 1 & 2 & 0 & \epsilon & \epsilon \\ \epsilon & 0 & \epsilon^2 & 0 & 0 \\ 0 & \epsilon & 0 & \epsilon^2 & 0 \\ \epsilon & \epsilon & 0 & 0 & \epsilon^2 \end{pmatrix}.$$

The eigenvalues of $\Psi\Psi^*$ are ϵ^2, $1+\epsilon^2$, $3+\epsilon^2$ with corresponding eigenvectors

$$\frac{1}{\sqrt{3}}\begin{pmatrix} 0 \\ 0 \\ -1 \\ -1 \\ 1 \end{pmatrix}, \frac{1}{\sqrt{2+2\epsilon^2}}\begin{pmatrix} -1 \\ 1 \\ -\epsilon \\ \epsilon \\ 0 \end{pmatrix}, \frac{1}{\sqrt{18+6\epsilon^2}}\begin{pmatrix} 3 \\ 3 \\ \epsilon \\ \epsilon \\ 2\epsilon \end{pmatrix}, \frac{1}{\sqrt{2+2\epsilon^2}}\begin{pmatrix} \epsilon \\ -\epsilon \\ -1 \\ 1 \\ 0 \end{pmatrix}, \frac{1}{\sqrt{2+\epsilon^2}}\begin{pmatrix} -\epsilon \\ 0 \\ 1 \\ 0 \\ 1 \end{pmatrix}.$$

These eigenvalues are the squares of the singular values of Ψ. We form the matrix

$U(\epsilon) \in \mathcal{M}(5\times 5)$ and $\Sigma \in \mathcal{M}(5\times 3)$ by

$$U(\epsilon) = \begin{pmatrix} 0 & -\frac{1}{\sqrt{2+2\epsilon^2}} & \frac{3}{\sqrt{18+6\epsilon^2}} & \frac{\epsilon}{\sqrt{2+2\epsilon^2}} & -\frac{\epsilon}{\sqrt{2+\epsilon^2}} \\ 0 & \frac{1}{\sqrt{2+2\epsilon^2}} & \frac{3}{\sqrt{18+6\epsilon^2}} & -\frac{\epsilon}{\sqrt{2+2\epsilon^2}} & 0 \\ -\frac{1}{\sqrt{3}} & -\frac{\epsilon}{\sqrt{2+2\epsilon^2}} & \frac{\epsilon}{\sqrt{18+6\epsilon^2}} & -\frac{1}{\sqrt{2+2\epsilon^2}} & \frac{1}{\sqrt{2+\epsilon^2}} \\ -\frac{1}{\sqrt{3}} & \frac{\epsilon}{\sqrt{2+2\epsilon^2}} & \frac{\epsilon}{\sqrt{18+6\epsilon^2}} & \frac{1}{\sqrt{2+2\epsilon^2}} & 0 \\ \frac{1}{\sqrt{3}} & 0 & \frac{2\epsilon}{\sqrt{18+6\epsilon^2}} & 0 & \frac{1}{\sqrt{2+\epsilon^2}} \end{pmatrix}, \Sigma = \begin{pmatrix} \epsilon & 0 & 0 \\ 0 & \sqrt{1+\epsilon^2} & 0 \\ 0 & 0 & \sqrt{3+\epsilon^2} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Finally, we consider the matrix

$$\Psi^*\Psi = \begin{pmatrix} 1+\epsilon^2 & 0 & 1 \\ 0 & 1+\epsilon^2 & 1 \\ 1 & 1 & \epsilon^2+2 \end{pmatrix}.$$

The eigenvalues are the same as for $\Psi\Psi^*$ with corresponding eigenvectors

$$\frac{1}{\sqrt{3}}\begin{pmatrix}-1\\-1\\1\end{pmatrix}, \frac{1}{\sqrt{2}}\begin{pmatrix}-1\\-1\\0\end{pmatrix}, \frac{1}{\sqrt{6}}\begin{pmatrix}1\\1\\2\end{pmatrix}.$$

We form the matrix $V \in \mathcal{M}(3\times 3)$ by

$$V = \begin{pmatrix} -\frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \\ -\frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & 0 & \frac{2}{\sqrt{6}} \end{pmatrix}$$

which, as expected, is independent of ϵ.

So in our example, we have,

$$M(\epsilon) = U(\epsilon)\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} V^*$$

the columns form the 1-tight frame of $\mathbb{R}^5$ that minimizes the least squares error for the given vectors $\{\psi_i(\epsilon)\}_{i=1}^3$. We set $\epsilon = 0$

$$U(\epsilon=0) = \begin{pmatrix} 0 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 & 0 \\ 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 & 0 \\ -\frac{1}{\sqrt{3}} & 0 & 0 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{3}} & 0 & 0 & \frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{3}} & 0 & 0 & 0 & \frac{1}{\sqrt{2}} \end{pmatrix}.$$

So we have,

$$M(\epsilon=0) = U(\epsilon=0)\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} V^*$$

$$= \begin{pmatrix} \frac{1}{2}+\frac{\sqrt{3}}{6} & -\frac{1}{2}+\frac{\sqrt{3}}{6} & \frac{\sqrt{3}}{3} \\ -\frac{1}{2}+\frac{\sqrt{3}}{6} & \frac{1}{2}+\frac{\sqrt{3}}{6} & \frac{\sqrt{3}}{3} \\ \frac{1}{3} & \frac{1}{3} & -\frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & -\frac{1}{3} \\ -\frac{1}{3} & -\frac{1}{3} & \frac{1}{3} \end{pmatrix}.$$

The columns are the closest orthogonal set that minimizes the least-squares error. Note that if we project these vectors into the original space of $\mathbb{R}^2$, we get our previous solution, that is we get the 1-tight frame in $\mathbb{R}^2$ that minimizes the least-squares error.

6.4 Geometrically uniform frames

Let H be a d-dimensional Hilbert space. Let $\mathcal{Q} = \{U_i \in \mathcal{L}(H) : 1 \leq i \leq N\}$ be a finite abelian group of N unitary linear operators. A set of N vectors $\{\phi_i\}_{i=1}^N \subset H$ is said to be *geometrically uniform* if there exists a $\phi \in H$ such that

$$\{\phi_i\}_{i=1}^N = \{U_i\phi\}_{U_i \in \mathcal{Q}}.$$

ϕ is usually referred to as the *generating vector*. A frame is said to be a *geometrically uniform frame* (GU) if it is also a geometrically uniform set of vectors.

