2012

# Basis identification through convex optimization

Dominic Donald Kramer
*Iowa State University*

**Basis identification through convex optimization**

by

Dominic Kramer

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Mathematics

Program of Study Committee:

Eric Weber, Major Professor

Scott Hansen

Fritz Keinert

Anastasios Matzavinos

Justin Peters

Iowa State University

Ames, Iowa

2012

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# ACKNOWLEDGMENTS

I would like to thank my father Donald and my brother Donovan. You always encouraged me to chase my dreams, and you will always live on with me in my heart. Next, I would like to thank my mother Diane. You are the strongest person I know, and your constant love and support has been a guiding light. Last, I would like to thank my girlfriend Sijia. You are always there for me, always see the best in me, and if not for you, I would not be half the person I am today. I am truly blessed for having all of you in my life.

Next I would like to thank my advisor Dr. Eric Weber for your insight, advice, guidance, support, and patience throughout my graduate career. I am grateful for everything you have taught me, for your expert advice, and for helping me have the opportunity to travel overseas, visit China, and work with your collaborators at the National University of Singapore.

Last I would like to thank the members of my committee: Dr. Scott Hansen, Dr. Fritz Keinert, Dr. Anastasios Matzavinos, and Dr. Justin Peters for all your time and your help during my graduate career. In particular, I would like to thank each of you for all of the great advice you have given me.

# ABSTRACT

Suppose one has a highly redundant spanning set for a finite-dimensional Hilbert space, knows that a subset of the spanning set is an orthonormal basis for the space, and wants to identify that subset. To identify such a subset, most standard models typically require that the elements of the spanning set are arranged in a particular order. This dissertation develops a novel convex optimization problem, inspired by compressed sensing, and conjectures that the set of minimizers of this problem can be used to identify bases in a given spanning set. In particular, given an injective matrix $X$, consider the problem of minimizing $\|XY\|_1$, the sum of the absolute values of all entries in the matrix product $XY$, subject to $Y$ being a left inverse of $X$. This dissertation shows that for a given injective matrix $X$, the set of such minimizers is a nonempty, compact, convex set and conjectures that the extreme points of this set can be used to find a subset of the rows of $X$ that is a basis for the domain of $X$. An analysis of this conjecture is given, with particular attention given to the case when the rows of $X$ are a concatenation of two orthonormal bases. In this case, it is shown that if a left inverse $Z$ of $X$ is a minimizer of $\|XY\|_1$ subject to $Y$ being a left inverse of $X$, and $Z$ identifies an orthonormal basis in the rows of $X$, in a way made precise herein, then $Z$ is an extreme point of the set of minimizers. Furthermore, conditions are developed that ensure that such a left inverse $Z$ exists. Last, some special cases are developed where the above conjecture is shown to hold.

# CHAPTER 1. INTRODUCTION

If one is interested in a particular finite dimensional vector space, often the first step to analyze that space is to construct a basis for the space. It is well known from linear algebra that every vector space has a basis and there are numerous techniques to construct a basis for a given vector space. Performing the Gram-Schmidt algorithm on a sequence of vectors will construct a basis for the span of those vectors, and performing Gaussian elimination on a matrix provides a way to construct a basis for the range of the matrix. With these methods, however, one starts with a set of vectors and typically constructs a completely new set of vectors that are a basis for the appropriate space.

Suppose instead that one has a sequence of vectors that span a vector space and wishes to find a subset of that sequence that is a basis for the space. This request is more restrictive than merely wanting to construct a basis for a vector space since the vectors that can be in the basis are limited to those that are already in the sequence provided. Such a restriction arises in areas such as signal and image processing where one may know that certain types of signals can be represented in terms of a particular, linearly dependent, set of waveforms but needs to find a linearly independent subset of those particular waveforms that can represent the signals in question.

The goal of this work is to develop and explore a technique that is able to identify a subset of a spanning set that is a basis with nice properties. Specifically, there may be many subsets of the given spanning set for a space that form a basis for that space, especially if the number of elements in the spanning set is significantly higher than the dimension of the space. Therefore, the technique developed in this document attempts to find a "nice" basis. What constitutes a nice basis depends on the situation in which the basis will be used. However, if the spanning set does contain an orthonormal basis, that basis should be identified by the technique.

To analyze some possible well established techniques, consider first Gaussian elimination. Notice this technique will not identify an orthonormal basis in a spanning set that contains one because Gaussian elimination will only identify the orthonormal basis if the elements of the spanning set are ordered correctly.

For example, consider the spanning set $\{x_1, \ldots, x_n, e_1, \ldots, e_n\} \subseteq \mathbb{R}^n$ where $\{e_i\}_{i=1}^n$ is the standard orthonormal basis and $\{x_i\}_{i=1}^n$ is another, nonorthogonal, basis. Then identifying the pivot columns of the matrix formed by performing Gaussian elimination on the matrix

$$\begin{bmatrix} | & & | & | & & | \\ x_1 & \ldots & x_n & e_1 & \ldots & e_n \\ | & & | & | & & | \end{bmatrix}$$

will identify the nonorthogonal basis $\{x_i\}_{i=1}^n$ and not the orthonormal basis $\{e_i\}_{i=1}^n$.

Next, as already mentioned above, the Gram-Schmidt algorithm cannot be used since it takes a spanning set for a space and generally constructs a completely new set, not a subset of the original, that is an orthonormal basis for the space. Additionally, brute force methods quickly become impractical in high dimensional spaces.

The technique to identify a basis in a spanning set developed in this work is inspired by the work done in the area of compressed sensing. That is, recall a sequence is a spanning set for $\mathbb{R}^n$ if and only if $X \in \mathbb{R}^{m\times n}$ is an injective matrix where the rows of $X$ are the elements of the sequence. Next, if $X$ is injective then there exists at least one $Y \in \mathbb{R}^{m\times n}$ such that $YX = I$. Furthermore, if $X$ is not bijective then there are infinitely many such $Y$. Now consider the optimization problem

$$\text{minimize} \left\|XY\right\|_1 \quad \text{subject to} \quad YX = I \tag{1.1}$$

where the function $\left\|\cdot\right\|_1 : \mathbb{R}^{m\times m} \to \mathbb{R}$ is defined as

$$\left\|A\right\|_1 = \sum_{i,j=1}^{m} |A_{i,j}|$$

It will be shown that, given an injective matrix $X \in \mathbb{R}^{m\times n}$, there does exist a matrix $Y_* \in \mathbb{R}^{m\times n}$ that achieves the minimum in problem (1.1). That is,

$$\left\|XY_*\right\|_1 = \inf_{YX=I} \left\|XY\right\|_1$$

Thus the term *minimize* in the statement of problem (1.1) is justified. Note however, in general, problem (1.1) does not have a unique minimizer. Instead, the collection of minimizers of problem (1.1) forms a convex subset of $\mathbb{R}^{n\times m}$. It will be shown that this set of minimizers is compact and therefore, by the Krein-Milman Theorem, is the closed convex hull of its extreme points.