6.4.1 Examples of GU vector sets

Consider the Hilbert space $H = \mathbb{R}^2$. First note that any two distinct vectors of the same length, as shown in Figure 1, is GU. The abelian group of unitary operators consists of just the identity map, and a reflection along the line of symmetry between the two vectors.

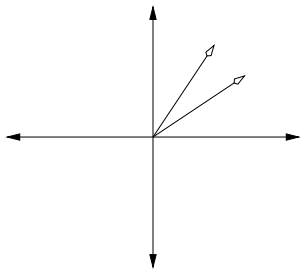

Figure 6.1: A GU set consisting of 2 vectors.

Now consider the generating vector $\phi = \begin{pmatrix} \frac{1}{2} \\ \\ \frac{\sqrt{3}}{2} \end{pmatrix}$ and the abelian group

$$\mathcal{Q} = \{I, \mathcal{R}_{\pi/3}, \mathcal{R}_{2\pi/3}\}$$

where $\mathcal{R}_{\pi/3}$ is the rotation by angle $\pi/3$ and $\mathcal{R}_{2\pi/3}$ is the rotation by angle $2\pi/3$. This GU set of vectors is depicted in Figure 2.

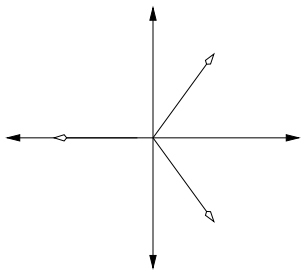

Figure 6.2: A GU set consisting of 3 vectors.

6.4.2 Properties of GU frames and a second solution

Geometrically uniform frames have several nice properties. If we have a frame, and if we still have a frame after removing a single vector, then clearly the frame bounds will change. However, it has been shown in [11] that if it is a GU frame, then the frame bounds change independently of which frame vector is removed.

Second, if we are given a GU frame, it has been shown in [12] that the corresponding tight frame that minimizes the least-squares error E also minimizes the probability of detection error P_e and inherent the geometrically uniform property. This is shown in the following theorems.

Theorem 6.5. *Let H be a finite dimensional Hilbert space. Let $\phi \in H$ and assume that $T = \{\phi_i\}_{i=1}^N = \{U_i\phi : U_i \in \mathcal{Q}\}$ is a GU frame for H. Additionally, assume that for any integer j, $T(j) = \{U_i\phi : U_i \in \mathcal{Q},\ i \neq j\}$ is still a frame for H. Then the frame bounds change independently of the choice of j, i.e. the frame bounds of $T(j)$ are the same for all j.*

Proof. Let $\{e_i\}_{i=1}^d$ be an orthonormal basis for H and let $\Phi \in \mathcal{M}(N \times d)$ be the Bessel map matrix corresponding to the vector set $\{U_i\phi\}_{i=1}^N$ with respect to the basis $\{e_i\}_{i=1}^d$. By observation 3 from section 2.2.1, the $d \times d$ frame operator matrix can be written as

$$S = \Phi^*\Phi = \sum_{i=1}^N \phi_i\phi_i^*.$$

Note that S is self-adjoint and positive, hence it has real nonzero eigenvalues. If T has frame bounds A and B, then we have for all $x \in H$,

$$A\|x\|^2 \leq \sum_{i=1}^N |\langle x, \phi_i\rangle|^2 \leq B\|x\|^2$$

and using the definition of the frame operator we have

$$\sum_{i=1}^N |\langle x, \phi_i\rangle|^2 = \|\Phi x\|^2_{l^2(\mathbb{Z}_N)} = \langle \Phi x, \Phi x\rangle = \langle x, \Phi^*\Phi x\rangle = \langle x, Sx\rangle$$

hence

$$A\|x\|^2 \leq \langle x, Sx\rangle \leq B\|x\|^2.$$

So we see that the frame bounds are

$$A = \min\{\lambda : \lambda \text{ is an eigenvalue of } S\}$$

and

$$B = \max\{\lambda : \lambda \text{ is an eigenvalue of } S\}.$$

So to prove the theorem, it suffices to show that the removal of a vector ϕ_j gives us a new frame operator whose set of eigenvalues doesn't dependent on the choice of ϕ_j. We can write this new frame operator with ϕ_j removed as

$$S(j) = \sum_{i=1}^N \phi_i\phi_i^* - \phi_j\phi_j^* = \sum_{i=1}^N U_i\phi\phi^*U_i^* - U_j\phi\phi^*U_j^*.$$

Note that since $\mathcal{Q}$ is a group we have

$$U_j^*\mathcal{Q} = \{U_j^*U_i\phi : 0 \leq i \leq N\} = \{U_i\phi : 0 \leq i \leq N\} = \mathcal{Q}.$$

So conjugating the original frame operator with U_j gives,

$$U_j^*SU_j = \sum_{i=1}^{N} U_j^*U_i\phi\phi^*U_i^*U_j = \sum_{i=1}^{N} U_j^*U_i\phi\phi^*(U_j^*U_i)^* = S.$$

Also, since similar matrices have the same eigenvalues, we consider $U_j^*S(j)U_j$ and get

$$U_j^*S(j)U_j = \sum_{i=1}^{N} U_j^*U_i\phi\phi^*U_i^*U_j - \phi\phi^* = S - \phi\phi^*.$$

The eigenvalues of $S-\phi\phi^*$ do not depend on j, hence the eigenvalues of $S(j)$ do not depend on j. It follows that the frame bounds for $T(j)$ are the same for all choices of j. □

Theorem 6.6. *Let*

$$T = \{U_i\phi : U_i \in \mathcal{Q}\}$$

be a GU tight frame in a d-dimensional Hilbert space H where $\phi \in H$ and $\|\phi\| = 1$. Suppose T is still a frame with $\phi_j = U_j\phi$ removed, i.e. suppose

$$T(j) = \{U_i\phi : U_i \in \mathcal{Q}, i \neq j\}$$

is a frame. Then $T(j)$ has frame bounds $A = \frac{N}{d} - 1$ and $B = \frac{N}{d}$.

We first start with a lemma.

Lemma 6.5. *Assume we have a GU frame in a d-dimensional Hilbert space with generating vector ϕ. Then the frame bounds satisfy $A \leq \frac{N}{d}\|\phi\|^2 \leq B$.*

Proof of Lemma 6.5. Let $\{e_i\}_{i=1}^d$ be an orthonormal basis for H and $S \in \mathcal{M}(d \times d)$ be the frame operator matrix with respect to the basis and let $\{\lambda_i\}_{i=1}^d$ represent the eigenvalues of S. We can write the frame operator matrix for a GU frame as

$$S = \sum_{i=1}^{N} U_i \phi \phi^* U_i^*.$$

We have,

$$\sum_{i=1}^{d} \lambda_i = Tr(S) = \sum_{i=1}^{N} Tr(U_i \phi \phi^* U_i^*) = \sum_{i=1}^{N} Tr(\phi \phi^*) = N\|\phi\|^2,$$

and

$$dA = d \min_i \{\lambda_i\} \leq \sum_{i=1}^{d} \lambda_i \leq d \max_i \{\lambda_i\} = dB.$$

Dividing by d gives the result we want. □

Proof of Theorem 6.6. By the lemma, we see that T has frame constant $A = \frac{N}{d}$. By observation 3 from section 2.2.1 we can write the frame operator matrix S with respect to some orthonormal basis of H as,

$$S = \sum_{i=1}^{N} \phi_i \phi_i^* = \frac{N}{d} I_d$$

where I_d is the $d \times d$ identity matrix. Let $T(j)$ be the frame where ϕ_j is removed. Then the corresponding frame operator is

$$S(j) = \sum_{i=1}^{N} \phi_i \phi_i^* - \phi_j \phi_j^* = \frac{N}{d} I_d - U_j \phi \phi^* U_j^*.$$

Again, we consider the similar matrix $U_j^* S(j) U_j$ and get,

$$U_j^* S(j) U_j = \frac{N}{d} I_d - \phi \phi^*.$$

Note that $\phi\phi^*$ is a matrix with only one nonzero eigenvalue of 1, since $\|\phi\| = 1$, with eigenvector ϕ. So by the above expression, it follows that $S(j)$ has distinct eigenvalues of $\frac{N}{d}$ and $\frac{N}{d} - 1$, hence the frames bounds are $A = \frac{N}{d} - 1$ and $B = \frac{N}{d}$. $\square$

It can be shown that if the given vectors $T = \{\phi_i\}_{i=1}^N \subset H$ is a GU frame, then the unique tight frame $\{e_i\}_{i=1}^N$ with frame constant A^2 that minimizes the least-squares error E also minimizes the probability of detection error P_e and is a GU tight frame. We must introduce several new definitions before proceeding with the proof.

6.4.3 Preliminaries

Definition 6.1. Given a GU frame $T = \{\phi_i\}_{i=1}^N$, we define the $N \times N$ *Gram matrix* G as the matrix with entries $G_{ij} = \langle \phi_i, \phi_j \rangle = \langle U_i\phi, U_j\phi \rangle = \langle \phi, U_i^* U_j \phi \rangle$.

Since $\mathcal{Q}$ is a group and $U_i^* U_j \in \mathcal{Q}$, we see that for fixed i the set $\{U_i^* U_j \phi : 1 \leq j \leq n\}$ is just a permutation of the set $\{U_j\phi : 1 \leq j \leq n\}$. As a side note, we see that all the columns of G have the same entries but are just permuted in a different order. A matrix of this type is called a *permuted matrix*. All GU vectors have a permuted Gram matrix. This in fact characterizes GU vectors. According to [4], if a set of vectors $T = \{\phi_i\}_{i=1}^N$ has a permuted Gram matrix, and has the property for all i and j that $\langle \phi_i, \phi_j \rangle = \langle \phi_j, \phi_i \rangle$, then T is a GU set.

6.4.4 Change of notation

Let $\mathcal{Q}$ be a finite abelian group of N elements. Then $\mathcal{Q}$ is isomorphic to a direct product of cyclic groups, i.e.

$$\mathcal{Q} \cong Q = \mathbb{Z}_{n_1} \otimes \ldots \otimes \mathbb{Z}_{n_p}$$

where the group operation on Q is componentwise modular addition and $N = \prod_{i=1}^{p} n_i$. Let $\phi \in H$ and $T = \{U_i\phi : U_i \in \mathcal{Q}\}$ a GU set of vectors. Since $\mathcal{Q} \cong Q$, for any $U_i \in \mathcal{Q}$ there corresponds a $q \in Q$ so that we can denote

$$\phi_i = U_i\phi = \phi(q).$$

In this notation, we can change the indices of the Gram matrix G to elements of Q, i.e. for $g, h \in Q$, the (g, h)th entry is $G_{g,h} = \langle \phi(g), \phi(h) \rangle$.

Definition 6.2. Define the *Gram function* $s : Q \to \mathbb{C}$ for all $g \in Q$ as,

$$s(g) = \langle \phi, \phi(g) \rangle.$$

We now illustrate the connection between the Gram function and the Gram matrix. Let $g, h \in Q$ and let U_i and U_j be the corresponding elements in $\mathcal{Q}$ respectively. Then note that

$$\begin{aligned} s(g-h) &= \langle \phi, \phi(g-h) \rangle = \langle \phi, U_i U_j^* \phi \rangle = \langle \phi, U_j^* U_i \phi \rangle \\ &= \langle U_j\phi, U_i\phi \rangle = \langle \phi(h), \phi(g) \rangle = G_{h,g}. \end{aligned}$$

6.4.5 Fourier transform of functions on Q

Definition 6.3. Given a function $f : Q \to \mathbb{C}$ we define the *Fourier Transform* of f for all $h \in Q$ as

$$\hat{f}(h) = \frac{1}{\sqrt{N}} \sum_{q \in Q} \langle h, q \rangle f(q)$$

where

$$\langle h, q \rangle = \prod_{i=1}^{p} e^{-2\pi i h_i q_i / n_i}$$

where $h_i,\ q_i \in \mathbb{Z}_{n_i}$ are the ith components of h and q respectively.

With this definition, we have for all $g, h, h' \in Q$,

$$\langle h, g \rangle = \langle g, h \rangle$$

$$\langle h, g \rangle^* = \langle -h, g \rangle = \langle h, -g \rangle$$

$$\langle h + h', g \rangle = \langle h, g \rangle \langle h', g \rangle$$

where * denotes complex conjugation. It is natural to define the $N \times N$ *Fourier transform matrix* $\mathcal{F}$ as the matrix with entries $\mathcal{F}_{g,h} = \frac{1}{\sqrt{N}} \langle h, g \rangle$ for indices $h, g \in Q$. With this definition, it is not hard to show that $\mathcal{F}$ is a unitary matrix. Note that a function $f : Q \to \mathbb{C}$ can be considered as a vector $\vec{f} = \{f(g)\}_{g \in Q}$ with Fourier transform $\vec{\hat{f}} = \mathcal{F}\vec{f}$.