Now the main conjecture that will be analyzed in this work asserts that if the rows of $X$ are all of unit length, then there exists an extreme point of the set of minimizers of problem (1.1) that can be used to identify a subset of the rows of $X$ that forms a "nice" basis for $\mathbb{R}^n$.

To explain this further, consider the following example. Let

$$X = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$$

Notice then $X$ is injective and the matrices

$$Y_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}, \qquad Y_2 = \begin{bmatrix} 0 & 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 0 & 0 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}, \qquad \text{and} \qquad Y_3 = \begin{bmatrix} 2 & 0 & -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ 0 & 2 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}$$

are all left inverses of $X$. Next notice

$$\begin{aligned} XY_1 &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & 0 \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 & 0 \end{bmatrix} \end{aligned}$$

and thus

$$\|XY_1\|_1 = 2 + 4\frac{1}{\sqrt{2}} = 2 + 2\sqrt{2}$$

Similarly, $\|XY_2\|_1 = 2 + 2\sqrt{2}$, but $\|XY_3\|_1 = 6 + 6\sqrt{2} > 2 + 2\sqrt{2}$. Furthermore notice that compared to $Y_3$, the matrices $Y_1$ and $Y_2$ are special in that for each matrix, two columns are columns of zeros and the other two nonzero columns form a basis for $\mathbb{R}^2$. Because of this, $Y_1$

and $Y_2$ are said to *correspond to a basis*. That is, for example, $Y_2$ corresponds to the basis formed from the third and fourth rows of $X$ since the third and fourth columns of $Y_2$ are not columns of zeros.

Now notice $Y_3$ cannot be a minimizer of problem (1.1) since $\|XY_3\|_1 > \|XY_1\|$. In fact, as will be proven later, $Y_1$ and $Y_2$ are both minimizers of problem (1.1). In fact, for the matrix $X$, it will be shown that the matrices $Y_1$ and $Y_2$ are extreme points of the set of minimizers of problem (1.1). Thus, for this particular $X$, there exists an extreme point of the minimizers of problem (1.1) that corresponds to a basis in the rows of $X$.

Again, for a given injective matrix $X$, the set of solutions to problem (1.1) is infinite, but can be written as the closed convex hull of its extreme points by the Krein-Milman Theorem. The main conjecture this work will analyze is that, like the example above, at least some of the extreme points of the set of minimizers of problem (1.1) are matrices that correspond to a basis. Moreover, if the rows of $X$ are a concatenation of two orthonormal bases, like the example above, the conjecture is that the left inverses that correspond to these orthonormal bases are minimizers of problem (1.1) and are, in fact, extreme points of the set of minimizers of problem (1.1).

Problem (1.1) has a relationship to frame theory, convex optimization, and compressed sensing. The first three chapters of this work consist of a brief introduction to these areas. The fifth chapter thoroughly describes problem (1.1) and describes its connection to compressed sensing. Chapter four develops the tools from convex optimization that will be needed to analyze problem (1.1). Chapter six contains a series of results that show for particular injective matrices, problem (1.1) does identify bases as conjectured. Last, chapter seven is dedicated to the case where the rows of $X$ in problem (1.1) are a concatenation of two orthonormal bases. This chapter proves that if the rows of $X$ are a concatenation of two orthonormal bases then $X$ satisfies necessary conditions for problem (1.1) to identify these orthonormal bases as described above. It is also established in chapter seven that if there exists a minimizer to problem (1.1) that corresponds to an orthonormal basis, then it is necessarily an extreme point of the set of minimizers of problem (1.1).

# CHAPTER 2. Preliminaries

## 2.1 Injective Matrices and the Moore-Penrose Inverse

Notice a matrix $X \in F^{m\times n}$, where $F$ is some field, is injective if and only the rows of $X$ form a spanning set for $F^n$. Thus, although this work will be working with spanning sets in finite-dimensional Hilbert spaces, and identifying subsets of those spanning sets that are bases, most results will be written in terms of injective matrices.

An important result from linear algebra is that a matrix is injective if and only if it has a left inverse. This can be seen by observing that if $X \in \mathbb{F}^{m\times n}$, with $\mathbb{F}$ either $\mathbb{R}$ or $\mathbb{C}$, is injective then $X^*$ is surjective, where $X^*$ denotes the conjugate transpose of $X$. Hence there exists $y_i$ for $i = 1, \ldots, n$ such that $X^*y_i = e_i$ where $\{e_i\}_{i=1}^n$ is the standard orthonormal basis for $\mathbb{F}^n$. Thus if $Y$ is the matrix whose columns are $y_i$, then $X^*Y = I$, and hence $Y^*X = I$. Therefore $Y^*$ is a left inverse of $X$. Conversely, if $YX = I$ for some $Y$ then $x \in \ker X$ implies $x = YXx = Y(Xx) = 0$ and $X$ is injective.

Now given an injective matrix, a left inverse of that matrix that is of particular importance is its Moore-Penrose inverse. To derive this left inverse the following lemma will be needed.

**Lemma 2.1.1.** *Let $X \in \mathbb{F}^{m\times n}$ be an injective matrix. Then $X^*X$ is invertible.*

**Proof.**

Suppose $y \in \ker X^*X$. Then $X^*Xy = 0$ and therefore $Xy \in \ker X^* = (\operatorname{range} X)^\perp$. Hence $Xy \in \operatorname{range} X \cap (\operatorname{range} X)^\perp = 0$ since $Xy \in \operatorname{range} X$ by construction. Thus $Xy = 0$ and hence $y = 0$ since $X$ is injective. Therefore $X^*X$ is injective. Thus $(X^*X)^* = X^*X$ is surjective and is therefore bijective.

□

The above lemma justifies the following definition.

**Definition 2.1.1.** *The* Moore-Penrose Inverse *of an injective matrix $X \in \mathbb{F}^{m\times n}$ (with $\mathbb{F}$ either $\mathbb{R}$ or $\mathbb{C}$) is defined as*

$$X^\dagger = (X^*X)^{-1}X^*$$

The following result shows that $X^\dagger$ is a left inverse of $X$.

**Proposition 2.1.2.** *Let $X \in \mathbb{F}^{m\times n}$ (with $\mathbb{F}$ either $\mathbb{C}$ or $\mathbb{R}$) be an injective matrix. Then $X^\dagger Xy = y$ for all $y \in \mathbb{F}^n$.*

**Proof.**

Let $y \in \mathbb{R}^n$. Then by direction calculation,

$$X^\dagger Xy = ((X^*X)^{-1}X^*)Xy = (X^*X)^{-1}(X^*X)y = y$$

□

## 2.2 Normed Linear Spaces

Given a vector space $V$ over the field $\mathbb{F}$ (with $\mathbb{F}$ either $\mathbb{R}$ or $\mathbb{C}$) a *norm* on $V$ is a function $\|\cdot\| : V \to \mathbb{R}$ that satisfies the following conditions for all $x, y \in V$ and $\alpha \in \mathbb{F}$,

1. $\|x\| \geq 0$
2. $\|x\| = 0$ if and only if $x = 0$
3. $\|\alpha x\| = |\alpha| \cdot \|x\|$
4. $\|x + y\| \leq \|x\| + \|y\|$ (the Triangle Inequality)

A vector space with a norm defined on it is called a *normed linear space*.