Lemma 6.6. *Let $T = \{\phi_i\}_{i=1}^{N}$ be a set of GU vectors in an d-dimensional Hilbert space H. Then the corresponding Gram matrix G is diagonalizable by the FT matrix $\mathcal{F}$.*

Proof. It suffices to show that the columns of $\mathcal{F}$ are eigenvectors of the Gram matrix G. Let $k \in Q$ be fixed and let $\mathcal{F}_k$ be the kth column of $\mathcal{F}$. Then the h^{th} component

of the vector $G\mathcal{F}_k$ is,

$$\begin{aligned}
[G\mathcal{F}_k]_h &= \sum_{g\in Q} G_{h,g}\mathcal{F}_{g,k} = \frac{1}{\sqrt{N}}\sum_{g\in Q}\langle k,g\rangle s(g-h) \\
&= \frac{1}{\sqrt{N}}\sum_{g\in Q}\langle k,g+h\rangle s(g) = \frac{1}{\sqrt{N}}\sum_{g\in Q}\langle k,h\rangle\langle k,g\rangle s(g) \\
&= \langle k,h\rangle\frac{1}{\sqrt{N}}\sum_{g\in Q}\langle k,g\rangle s(g) = \langle k,h\rangle\hat{s}(k) \\
&= \sqrt{N}\hat{s}(k)\frac{1}{\sqrt{N}}\langle k,h\rangle = \sqrt{N}\hat{s}(k)\mathcal{F}_{h,k}.
\end{aligned}$$

So, we see that

$$G\mathcal{F}_k = \sqrt{N}\hat{s}(k)\mathcal{F}_k.$$

Using the fact that $\mathcal{F}$ is unitary it is not hard to show that $\mathcal{F}^*G\mathcal{F}$ is a diagonal matrix with diagonal components $\sqrt{N}\hat{s}(k)$ for $k \in Q$.

Note that the Gram matrix can be written as

$$G = \Phi\Phi^*$$

where Φ is the Bessel map matrix for the set T. So G is nonnegative and self-adjoint, hence $\hat{s}(k)$ is both real and nonnegative for all $k \in Q$. □

We have the following lemma from [31].

Lemma 6.7. *Let H be a d-dimensional Hilbert space and $\{\phi_i\}_{i=1}^N \subset H$ be a frame for H with corresponding weights $\{\rho_i\}_{i=1}^N$. Define the operators $\{W_i\}_{i=1}^N \subset \mathcal{L}(H)$ for all $x \in H$ by*

$$W_i x = \rho_i\langle\phi_i, x\rangle\phi_i.$$

Let $\{e_i\}_{i=1}^N$ be a tight frame corresponding to a POM Π defined for all $x \in H$ and

$1 \leq i \leq N$ *by*

$$\Pi(i)x = \langle e_i, x\rangle e_i.$$

Then $\{e_i\}_{i=1}^N$ *minimizes the probability of detection error*

$$P_e = 1 - \sum_{i=1}^{N} \rho_i |\langle \phi_i, e_i\rangle|^2$$

if

1. $\Pi(i)(W_j - W_i)\Pi(j) = 0 \ \forall \ i, \ j = 1, \ldots, N$
2. $\sum_{i=1}^{N} \Pi(i)W_i - W_j \geq 0 \ \forall \ j = 1, \ldots, N.$

6.5 Minimizers of P_e

We are now in a position to prove a second solution to the quantum detection problem.

Theorem 6.7. *Let* H *be an* d*-dimensional Hilbert space and assume that* $T = \{\phi_i\}_{i=1}^N \subset H$ *is a GU frame for* H*, and let* $A > 0$ *be given. Then the unique tight frame* $\{e_i\}_{i=1}^N$ *with frame constant* A^2 *that minimizes the least-squares error* E *also minimizes the probability of detection error*

$$P_e = 1 - \sum_{i=1}^{N} \frac{1}{N} |\langle \psi_i, e_i\rangle|^2$$

and is a GU tight frame.

Proof. We will first show that the tight frame $\{e_i\}_{i=1}^N$ that minimizes the least-squares error E is also GU. Let $T = \{\phi_i\}_{i=1}^N$ be a GU frame for H and let $A > 0$ be given. Let $\Phi \in \mathcal{M}(N \times d)$ be the Bessel map matrix corresponding to the vector

set T with respect to some orthonormal basis $\{e_i\}_{i=1}^d$ for H. Let $\Phi^* = U\Sigma V^*$ be the singular value decomposition of Φ^*. The corresponding Gram matrix can be written as

$$G = \Phi\Phi^* = V\Sigma^* U^* U\Sigma V^* = V\Sigma^*\Sigma V^*.$$

From Lemma 6.6, we see that we must have $V = \mathcal{F}$ and that the singular values are $N^{1/4}\sqrt{\hat{s}(k)}$ for $k \in Q$, i.e. we have

$$\Phi^* = U\Sigma\mathcal{F}^* = N^{1/4}\sum_{g\in Q}\sqrt{\hat{s}(g)}u_g(\mathcal{F}_g)^* \tag{1}$$

where u_g is the gth column of U and $\mathcal{F}_g$ is the gth column of $\mathcal{F}$.

By Theorem 6.1 and observation 1 from section 2.2.1, we know that the conjugate Bessel map matrix whose columns are the A^2 tight frame that minimizes the least-squares error E has the form

$$\begin{aligned}
F^* &= \sum_{i=1}^{Rank(\Phi)} u_i v_i^* = \sum_{h\in Q,\hat{s}(h)\neq 0} u_h(\mathcal{F}_h)^* \\
&= \sum_{h\in Q,\hat{s}(h)\neq 0} \begin{pmatrix} u_h(1) \\ \vdots \\ u_h(d) \end{pmatrix} \frac{1}{\sqrt{N}}(\langle h, g_1\rangle^*, \ldots, \langle h, g_N\rangle^*) \\
&= \sum_{h\in Q,\hat{s}(h)\neq 0} \frac{1}{\sqrt{N}} \begin{pmatrix} u_h(1)\langle h, g_1\rangle^* & \ldots & u_h(1)\langle h, g_N\rangle^* \\ \vdots & \ddots & \vdots \\ u_h(d)\langle h, g_1\rangle^*, & \ldots & u_h(d)\langle h, g_N\rangle^* \end{pmatrix}
\end{aligned}$$

where $u_h(i)$ corresponds to the ith component of the vector u_h and g_i is the ith element of Q. Hence the gth column can be written as

$$e(g) = F_g^* = \frac{1}{\sqrt{N}}\sum_{h\in Q,\hat{s}(h)\neq 0}\langle h, g\rangle^* u_h. \tag{2}$$