### 2.2.1 Norms on Matrices

Notice if $\|\cdot\|$ is a norm on $\mathbb{F}^n$ (with $\mathbb{F}$ either $\mathbb{R}$ or $\mathbb{C}$) this norm can be extended to $\mathbb{F}^{m\times n}$ by thinking of a matrix in $\mathbb{F}^{m\times n}$ as a vector in $\mathbb{F}^{mn}$. This is illustrated formally in the proof of the next result.

**Proposition 2.2.1.** *Given two positive integers $m$ and $n$, let $\|\cdot\|$ be a norm on $\mathbb{F}^{mn}$. If $\psi : \mathbb{F}^{m\times n} \to \mathbb{F}^{mn}$ is a linear bijection, then $|||\cdot||| : \mathbb{F}^{m\times n} \to \mathbb{R}$ defined as $|||X||| = \|\psi(X)\|$ is a norm on $\mathbb{F}^{m\times n}$.*

**Proof.**

For any $X \in \mathbb{F}^{m\times n}$ one has that $|||X||| = \|\psi(X)\| \geq 0$ since $\|\cdot\|$ is a norm. Next if $0 = |||X||| = \|\psi(X)\|$ then $\psi(X) = 0$ and therefore $X = 0$ since $\psi$ is a bijection. Also, for any $\alpha \in \mathbb{F}$, one has that $|||\alpha X||| = \|\psi(\alpha X)\| = \|\alpha\psi(X)\| = |\alpha| \cdot \|\psi(X)\| = |\alpha| \cdot |||X|||$. Last, for any $Y \in \mathbb{F}^{m\times n}$,

$$|||X + Y||| = \|\psi(X + Y)\| = \|\psi(X) + \psi(Y)\| \leq \|\psi(X)\| + \|\psi(Y)\| = |||X||| + |||Y|||$$

Thus $|||\cdot|||$ is a norm on $\mathbb{F}^{m\times n}$.

□

Now consider the following norms on $\mathbb{F}^{m\times n}$ with $\mathbb{F}$ either $\mathbb{R}$ or $\mathbb{C}$.

**Definition 2.2.1.** *Define $\|\cdot\|_1, \|\cdot\|_2, \|\cdot\|_\infty : \mathbb{F}^{m\times n} \to \mathbb{R}$ (with $\mathbb{F}$ either $\mathbb{R}$ or $\mathbb{C}$) as*

$$\|X\|_1 = \sum_{j=1}^{n}\sum_{i=1}^{m} |X_{i,j}|$$

$$\|X\|_2 = \left(\sum_{j=1}^{n}\sum_{i=1}^{m} |X_{i,j}|^2\right)^{1/2}$$

$$\|X\|_\infty = \max\{|X_{i,j}| : i = 1, \ldots, m \text{ and } j = 1, \ldots, n\}$$

Using the previous result, the next result shows that all of the functions defined above are norms on $\mathbb{F}^{m\times n}$.

**Proposition 2.2.2.** *Each of the functions $\|\cdot\|_1, \|\cdot\|_2, \|\cdot\|_\infty : \mathbb{F}^{m\times n} \to \mathbb{R}$ are norms on $\mathbb{F}^{m\times n}$.*

**Proof.**

Let $\sigma : \{1, \ldots, mn\} \to \{1, \ldots, m\} \times \{1, \ldots, n\}$ be a bijection. Such a bijection exists since both sets in question are finite and have the same cardinality. Next define $\psi : \mathbb{F}^{m\times n} \to \mathbb{F}^{mn}$ by $\psi(X)_i = X_{\sigma(i)}$ for $i = 1, \ldots, mn$. Notice then if $\alpha \in \mathbb{F}$, $X \in \mathbb{F}^{m\times n}$, and $i = 1, \ldots, mn$ then $\psi(\alpha X)_i = (\alpha X)_{\sigma(i)} = \alpha X_{\sigma(i)} = \alpha\psi(X)_i$. Further if $X, Y \in \mathbb{F}^{m\times n}$ and $i = 1, \ldots, mn$ then

$\psi(X+Y)_i = (X+Y)_{\sigma(i)} = X_{\sigma(i)} + Y_{\sigma(i)} = \psi(X)_i + \psi(Y)_i$. Therefore $\psi$ is linear. Furthermore it is a bijection since $\sigma$ is a bijection.

Therefore since $\|X\|_i = \|\psi(X)\|_i$ for $i = 1, 2, \infty$ and any $X \in \mathbb{F}^{m\times n}$ where the right-hand norm is the standard $\ell_i$ norm in $\mathbb{F}^{mn}$ for $i = 1, 2, \infty$, it follows from the previous result that the functions in question are norms on $\mathbb{F}^{m\times n}$.

□

In addition to being a norm, the function $\|\cdot\|_1$ is submultiplicative.

**Proposition 2.2.3.** *The function $\|\cdot\|_1 : \mathbb{F}^{m\times n} \to \mathbb{R}$ (with $\mathbb{F}$ either $\mathbb{R}$ or $\mathbb{C}$) satisfies $\|XY\|_1 \leq \|X\|_1 \cdot \|Y\|_1$ for any $X \in \mathbb{F}^{m\times n}$ and $Y \in \mathbb{F}^{n\times N}$ where $N$ is any positive integer.*

**Proof.**

Let $X \in \mathbb{F}^{m\times n}$ and $Y \in \mathbb{R}^{n\times N}$. Then if $\widetilde{x}_j$ denotes the $j$th column of $X$ and $z_\ell$ denotes of $\ell$th row of $Z$,

$$\begin{aligned}
\|X\|_1 \cdot \|Y\|_1 &= \left(\sum_{j=1}^{n}\sum_{i=1}^{m} |X_{i,j}|\right)\left(\sum_{k=1}^{N}\sum_{\ell=1}^{n} |Z_{\ell,k}|\right) \\
&= \left(\sum_{j=1}^{n}\sum_{i=1}^{m} |X_{i,j}|\right)\left(\sum_{\ell=1}^{n}\sum_{k=1}^{N} |Z_{\ell,k}|\right) \\
&= \left(\sum_{j=1}^{n} \|\widetilde{x}_j\|_1\right)\left(\sum_{\ell=1}^{n} \|z_\ell\|_1\right) \\
&= \sum_{j=1}^{n} \|\widetilde{x}_j\|_1 \cdot \|z_j\|_1 + \sum_{\substack{j,\ell=1 \\ j\neq\ell}}^{n} \|\widetilde{x}_j\|_1 \cdot \|z_\ell\|_1 \\
&\geq \sum_{j=1}^{n} \|\widetilde{x}_j\|_1 \cdot \|z_j\|_1
\end{aligned}$$

and

$$\begin{aligned}
\|XY\|_1 &= \sum_{j=1}^{N}\sum_{i=1}^{m} |(XY)_{i,j}| \\
&= \sum_{j=1}^{N}\sum_{i=1}^{m} \left|\sum_{k=1}^{n} X_{i,k}Z_{k,j}\right| \\
&\leq \sum_{j=1}^{N}\sum_{i=1}^{m}\sum_{k=1}^{n} |X_{i,k}Z_{k,j}|
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{j=1}^{N} \sum_{k=1}^{n} \sum_{i=1}^{m} |X_{i,k} Z_{k,j}| \\
&= \sum_{j=1}^{N} \sum_{k=1}^{n} |Z_{k,j}| \sum_{i=1}^{m} |X_{i,k}| \\
&= \sum_{j=1}^{N} \sum_{k=1}^{n} \|\widetilde{x}_k\|_1 \cdot |Z_{k,j}| \\
&= \sum_{k=1}^{n} \sum_{j=1}^{N} \|\widetilde{x}_k\|_1 \cdot |Z_{k,j}| \\
&= \sum_{k=1}^{n} \|\widetilde{x}_k\|_1 \sum_{j=1}^{N} |z_{k,j}| \\
&= \sum_{k=1}^{n} \|\widetilde{x}_k\|_1 \cdot \|z_k\|_1
\end{aligned}
$$

Thus $\|XY\|_1 \leq \|X\|_1 \cdot \|Y\|_1$.

□

### 2.2.2 Norm Equivalence

The presence of a norm in a normed linear space introduces a topology to the space that allows one to define what it means for a sequence of elements in that space to converge to an element of that space. Specifically, a sequence $\{x_n\}_{n=1}^{\infty}$ in a normed linear space $V$ with norm $\|\cdot\|$ is said to *converge* to an element $x \in V$ if, given $\varepsilon > 0$, there exists $N$ such that $\|x_n - x\| < \varepsilon$ for all $n \geq N$.

It appears, at first sight, if a vector space has two norms defined on it then a sequence may converge with respect to one norm but not converge with respect to the other norm. If the vector space is infinite-dimensional this indeed may be the case. However, if the space is finite-dimensional the choice of norm does not affect the convergence of the sequence. The following makes this precise.

Two norms $\|\cdot\|_a$ and $\|\cdot\|_b$ defined a vector space $V$ are said to be *equivalent* if there exists constants $0 < A \leq B < \infty$ such that

$$A \|x\|_a \leq \|x\|_b \leq B \|x\|_a$$

for all $x \in V$.

**Proposition 2.2.4.** *Let $\|\cdot\|_a$ and $\|\cdot\|_b$ be two equivalent norms defined on a vector space $V$. Then a sequence $\{x_n\}_{n=1}^n$ converges to $x$ with respect to the norm $\|\cdot\|_a$ if and only if the sequence converges to $x$ with respect to the norm $\|\cdot\|_b$.*

**Proof.**

Let $\|\cdot\|_c$ and $\|\cdot\|_d$ be norms defined on $V$ and suppose there exists a constant $0 < C < \infty$ such that

$$\|x\|_c \leq C \|x\|_d$$

for all $x \in V$. Notice then if $\{x_n\}_{n=1}^\infty$ converges to $x$ with respect to $\|\cdot\|_d$ then $\{x_n\}_{n=1}^\infty$ converges to $x$ with respect to $\|\cdot\|_c$. To see why, suppose $\{x_n\}_{n=1}^\infty$ converges to $x$ with respect to $\|\cdot\|_d$. Then, given $\varepsilon > 0$, there exists $N$ such that $\|x_n - x\|_c < \varepsilon/C$ for all $n \geq N$. Thus

$$\|x_n - x\|_c \leq C \|x_n - x\|_d < \varepsilon$$

for all $n \geq N$. Therefore $\{x_n\}_{n=1}^\infty$ converges to $x$ with respect to $\|\cdot\|_c$.

Now because $\|\cdot\|_a$ and $\|\cdot\|_b$ are equivalent there exists constants $0 < A \leq B < \infty$ such that

$$A \|x\|_a \leq \|x\|_b \leq B \|x\|_a$$

for all $x \in V$. Therefore if $\{x_n\}_{n=1}^\infty$ converges to $x$ with respect to $\|\cdot\|_a$ then, by the above claim, $\{x_n\}_{n=1}^\infty$ converges to $x$ with respect to $\|\cdot\|_b$. Conversely since $A > 0$ notice

$$\|x\|_a \leq \frac{1}{A} \|x\|_b$$

for all $x \in V$. Therefore if $\{x_n\}_{n=1}^\infty$ converges to $x$ with respect to $\|\cdot\|_b$ then again, by the above claim, $\{x_n\}_{n=1}^\infty$ converges to $x$ with respect to $\|\cdot\|_a$.

□

The following well-known result states that any two norms on a finite-dimensional vector space are necessarily equivalent. For a proof of the result see Corollary 5.4.5 in [18].

**Corollary 5.4.5, [18].** *Let $V$ be a finite-dimensional vector space and let $\|\cdot\|_a$ and $\|\cdot\|_b$ be two norms defined on $V$. Then $\|\cdot\|_a$ and $\|\cdot\|_b$ are equivalent.*

### 2.2.3 Equicontinuity

This section defines what it means for a family of functions in a normed linear space to be *equicontinuous* and develops two results that will be important later. The definition provided is based on the one given in [26] and the proof of the results are based on those given in [13].

**Definition 2.2.2.** *A family of real-valued functions $\mathcal{F}$ in a normed linear space $X$ with norm $\|\cdot\|$ is said to be* equicontinuous *at $x \in X$ if, given $\varepsilon > 0$, there exists $\delta > 0$ such that for all $y \in X$ and all $f \in \mathcal{F}$ if $\|x - y\| < \delta$ then $|f(x) - f(y)| < \varepsilon$.*

**Proposition 2.2.5.** *If $\mathcal{F}$ is an equicontinuous family of real-valued functions on a normed linear space $X$ with norm $\|\cdot\|$, define the function*

$$g(x) := \inf_{f \in \mathcal{F}} f(x)$$

*If $g(x)$ is finite for each $x \in X$ then $g$ is continuous on $X$.*

**Proof.**

Let $x_0 \in X$ and $\varepsilon > 0$ be given. Then, because the family $\mathcal{F}$ is equicontinuous, there exists $\delta > 0$ such that if $\|x - x_0\| < \delta$ for some $x \in X$ then $|f(x) - f(x_0)| < \varepsilon/2$ for all $f \in \mathcal{F}$. Now by the construction of $g$ there exists $f, h \in \mathcal{F}$ such that

$$\begin{aligned} g(x_0) \leq & f(x_0) < g(x_0) + \frac{\varepsilon}{2} \\ g(x) \leq & \ h(x) < \ g(x) + \frac{\varepsilon}{2} \end{aligned}$$