We want to show that for $U_i \in \mathcal{Q}$, corresponding to $g' \in Q$, that

$$U_i e(g) = U_i F_g^* = F_{g+g'}^* = e(g+g')$$

where addition of the indices is modular addition in each component of $Q = Z_{n_1} \otimes \ldots \otimes Z_{n_p}$. This would show that the A^2-tight frame $\{e(g)\}_{g\in Q}$ is GU. In order to analyze $U_i e(g)$, inspection of equation (2) tells us that we first must look at $U_i u_h$, hence we now want to find an expression for u_h. Since $\mathcal{F}$ is unitary, the column vectors are orthonormal, hence multiplying by the column vector $\mathcal{F}_h$ on both sides of (1) gives us,

$$\Phi^* \mathcal{F}_h = N^{1/4} \sum_{g\in Q} \sqrt{\hat{s}(g)} u_g \mathcal{F}_g^* \mathcal{F}_h = N^{1/4} \sqrt{\hat{s}(h)} u_h. \tag{3}$$

We write,

$$\Phi^* \mathcal{F}_h = \frac{1}{\sqrt{N}} \begin{pmatrix} | & | & \\ \phi(g_1) & \phi(g_2) & \cdots \\ | & | & \end{pmatrix} \begin{pmatrix} \langle h, g_1 \rangle \\ \langle h, g_2 \rangle \\ \vdots \end{pmatrix} = \hat{\phi}(h)$$

where we define the Fourier transform of $\phi : Q \to \mathbb{C}^N$ by

$$\hat{\phi}(h) = \frac{1}{\sqrt{N}} \sum_{g\in Q} \langle h, g \rangle \phi(g).$$

So solving for u_h in equation (3) gives us,

$$u_h = \frac{1}{N^{1/4}\sqrt{\hat{s}(h)}} \hat{\phi}(h) \tag{4}$$

for values of h such that $\hat{s}(h) \neq 0$. So we see that in order to find an expression for $U_i u_h$, we must determine an expression for $U_i \hat{\phi}(h)$. First note that for any $g \in Q$

and $\phi(g) \in T$, where $U_k \in \mathcal{Q}$ corresponds to the element $g \in Q$, we have $U_i\phi(g) = U_iU_k\phi = \phi(g+g')$. So we have

$$\begin{aligned} U_i\hat{\phi}(h) &= \frac{1}{\sqrt{n}}\sum_{k\in Q}\langle h,k\rangle U_i\phi(k) = \frac{1}{\sqrt{n}}\sum_{k\in Q}\langle h,k\rangle\phi(k+g') \\ &= \frac{1}{\sqrt{n}}\sum_{k\in Q}\langle h,k-g'\rangle\phi(k) = \langle h,-g'\rangle\frac{1}{\sqrt{n}}\sum_{k\in Q}\langle h,k\rangle\phi(k) \\ &= \langle h,-g'\rangle\hat{\phi}(h) \end{aligned}$$

hence applying U_i to equation (4) gives

$$U_iu_h = \frac{1}{N^{1/4}\sqrt{\hat{s}(h)}}U_i\hat{\phi}(h) = \langle h,-g'\rangle\frac{1}{n^{1/4}\sqrt{\hat{s}(h)}}\hat{\phi}(h) = \langle h,-g'\rangle u_h.$$

So computing U_ie_g by using the above expression and equation (2) gives us

$$\begin{aligned} U_ie(g) &= U_iF_g^* = \frac{1}{\sqrt{N}}\sum_{h\in Q,\hat{s}(h)\neq 0}\langle h,g\rangle^*U_iu_h = \frac{1}{\sqrt{N}}\sum_{h\in Q,\hat{s}(h)\neq 0}\langle h,g\rangle^*\langle h,-g'\rangle u_h \\ &= \frac{1}{\sqrt{N}}\sum_{h\in Q,\hat{s}(h)\neq 0}\langle h,-g\rangle\langle h,-g'\rangle u_h = \frac{1}{\sqrt{N}}\sum_{h\in Q,\hat{s}(h)\neq 0}\langle h,-g-g'\rangle u_h \\ &= \frac{1}{\sqrt{N}}\sum_{h\in Q,\hat{s}(h)\neq 0}\langle h,g+g'\rangle^*u_h = F_{g+g'}^* = e(g+g') \end{aligned}$$

which is what we wanted to show. Hence $\{e(g)\}_{g\in Q}$ is GU.

Now we want to show that $\{e(g)\}_{g\in Q}$ minimizes the probability of a detection error. We will show that the $\{e(g)\}_{g\in Q}$ satisfies the conditions of Lemma 6.7, hence minimizes P_e.

Note that we can write the singular value decomposition of the conjugate Bessel map matrix for $\{e(g)\}_{g\in Q}$ as

$$F^* = \sum_{g\in Q}u_g\mathcal{F}_g^* = U\Upsilon\mathcal{F}^* \tag{5}$$

where $\Upsilon \in \mathcal{M}(d \times N)$ has singular values of 1 along the diagonal. So by equation (1) and the expression of F^* given above we have,

$$\begin{aligned} F\Phi^* &= N^{1/4} \sum_{k \in Q} \mathcal{F}_k u_k^* \sum_{g \in Q} \sqrt{\hat{s}(g)} u_g \mathcal{F}_g^* = N^{1/4} \sum_{k,g \in Q} \sqrt{\hat{s}(g)} \mathcal{F}_k u_k^* u_g \mathcal{F}_g^* \\ &= N^{1/4} \sum_{k,g \in Q} \sqrt{\hat{s}(g)} \mathcal{F}_k \langle u_k, u_g \rangle \mathcal{F}_g^* = N^{1/4} \sum_{k, \in Q} \sqrt{\hat{s}(k)} \mathcal{F}_k \mathcal{F}_k^* \\ &= \mathcal{F} \Upsilon^* \Sigma \mathcal{F}^*. \end{aligned}$$

So we see that $F\Phi^*$ is self-adjoint. Note that the components of the matrix $F\Phi^*$ are given, for any $g, h \in Q$, by $[F\Phi^*]_{g,h} = \langle e(g), \phi(h) \rangle$. Since $F\Phi^*$ is self adjoint we have for all $g,\ h \in Q$,

$$\langle \phi(h), e(g) \rangle = \langle e(g), \phi(h) \rangle^* = [(F\Phi^*)^*]_{g,h} = [F\Phi^*]_{h,g} = \langle e(h), \phi(g) \rangle. \tag{6}$$

Given any $g \in Q$ let U_i be the corresponding element of $\mathcal{Q}$. Since $F\Phi^* = \mathcal{F}\Upsilon^*\Sigma\mathcal{F}^*$ the diagonal elements of $F\Phi^*$ for any $g \in Q$ are

$$[F\Phi^*]_{g,g} = \langle e(g), \phi(g) \rangle = \langle (\mathcal{F}^*)_g, \Upsilon^* \Sigma (\mathcal{F}^*)_g \rangle \tag{7}$$

where $(\mathcal{F}^*)_g$ is the gth column of $\mathcal{F}^*$. Also,

$$\langle e(g), \phi(g) \rangle = \langle U_i e, U_i \phi \rangle = \langle e, U_i^* U_i \phi \rangle = \langle e, \phi \rangle. \tag{8}$$

So we see that all of the diagonal elements of $F\Phi^*$ are constant. Note also that by equation (6),

$$\langle e, \phi \rangle = \langle e(g), \phi(g) \rangle = \langle \phi(g), e(g) \rangle = \langle \phi, e \rangle = \langle e, \phi \rangle^*,$$

hence $\langle e, \phi \rangle \in \mathbb{R}$.

We now define the operators $\{W_g\}_{g\in Q}$ and $\{\Pi_g\}_{g\in Q}$ for all $x \in H$ as

$$W(g)(x) = \langle \phi(g), x\rangle \phi(g)$$

and

$$\Pi(g)(x) = \langle e(g), x\rangle e(g).$$

Note that we can write $W(g) = \phi(g)\phi(g)^*$ and $\Pi(g) = e(g)e(g)^*$. We now examine the first condition of Lemma 4. We have for all $g,\ h \in Q$,

$$\begin{aligned}
\Pi(g)(W(h) - W(g))\Pi(h) &= \Pi(g)W(h)\Pi(h) - \Pi(g)W(g)\Pi(h) \\
&= e(g)e(g)^*\phi(h)\phi(h)^*e(h)e(h)^* - e(g)e(g)^*\phi(g)\phi(g)^*e(h)e(h)^* \\
&= e(g)\langle e(g), \phi(h)\rangle\langle \phi(h), e(h)\rangle e(h)^* - e(g)\langle e(g), \phi(g)\rangle\langle \phi(g), e(h)\rangle e(h)^* \\
&= e(g)[\langle e(g), \phi(h)\rangle\langle \phi(h), e(h)\rangle - \langle e(g), \phi(g)\rangle\langle \phi(g), e(h)\rangle]e(h)^* \\
\text{by (6) and (8)} \quad &= e(g)[\langle \phi(g), e(h)\rangle\langle \phi, e\rangle - \langle e, \phi\rangle\langle \phi(g), e(h)\rangle]e(h)^* = 0.
\end{aligned}$$

So condition 1 of Lemma 4 is satisfied.

Now we need to show condition 2, i.e. we want to show that for all $h \in Q$,

$$\sum_{g\in Q} \Pi(g)W(g) - \phi(h)\phi(h)^* \geq 0.$$

By equations (1), (5) and (8) we have,

$$\begin{aligned}
\sum_{g\in Q} \Pi(g)W(g) &= \sum_{g\in Q} e(g)e(g)^*\phi(g)\phi(g)^* = \sum_{g\in Q} e(g)\langle e(g), \phi(g)\rangle\phi(g)^* \\
&= \langle e, \phi\rangle \sum_{g\in Q} e(g)\phi(g)^* = \langle e, \phi\rangle F^*\Phi \\
&= \langle e, \phi\rangle (U\Upsilon\mathcal{F}^*)(\mathcal{F}\Sigma^t U^*) = \langle e, \phi\rangle U\Upsilon\Sigma^t U^* \\
&= \langle e, \phi\rangle U\overline{\Sigma}U^*
\end{aligned}$$

where $\overline{\Sigma} = \Upsilon\Sigma^t \in \mathcal{M}(d \times d)$ is the diagonal matrix with singular values the same as Σ. Since $\Phi^* = U\Sigma\mathcal{F}^*$, we have for any $g \in Q$,

$$\phi(g) = U\Sigma(\mathcal{F}^*)_g$$

where $(\mathcal{F}^*)_g$ is the gth column of $\mathcal{F}^*$. Hence

$$\phi(g)\phi(g)^* = U\Sigma(\mathcal{F}^*)_g(\mathcal{F}^*)_g^*\Sigma^t U^*.$$

Hence to satisfy condition 2 of Lemma 4, for a fixed $h \in Q$ we want to show that the $d \times d$ matrix

$$\begin{aligned} T &= \prod_{g \in Q} \Pi(g)W_g - \phi(h)\phi(h)^* = \langle e, \phi\rangle U\overline{\Sigma}U^* - U\Sigma(\mathcal{F}^*)_g(\mathcal{F}^*)_g^*\Sigma^t U^* \\ &= U[\langle e, \phi\rangle\overline{\Sigma} - \Sigma(\mathcal{F}^*)_g(\mathcal{F}^*)_g^*\Sigma^t]U^* \end{aligned}$$

is positive. We will first show that $T' = \langle e, \phi\rangle\overline{\Sigma} - \Sigma(\mathcal{F}^*)_g(\mathcal{F}^*)_g^*\Sigma^t$ is positive. Let $x \in \mathbb{C}^d$. Define $\Sigma^{1/2}$ to be the $d \times N$ matrix with diagonal singular values of $(\sqrt{N}\hat{s}(k))^{1/2}$ for $k \in Q$. Then,

$$\begin{aligned} \langle x, T'x\rangle &= \langle e, \phi\rangle\langle x, \overline{\Sigma}x\rangle - \langle x, \Sigma(\mathcal{F}^*)_g(\mathcal{F}^*)_g^*\Sigma^t x\rangle \\ &= \langle e, \phi\rangle\langle x, \overline{\Sigma}x\rangle - \langle \Sigma^t x, (\mathcal{F}^*)_g\langle(\mathcal{F}^*)_g, \Sigma^t x\rangle\rangle \\ &= \langle e, \phi\rangle\langle x, \overline{\Sigma}x\rangle - \langle \Sigma^t x, (\mathcal{F}^*)_g\rangle\langle(\mathcal{F}^*)_g, \Sigma^t x\rangle \\ &= \langle e, \phi\rangle\langle x, \overline{\Sigma}x\rangle - |\langle \Sigma^t x, (\mathcal{F}^*)_g\rangle|^2 \\ &= \langle e, \phi\rangle\langle x, \overline{\Sigma}x\rangle - |\langle (\Sigma^{1/2})^t x, \Upsilon^*\Sigma^{1/2}(\mathcal{F}^*)_g\rangle|^2 \\ \text{by Cauchy Schwartz} \quad &\geq \langle e, \phi\rangle\langle x, \overline{\Sigma}x\rangle - \|(\Sigma^{1/2})^t x\|^2\|\Upsilon^*\Sigma^{1/2}(\mathcal{F}^*)_g\|^2 \\ &= \langle e, \phi\rangle\langle x, \overline{\Sigma}x\rangle - \langle x, \overline{\Sigma}x\rangle\langle(\mathcal{F}^*)_g, \Upsilon^*\Sigma(\mathcal{F}^*)_g\rangle \\ \text{by (7) and (8)} \quad &= \langle e, \phi\rangle\langle x, \overline{\Sigma}x\rangle - \langle x, \overline{\Sigma}x\rangle\langle e, \phi\rangle = 0 \end{aligned}$$