Furthermore since $\mathcal{F}$ is equicontinuous on $X$

$$\begin{aligned} -\frac{\varepsilon}{2} < & f(x) - f(x_0) < \frac{\varepsilon}{2} \\ -\frac{\varepsilon}{2} < & h(x_0) - h(x) < \frac{\varepsilon}{2} \end{aligned}$$

Therefore

$$\begin{aligned} g(x) < \ f(x) < & f(x_0) + \frac{\varepsilon}{2} < g(x_0) + 2 \cdot \frac{\varepsilon}{2} = g(x_0) + \varepsilon \\ g(x_0) < & h(x_0) < \ h(x) + \frac{\varepsilon}{2} < \ h(x) + 2 \cdot \frac{\varepsilon}{2} = \ g(x) + \varepsilon \end{aligned}$$

That is $g(x) - g(x_0) < \varepsilon$ and $g(x_0) - g(x) < \varepsilon$ and hence $|g(x) - g(x_0)| < \varepsilon$. Since this is true for all $x \in X$ such that $\|x - x_0\| < \delta$ it follows that $g$ is continuous at $x_0$ and since $x_0 \in X$ was arbitrary, $g$ is continuous on $X$.

□

**Proposition 2.2.6.** *Let $X$ and $Y$ be finite dimensional normed linear spaces with norms $\|\cdot\|_X$ and $\|\cdot\|_Y$ respectively. Next let $A \subseteq X$ and $B \subseteq Y$ both be nonempty and let $f : A \times B \to \mathbb{R}$ be a uniformly continuous function on $A \times B$. Last define*

$$g(x) := \inf_{y \in B} f(x, y)$$

*Then if $g(x)$ is finite for all $x \in A$ then $g$ is continuous on $A$.*

**Proof.**

Since $X$ and $Y$ are finite dimensional normed linear spaces so is $X \times Y$ and thus all norms on $X \times Y$ are equivalent. For convenience define the norm $|||\cdot|||$ on $X \times Y$ as

$$|||(x, y)||| := \max\left\{\|x\|_X, \|y\|_Y\right\}$$

Next for $y \in B$ define $g_y(x) := f(x, y)$. Notice $\{g_y\}_{y \in B}$ is equicontinuous on $X$. To see why notice, given $\varepsilon > 0$, the fact that $f$ is uniformly continuous of $A \times B$ implies there exists $\delta > 0$ such that if $(x, y), (x', y') \in A \times B$ such that $|||(x, y) - (x', y')||| < \delta$ then $|f(x, y) - f(x', y')| < \varepsilon$.

In particular, given any $y \in B$, considering the points $(x, y)$ and $(x', y)$, one has if

$$\begin{aligned}
\|x - x'\|_X &= \max\left\{\|x - x'\|_X, \|0\|_Y\right\} \\
&= |||(x - x', y - y)||| \\
&= |||(x, y) - (x', y)||| \\
&< \delta
\end{aligned}$$

then $|g_y(x') - g_y(x)| = |f(x', y) - f(x, y)| < \varepsilon$. Therefore since $x \in A$ and $y \in B$ were arbitrary if follows that $\{g_y\}_{y \in B}$ is equicontinuous on $A$. Now notice

$$g(x) = \inf_{y \in B} f(x, y) = \inf_{y \in B} g_y(x)$$

Thus by the previous proposition, $g$ is continuous on $B$.

□

## 2.3 Hilbert Spaces

An *inner product* on the vector space $V$ over the field $\mathbb{F}$ (with $\mathbb{F}$ either $\mathbb{R}$ or $\mathbb{C}$) is a function $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{F}$ that satisfies the following conditions for all $x, y, z \in V$ and $\alpha, \beta \in \mathbb{F}$.

1. $\langle x, x \rangle \geq 0$
2. $\langle x, x \rangle = 0$ if and only if $x = 0$
3. $\langle x, y \rangle = \overline{\langle y, x \rangle}$
4. $\langle \alpha x + \beta y, z \rangle = \alpha \langle x, z \rangle + \beta \langle y, z \rangle$

A vector space on which an inner product is defined is called an *inner product space*. Note that an inner product on $V$ induces a norm defined as $\|x\| := \sqrt{\langle x, x \rangle}$. Therefore every inner product space is necessarily a normed linear space. The converse is not true in general.

Recall that a sequence $\{x_n\}_{i=1}^{\infty}$ in a normed linear space $V$ with norm $\|\cdot\|$ is said to converge to an element $x \in V$ if for every $\varepsilon > 0$ there exists an integer $N$ such that $\|x_n - x\| < \varepsilon$ for all $n \geq N$. If, in fact, for every $\varepsilon > 0$ there exists an integer $N$ such that that $\|x_n - x_m\| < \varepsilon$ for all $n, m \geq N$ then the sequence $\{x_n\}_{n=1}^{\infty}$ is said to be a *Cauchy sequence*.

Note that if a sequence is a Cauchy sequence in $V$ it need not converge to an element of $V$. For example, considering the vector space $\mathbb{Q}$, define the sequence $\{x_n\}_{n=1}^{\infty}$ as

$$x_n = \sum_{k=1}^{n} \frac{(-1)^k}{k}$$

Notice $x_n \in \mathbb{Q}$ for each positive integer $n$ as $x_n$ is a finite sum of rational numbers. However, as seen by examining the Maclaurin series of the function $\ln(\cdot)$,

$$\lim_{n \to \infty} x_n = \lim_{n \to \infty} \sum_{k=1}^{n} \frac{(-1)^k}{k} = \sum_{k=1}^{\infty} \frac{(-1)^k}{k} = \ln(2)$$

but $\ln(2) \notin \mathbb{Q}$. A normed linear space in which this does not occur is said to be *complete*. Specifically, a normed linear space $V$ is complete if every Cauchy sequence in $V$ converges to an element of $V$. An inner product space that is complete is of particular importance and is called a *Hilbert space*. The properties of these spaces are paramount to the rest of this work. The symbol $\mathcal{H}$ will be used to denote a general Hilbert space.

For a Hilbert space, the inner product on the space geometrically describes the angle between elements of the space. Furthermore its induced norm describes the lengths of elements. Last completeness is required in the definition of a Hilbert space to guarantee that Cauchy sequences in the space necessarily converge to an element in the space.

In a Hilbert space $\mathcal{H}$, two elements $x, y \in \mathcal{H}$ are said to be *orthogonal* if $\langle x, y\rangle = 0$. Geometrically this extends the concept of perpendicular vectors in two and three dimensional spaces. Next, an element $x \in \mathcal{H}$ is a unit vector, or is said to be *normalized*, if $\|x\| = 1$. Two vectors $x, y \in \mathcal{H}$ that are orthogonal and are both unit vectors are said to be *orthonormal*.

If $\mathcal{I}$ is an index set, a sequence $\{u_i\}_{i\in\mathcal{I}}$ in a Hilbert space $\mathcal{H}$ is an *orthonormal basis* for $\mathcal{H}$ if

$$\langle u_i, u_j\rangle = \delta_{i,j} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{else} \end{cases}$$

for each $i, j \in \mathcal{I}$ and for each $x \in \mathcal{H}$ there exist unique coefficients $\{a_i\}_{i\in\mathcal{I}}$ such that

$$x = \sum_{i\in\mathcal{I}} a_i u_i$$

[11].

A Hilbert space is said to be *separable* if it has a countable dense subset. A fundamental result is that every separable Hilbert space has an orthonormal basis. See Theorem 3.2.4 [11] for more details. Notice if $\mathcal{H}$ is finite-dimensional then it is necessarily separable.

Next, an orthonormal basis for a Hilbert space has the very important property that it satisfies *Parseval's Equation*,

$$x = \sum_{i\in\mathcal{I}} \langle x, u_i\rangle\, u_i \qquad \text{for all } x \in \mathcal{H} \tag{2.1}$$

Corollary 3.2.3, [11], as well as *Parseval's Identity*,

$$\|x\|^2 = \sum_{i\in\mathcal{I}} |\langle x, u_i\rangle|^2 \qquad \text{for all } x \in \mathcal{H} \tag{2.2}$$

Proposition 1.71, [17]. Thus every element of a Hilbert space can be written in an expansion in terms of a orthonormal basis $\{u_i\}_{i\in\mathcal{I}}$ and the coefficients of this expansion are calculated via inner products with the elements of the orthonormal basis. Furthermore the length of an

element is also calculated in terms of these inner products. Such expansions will be the topic of the next chapter.

# CHAPTER 3. Frame Theory

The fact that the coefficient representation of any element in a finite-dimensional Hilbert space $\mathcal{H}$, in terms of an orthonormal basis $\{u_i\}_{i=1}^n$ for $\mathcal{H}$, can be calculated solely using inner products with the elements of the orthonormal basis is very useful in areas such as image and signal processing where it is essential that coefficients can be calculated quickly and efficiently.

Specifically, to transmit an element $x \in \mathcal{H}$, its coefficients $\{\langle x, u_i\rangle\}_{i\in\mathcal{I}}$ are transmitted. A receiver, with a prior knowledge of the sequence $\{u_i\}_{i\in\mathcal{I}}$, then reconstructs the element from these coefficients using Parseval's equation (2.1). In reality though, during transmission, noise or errors can be introduced in the coefficients and, as such, the receiver does not receive the correct coefficients but instead receives a perturbed version of these coefficients. Furthermore, some coefficients may, in fact, be lost altogether. In the event that even a single coefficient is lost or corrupted, the fact that the elements of an orthonormal basis are linearly independent means that the element reconstructed from the coefficients will in general be very different from the transmitted element. For a more thorough description of this process see [14] or [6].

To remedy this, instead of using an orthonormal set, a set of elements $\mathbb{X} = \{x_i\}_{i=1}^m$ that are linearly dependent may be used to represent the element $x$. If $\mathbb{X}$ is a spanning set then there are coefficients $\{a_i\}_{i=1}^m$ such that

$$x = \sum_{i=1}^{m} a_i x_i$$

Furthermore, notice the linear dependence of the elements of $\mathbb{X}$ implies the coefficients $\{a_i\}_{i=1}^m$ contain redundancy. Thus, if noise is introduced into these coefficients upon transmission, perhaps the noise is spread across the redundant aspects of the coefficients so that the reconstructed element, although not necessarily the transmitted element, is close to the transmitted element.

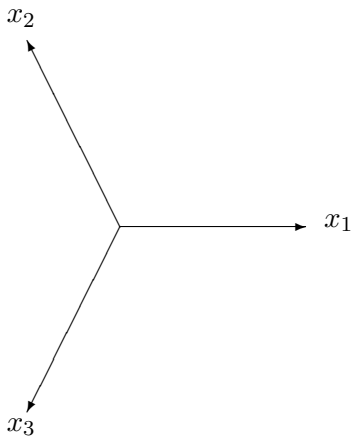


Figure 3.1 The Mercedes-Benz frame.

The above idea would be very beneficial since the addition of noise to the coefficients is inevitable in many real-world applications. However, if the linearly dependent spanning set $\mathbb{X}$ is used instead of the orthonormal basis, how can the coefficients representing an element be calculated? In particular, can the coefficients still be calculated only using inner products with elements of $\mathbb{X}$. The answer is, in fact, yes for a certain collection of linearly dependent spanning sets called Parseval frames.

## 3.1 The Mercedes-Benz Frame

Before defining what is a Parseval frame, consider the sequence of vectors in $\mathbb{R}^2$,

$$\{x_1, x_2, x_3\} = \left\{ \sqrt{\frac{2}{3}} \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \sqrt{\frac{2}{3}} \begin{bmatrix} -\frac{1}{2} \\ \frac{\sqrt{3}}{2} \end{bmatrix}, \sqrt{\frac{2}{3}} \begin{bmatrix} -\frac{1}{2} \\ -\frac{\sqrt{3}}{2} \end{bmatrix} \right\}.$$

Figure (3.1) illustrates these vectors. This sequence is called the Mercedes-Benz frame since graphically the elements of the sequence resemble the logo for the Mercedes-Benz Corporation. Notice for any $x = (a \;\; b)^T \in \mathbb{R}^n$

$$\begin{aligned} \sum_{i=1}^{3} \langle x, x_i \rangle &= \langle x, x_1 \rangle \, x_1 + \langle x, x_2 \rangle \, x_2 + \langle x, x_3 \rangle \, x_3 \\ &= \frac{2}{3} a \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \frac{2}{3} \left( -\frac{1}{2} a + \frac{\sqrt{3}}{2} b \right) \begin{bmatrix} -\frac{1}{2} \\ \frac{\sqrt{3}}{2} \end{bmatrix} + \frac{2}{3} \left( -\frac{1}{2} a - \frac{\sqrt{3}}{2} b \right) \begin{bmatrix} -\frac{1}{2} \\ -\frac{\sqrt{3}}{2} \end{bmatrix} \end{aligned}$$

$$
\begin{aligned}
&= \frac{2}{3}a \begin{bmatrix} 1+\frac{1}{4}+\frac{1}{4} \\ -\frac{\sqrt{3}}{4}+\frac{\sqrt{3}}{4} \end{bmatrix} + \frac{2}{3}b \begin{bmatrix} -\frac{\sqrt{3}}{4}+\frac{\sqrt{3}}{4} \\ \frac{3}{4}+\frac{3}{4} \end{bmatrix} \\
&= a \begin{bmatrix} 1 \\ 0 \end{bmatrix} + b \begin{bmatrix} 0 \\ 1 \end{bmatrix} \\
&= x
\end{aligned}
$$