hence T' is a positive matrix. Now, given any $x \in \mathbb{C}^d$ we have

$$\langle x, Tx \rangle = \langle x, UT'U^*x \rangle = \langle U^*x, T'U^*x \rangle \geq 0$$

hence T is a positive operator. So we see that condition 2 of Lemma 4 is satisfied. It follows that $\{e(g)\}_{g \in Q}$ minimizes the probability of detection error P_e. $\square$

Appendix A

Penrose-Moore pseudo inverse

Let $A \in \mathcal{M}(n \times m)$. The *Penrose-Moore pseudo inverse* is the unique matrix $A^\dagger \in \mathcal{M}(m \times n)$ such that,

1. $AA^\dagger A = A$
2. $A^\dagger AA^\dagger = A^\dagger$
3. $(AA^\dagger)^* = AA^\dagger$
4. $(A^\dagger A)^* = A^\dagger A$.

If $A = U\Sigma V^*$ is the singular decomposition of A, where $U \in \mathcal{M}(n \times n)$, $\Sigma \in \mathcal{M}(n \times m)$, $V \in \mathcal{M}(m \times m)$, then the solution to the above is $A^\dagger = V\Sigma^\dagger U^*$ where $\Sigma^\dagger \in \mathcal{M}(m \times n)$ is of the form

$$\Sigma^\dagger = \begin{pmatrix} 1/\sigma_1 & & \\ & 1/\sigma_2 & \\ & & \ddots \end{pmatrix}$$

where $\{\sigma_i\}$ are the singular values of A.

Also, if $A \in \mathcal{M}(n \times m)$ with singular value decomposition $A = U\Sigma V^*$ we define $A^{1/2}$ by $A^{1/2} = U\Sigma^{1/2}V^*$ where $\Sigma^{1/2}$ has diagonals $\{\sqrt{\sigma_i}\}$, i.e. the singular values of $A^{1/2}$ are $\{\sqrt{\sigma_i}\}$.

Appendix B

Hilbert space definitions

Let $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$.

Definition B.1. Let H be a vector space over $\mathbb{K}$. A real-valued function $\|\cdot\|$ on H is a *norm* if:

1. For all $x \in H$, $\|x\| \geq 0$.
2. $\|x\| = 0$ if and only if $x = 0$.
3. For all $x, y \in H$, $\|x + y\| \leq \|x\| + \|y\|$.
4. For all $x \in H$ and $a \in \mathbb{K}$, $\|ax\| = |a|\|x\|$.

Definition B.2. Let H be a vector space with norm $\|\cdot\|$. H is *complete* if every Cauchy sequence in H converges, that is if a sequence $\{f_i\}_{i\in\mathbb{Z}} \subset H$ has the property that for any $\epsilon > 0$, there exists an $N > 0$ such that for all $n, m \geq N$,

$$\|f_n - f_m\| \leq \epsilon$$

then there exists an $f \in H$ such that

$$\lim_{n\to\infty} \|f_n - f\| = 0.$$

Definition B.3. A *Hilbert space* H is a Banach space over $\mathbb{K}$ with a $\mathbb{K}$-valued function $\langle\cdot,\cdot\rangle$ defined on $H \times H$, called an *inner product*, that has the properties:

1. For all $\alpha, \beta \in \mathbb{K}$ and $x, y, z \in H$,

$$\langle \alpha x + \beta y, z \rangle = \alpha \langle x, z \rangle + \beta \langle y, z \rangle.$$

2. For all $x, y \in H$,

$$\langle x, y \rangle = \langle y, x \rangle^*$$

where $*$ indicates complex conjugation.

3. For all $x \in H$, $\langle x, x \rangle = \|x\|^2$.

Definition B.4. Let H be a Hilbert space. H is *separable* if there exists a countable dense set X in H.

Separable Hilbert spaces always have an orthonormal basis, as the following theorem shows.

Theorem B.1. *Let H be a separable Hilbert space. Then there exists a complete orthonormal set in H.*

Definition B.5. Let H be a Hilbert space and $T : H \to H$ a linear operator. T is *bounded* if there exists a constant $A \in \mathbb{R}$ such that for all $x \in H$,

$$\|T(x)\| \leq A\|x\|.$$

For bounded linear operators T, the adjoint T^* is defined for all $x, y \in H$ by

$$\langle Tx, y \rangle = \langle x, T^* y \rangle.$$

However, if T is not bounded and defined on only a dense subset of H, more care must be used to define the adjoint. The following gives a more general definition of the adjoint, which reduces to the above definition when the operator T is bounded.

Definition B.6. Let H be a Hilbert space and T a linear operator defined on a dense subset $\mathrm{Dom}(T)$ of H. Define $\mathrm{Dom}(T^*)$ as the set of all $y \in H$ such that the operator T_y defined for all $x \in H$ by

$$T_y(x) = \langle Tx, y \rangle$$

is bounded. Then by the Hahn-Banach theorem, we can extend T_y to all of H, and hence there exists a unique element $T^*y \in H$ such that

$$T_y(x) = \langle Tx, y \rangle = \langle x, T^*y \rangle.$$

T^* is defined to be the *adjoint* of T, with domain $\mathrm{Dom}(T^*)$.

Definition B.7. Let H be a Hilbert space and T a linear operator defined on a dense subset of H. T is self-adjoint if $T = T^*$.

B.1 Examples of Hilbert spaces

Let $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$.