Thus the elements of the Mercedes-Benz frame satisfy Parseval's Equation (2.1). That is the coefficients of $x$ can be directly calculated from the inner products of $x$ with the elements of the Mercedes-Benz frame just as with an orthonormal basis. However, notice the Mercedes-Benz frame is not an orthonormal basis. The above work describes a reconstruction formula for the Mercedes-Benz frame since it shows how any element $x \in \mathcal{H}$ can be reconstructed from the coefficients $\{\langle x, x_i\rangle\}_{i=1}^{3}$. Next, notice

$$
\begin{aligned}
\sum_{i=1}^{3} |\langle x, x_i\rangle|^2 &= \frac{2}{3}\left(a^2 + \left(-\frac{1}{2}a + \frac{\sqrt{3}}{2}b\right)^2 + \left(-\frac{1}{2}a - \frac{\sqrt{3}}{2}b\right)^2\right) \\
&= \frac{2}{3}\left(a^2 + \frac{1}{4}a^2 - \frac{\sqrt{3}}{2}ab + \frac{3}{4}b^2 + \frac{1}{4}a^2 + \frac{\sqrt{3}}{2}ab + \frac{3}{4}b^2\right) \\
&= a^2 + b^2 \\
&= \|x\|^2
\end{aligned}
$$

Thus the Mercedes-Benz frame also satisfies Parseval's Identity (2.2). A natural question then is, does a sequence satisfy Parseval's Identify if and only if it satisfies Parseval's Equation (2.1), just as the Mercedes-Benz frame. The answer is, in fact, yes and this prompts the development of Parseval frames.

## 3.2 Parseval Frames

**Definition 3.2.1.** *For a countable index set $\mathcal{I}$, a sequence $\{x_i\}_{i\in\mathcal{I}}$ of elements in a Hilbert space $\mathcal{H}$ is a* Parseval frame *for $\mathcal{H}$ if for every $x \in \mathcal{H}$,*

$$\|x\|^2 = \sum_{i\in\mathcal{I}} |\langle x, x_i\rangle|^2$$

Notice the Mercedes-Benz frame is a Parseval frame, and it turns out every Parseval frame has a nice reconstruction formula like the Mercedes-Benz frame. First, however, notice that

the definition of a Parseval frame does not assume $\mathcal{H}$ is finite-dimensional. However, since this work will be focused only on finite-dimensional Hilbert spaces, particularly $\mathbb{R}^n$, the results henceforth will be developed only for finite-dimensional Hilbert spaces. Many of the results ahead also can be extended to the infinite-dimensional case, and when a result cannot be, a note will be given. Generally in infinite-dimensional Hilbert spaces, care must be taken when dealing with convergence issues, and these issues do not appear in the finite-dimensional case.

**Proposition 3.2.1. [Proposition 3.11 [17]]** *Let $\mathbb{X} = \{x_i\}_{i=1}^k$ be a sequence of elements in a finite-dimensional Hilbert space $\mathcal{H}$. Then $\mathbb{X}$ is a Parseval frame if and only if for every $x \in \mathcal{H}$*

$$x = \sum_{i=1}^{k} \langle x, x_i \rangle \, x_i \tag{3.1}$$

**Proof.**

For the reverse direction, suppose $x \in \mathcal{H}$ satisfies equation (3.1). Then

$$\begin{aligned} \|x\|^2 &= \langle x, x \rangle \\ &= \left\langle \sum_{i=1}^{k} \langle x, x_i \rangle \, x_i, x \right\rangle \\ &= \sum_{i=1}^{k} \langle x, x_i \rangle \, \langle x_i, x \rangle \\ &= \sum_{i=1}^{k} \langle x, x_i \rangle \, \overline{\langle x, x_i \rangle} \\ &= \sum_{i=1}^{k} |\langle x, x_i \rangle|^2 \end{aligned}$$

Conversely let $\mathbb{F}$ (with $\mathbb{F}$ either $\mathbb{R}$ or $\mathbb{C}$) denote the field of scalars for $\mathcal{H}$ and define $\Theta : \mathcal{H} \to \mathbb{F}^k$ as

$$\Theta x = \{\langle x, x_i \rangle\}_{i=1}^k = \sum_{i=1}^{n} \langle x, x_i \rangle \, e_i$$

where $\{e_i\}_{i=1}^n$ is the standard orthonormal basis. Notice then if $\mathbb{X}$ is a Parseval frame for $\mathcal{H}$ then

$$\|x\|^2 = \sum_{i=1}^{k} |\langle x, x_i \rangle|^2 = \|\Theta x\|^2$$

for all $x \in \mathcal{H}$. Therefore $\Theta$ is an isometry (preserves lengths) and thus preserves inner products. Thus if $\{u_j\}_{j=1}^k$ is an orthonormal basis for $\mathcal{H}$, it follows for any $x \in \mathcal{H}$,

$$\begin{aligned}
x &= \sum_{j=1}^{k} \langle x, u_j \rangle \, u_j \\
&= \sum_{j=1}^{k} \langle \Theta x, \Theta u_j \rangle \, u_j \\
&= \sum_{j=1}^{k} \left\langle \Theta x, \sum_{i=1}^{n} \langle u_j, x_i \rangle \, e_i \right\rangle u_j \\
&= \sum_{j=1}^{k} \sum_{i=1}^{n} \langle \Theta x, e_i \rangle \, \overline{\langle u_j, x_i \rangle} u_j \\
&= \sum_{j=1}^{k} \sum_{i=1}^{n} \langle x, \Theta^* e_i \rangle \, \langle x_i, u_j \rangle \, u_j \\
&= \sum_{j=1}^{k} \langle x, x_i \rangle \sum_{i=1}^{n} \langle x_i, u_j \rangle \, u_j \\
&= \sum_{j=1}^{k} \langle x, x_i \rangle \, x_i
\end{aligned}$$

□

Therefore if a sequence is a Parseval frame for a Hilbert space then one can expand any element of the Hilbert space in terms the elements in the Parseval frame, and the coefficients in this expansion can be obtained directly from taking inner products with the elements of the Parseval frame. Furthermore, to be a Parseval frame a sequence need not be an orthonormal basis as illustrated by the Mercedes-Benz frame. Thus it is possible to construct a Parseval frame with redundancy that still has the nice reconstruction properties of orthonormal bases.

Notice an orthonormal basis is necessarily a Parseval frame, and Parseval frames generalize orthonormal bases since a Parseval frame satisfies Parseval's Equation (2.1) and Parseval's Identity (2.2) as shown above. One may then ask if there is a way to generalize general bases just as Parseval frames generalize orthonormal bases. This is the topic of the next section.

## 3.3 General Frames

**Definition 3.3.1.** *For a countable index set $\mathcal{I}$, a sequence $\{x_i\}_{i\in\mathcal{I}}$ of elements in a Hilbert space $\mathcal{H}$ is a* frame *for the space if there exists constants $0 < A \leq B < \infty$ such that for all $x \in \mathcal{H}$*

$$A\|x\|^2 \leq \sum_{i\in\mathcal{I}} |\langle x, x_i\rangle|^2 \leq B\|x\|^2$$

Notice setting $A = B = 1$ recovers the definition of a Parseval frame. Thus a Parseval frame is a frame, as defined above. The definition of a frame is given the way it is so that certain operators constructed from the frame are bounded. These operators will be discussed later. For now notice the definition generalizes the definition of a Parseval frame. It turns out this generality allows greater flexible for a sequence to be a frame, but still has enough structure to guarantee a frame has a reconstruction formula. This formula, however, will turn out to be more general than the reconstruction formula for Parseval frames.