1. Let $K \subset \mathbb{Z}$. Define $l^2(K)$ by

$$l^2(K) = \left\{ \{a_i\}_{i \in K} \subset \mathbb{K} : \sum_{i \in K} |a_i|^2 < \infty \right\}.$$

For all $\{a_i\}_{i \in K}, \{b_i\}_{i \in K} \in l^2(K)$, we have the inner product defined by

$$\langle \{a_i\}_{i \in K}, \{b_i\}_{i \in K} \rangle = \sum_{i \in K} a_i b_i^*$$

where $*$ represents complex conjugation. With this inner product, $l^2(K)$ is a Hilbert space. In section 2.2 dealing with finite frames, $K = \mathbb{Z}_N$ where

$$\mathbb{Z}_N = \{0, 1, 2, \ldots, N-1\}.$$

2. Let $\Omega \subset \mathbb{R}$ and consider the space

$$L^2(\Omega) = \left\{ f \text{ measurable functions on } \Omega : \int_\Omega |f|^2 < \infty \right\}.$$

$L^2(\Omega)$ is a Hilbert space with inner product defined for all $f, g \in L^2(\Omega)$ by

$$\langle f, g \rangle = \int_\Omega f(x) g^*(x) dx.$$

BIBLIOGRAPHY

[1] John J. Benedetto and Matthew Fickus, *Finite normalized tight frames*, Adv. Comput. Math. **18** (2003), no. 2-4, 357–385, Frames. MR MR1968126 (2004c:42059)

[2] Charles H. Bennett, *Quantum cryptography using any two nonorthogonal states*, Phys. Rev. Lett. **68** (1992), no. 21, 3121–3124. MR 1 163 546

[3] Sterling K. Berberian, *Notes on spectral theory*, Van Nostrand Mathematical Studies, No. 5, D. Van Nostrand Co., Inc., Princeton, N.J.-Toronto, Ont.-London, 1966. MR 32 #8170

[4] Howard E. Brandt, *Positive operator valued measure in quantum information processing*, Amer. J. Phys. **67** (1999), no. 5, 434–439. MR 2000a:81023

[5] ———, *Quantum measurement with a positive operator-valued measure*, J. Opt. B Quantum Semiclass. Opt. **5** (2003), no. 3, S266–S270, Wigner centennial (Pécs, 2002). MR 2004g:81012

[6] Ingrid Daubechies, *Ten lectures on wavelets*, CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 61, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1992. MR MR1162107 (93e:42045)

[7] Chandler Davis, *Geometric approach to a dilation theorem*, Linear Algebra and Appl. **18** (1977), no. 1, 33–43. MR 58 #7179

[8] R. J. Duffin and A. C. Schaeffer, *A class of nonharmonic Fourier series*, Trans. Amer. Math. Soc. **72** (1952), 341–366. MR MR0047179 (13,839a)

[9] Yonina C. Eldar, *Least-squares inner product shaping*, Linear Algebra Appl. **348** (2002), 153–174. MR 2003f:65074

[10] ———, *von Neumann measurement is optimal for detecting linearly independent mixed quantum states*, Phys. Rev. A (3) **68** (2003), no. 5, 052303, 4. MR 2004k:81047

[11] Yonina C. Eldar and Helmut Bölcskei, *Geometrically uniform frames*, IEEE Trans. Inform. Theory **49** (2003), no. 4, 993–1006. MR 2004f:94015

[12] Yonina C. Eldar and G. David Forney, Jr., *On quantum detection and the square-root measurement*, IEEE Trans. Inform. Theory **47** (2001), no. 3, 858–872. MR 2002f:94001

[13] ———, *Optimal tight frames and quantum measurement*, IEEE Trans. Inform. Theory **48** (2002), no. 3, 599–610. MR 2003c:94006

[14] Seymour Goldberg, *Unbounded linear operators*, Dover Publications Inc., New York, 1985, Theory and applications, Reprint of the 1966 edition. MR MR810617 (86k:47001)

[15] David Griffiths, *Introduction to quantum mechanics*, Prentice Hall, New Jersey, 1995.

[16] David Halliday, Robert Resnick, and Jearl Walker, *Fundamentals of physics*, fifth ed., John Wiley & Sons, Inc., 1997.

[17] Paul Hausladen and William K. Wootters, *A "pretty good" measurement for distinguishing quantum states*, J. Modern Opt. **41** (1994), no. 12, 2385–2390. MR 95j:81030

[18] Carl W. Helstrom, *Quantum detection and estimation theory*, J. Statist. Phys. **1** (1969), 231–252. MR 40 #3855

[19] Carl W. Helstrom and Robert S. Kennedy, *Noncommuting observables in quantum detection and estimation theory*, IEEE Trans. Information Theory **IT-20** (1974), 16–24. MR 50 #16070

[20] Roderick B. Holmes and Vern I. Paulsen, *Optimal frames for erasures*, Linear Algebra Appl. **377** (2004), 31–51. MR MR2021601 (2004j:42028)

[21] Jerry B. Marion and Stephen T. Thornton, *Classical dynamics of particles and systems*, fourth ed., Harcourt Brace & Company, 1995.

[22] Soo-Change Pei and Min-Hung Yeh, *An introduction to discrete finite frames*, Signal Processing magazine IEEE **14** (1997), no. 6, 84–96.

[23] Asher Peres and Daniel R. Terno, *Optimal distinction between non-orthogonal quantum states*, J. Phys. A **31** (1998), no. 34, 7105–7111. MR 99f:81034

[24] Asher Peres and William K. Wootters, *Optimal detection of quantum information*, Phys. Rev. Lett. **66** (1991), 1119–1122.

[25] J. Preskill, *Lecture notes: http://www.theory.caltech.edu/people/preskill/ph229/.*

[26] Wulf Rossmann, *Lie groups*, Oxford Graduate Texts in Mathematics, vol. 5, Oxford University Press, Oxford, 2002, An introduction through linear groups. MR MR1889121 (2003f:22001)

[27] Walter Rudin, *Functional analysis*, second ed., International Series in Pure and Applied Mathematics, McGraw-Hill Inc., New York, 1991. MR MR1157815 (92k:46001)

[28] 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 MR1984549 (2004d:42053)

[29] A Tonomura, J Endo, T Matsuda, T Kawasaki, and H Ezawa, *Demonstration of single-electron build-up of an interference pattern*, American Journal of Physics **57** (1989), no. 1, 117–120.

[30] John von Neumann, *Mathematical foundations of quantum mechanics*, Princeton University Press, Princeton, 1955, Translated by Robert T. Beyer. MR 16,654a

[31] Horace P. Yuen, Robert S. Kennedy, and Melvin Lax, *Optimum testing of multiple hypotheses in quantum detection theory*, IEEE Trans. Information Theory **IT-21** (1975), 125–134. MR 53 #1370