Again the definition of a frame does not require the Hilbert space in question to be finite-dimensional. However, since this work will be working with finite-dimensional spaces, the following results will focus on finite-dimensional spaces. In particular, a sequence is a frame for a finite-dimensional Hilbert space if and only if it is a spanning set for that space. This is not true, however, infinite-dimensional Hilbert spaces. Given a sequence $\mathbb{X} = \{x_i\}_{i=1}^k$ in finite-dimensional Hilbert space $\mathcal{H}$ over the field $\mathbb{F}$ (with $\mathbb{F}$ either $\mathbb{R}$ or $\mathbb{C}$) the span of $\mathbb{X}$ is defined as

$$\operatorname{span}\{x_i\}_{i=1}^k := \left\{ \sum_{i=1}^k a_i x_i \,:\, a_i \in \mathbb{F} \text{ for all } i = 1, \ldots, k \right\}.$$

**Proposition 3.3.1. [Proposition 3.18 [17]]** *A sequence $\{x_i\}_{i=1}^k$ in a finite-dimensional Hilbert space $\mathcal{H}$ over the field $\mathbb{F}$ (with $\mathbb{F}$ either $\mathbb{R}$ or $\mathbb{C}$) is a frame for $\mathcal{H}$ if and only if* $\operatorname{span}\{x_i\}_{i=1}^k = \mathcal{H}$.

**Proof.**

Let $x \in \mathcal{H}$ and notice by the Cauchy-Schwarz inequality,

$$\sum_{i=1}^k |\langle x, x_i\rangle|^2 \leq \sum_{i=1}^k \|x\|^2 \cdot \|x_i\|^2 = B\|x\|^2$$

where

$$B = \sum_{i=1}^{k} \|x_i\|^2$$

Therefore, in a finite-dimensional Hilbert space, a sequence necessarily has a finite upper bound in definition (3.3.1). Now let $\Theta \in \mathbb{F}^{k \times \dim \mathcal{H}}$ be the matrix whose $i$th row is $x_i^*$ and notice

$$\sum_{i=1}^{k} |\langle x, x_i \rangle|^2 = \|\Theta x\|^2$$

Thus the proposition asserts that $\Theta^*$ is surjective if and only if there exists $0 < A < \infty$ such that

$$\|\Theta x\|^2 \geq A$$

for all $x \in \mathcal{H}$ with $\|x\| = 1$. Thus, to prove the contrapositive, $\Theta^*$ is not surjective if and only if $\Theta$ is not injective if and only if there exists $u \in \mathcal{H}$ with $u \neq 0$ such that $\Theta u = 0$ if and only if there exists $v \in \mathcal{H}$ with $\|v\| = 1$ such that $\Theta v = 0$.

Next if there exists $v \in \mathcal{H}$ with $\|v\| = 1$ such that $\Theta v = 0$ then there cannot exist a constant $0 < A < \infty$ such that $\|\Theta x\|^2 \geq A$ for all $x \in \mathcal{H}$ with $\|x\| = 1$.

Conversely, if there does not exist a constant $0 < A < \infty$ such $\|\Theta x\|^2 \geq A$ for all $x \in \mathcal{H}$ with $\|x\| = 1$ then there exists a sequence $\{w_n\}_{n=1}^{\infty} \subseteq \mathcal{H}$ with $\|w_n\| = 1$ for all $n$ such that $\lim_{n\to\infty} \|\Theta w_n\| = 0$. Next since the sequence $\{w_n\}_{n=1}^{\infty}$ is bounded, by construction, it contains a convergent subsequence by the Bolzano-Weierstrass Theorem. Let $\{w_{n_i}\}_{i=1}^{\infty}$ denote this subsequence and let $w \in \mathcal{H}$ denote the element it converges to. Then by the continuity of $\|\cdot\|$,

$$0 = \lim_{i\to\infty} \|\Theta w_{n_i}\| = \left\| \Theta \lim_{i\to\infty} w_{n_i} \right\| = \|\Theta w\|$$

Hence $\Theta w = 0$. Next

$$\|w\| = \left\| \lim_{i\to\infty} w_{n_i} \right\| = \lim_{i\to\infty} \|w_{n_i}\| = 1$$

Therefore there exists $w \in \mathcal{H}$ with $\|w\| = 1$ such that $\Theta w = 0$.

□

To develop the reconstruction formula that frames posses, some operators will be needed. First, given a sequence $\mathbb{X} = \{x_i\}_{i=1}^{k}$ (not necessarily a frame) in a finite-dimensional Hilbert

space $\mathcal{H}$ over the field $\mathbb{F}$ (with $\mathbb{F}$ either $\mathbb{R}$ or $\mathbb{C}$) the *analysis operator* $\Theta_{\mathbb{X}} : \mathcal{H} \to \mathbb{F}^k$ of $\mathbb{X}$ is defined as

$$\Theta_{\mathbb{X}} x := \{\langle x, x_i \rangle\}_{i=1}^k$$

for $x \in \mathcal{H}$. If $\{e_i\}_{i=1}^n$ is the standard orthonormal basis for $\mathbb{F}^k$ defined by setting $(e_i)_j = \delta_{i,j}$ where $\delta_{i,j}$ denotes the Kronecker delta then, in terms of this basis, $\Theta_{\mathbb{X}}$ has the matrix form

$$\begin{bmatrix} - & x_1^* & - \\ & \vdots & \\ - & x_k^* & - \end{bmatrix}$$

where $x^*$ indicates conjugate transpose of the vector $x$. To see why this is the case notice for fixed $i$ and $j$,

$$\begin{aligned} (\Theta_{\mathbb{X}})_{i,j} &= e_i^T \Theta_{\mathbb{X}} e_j \\ &= \langle \Theta_{\mathbb{X}} e_j, e_i \rangle \\ &= \left\langle \{\langle e_j, x_k \rangle\}_{k=1}^k, e_i \right\rangle \\ &= \langle e_j, x_i \rangle \\ &= \overline{(x_i)_j} \end{aligned}$$

Notice if $\mathbb{X}$ spans $\mathcal{H}$ then $\Theta_{\mathbb{X}}$ is injective. To see why notice if $x \in \ker \Theta_{\mathbb{X}}$ then $0 = \Theta_{\mathbb{X}} x = \{\langle x, x_i \rangle\}_{i=1}^k$ and hence $\langle x, x_i \rangle = 0$ for all $i = 1, \ldots, k$. Now let $y \in \mathcal{H}$. Then since $\mathbb{X}$ spans $\mathcal{H}$ there exists $\{a_i\}_{i=1}^k$ such that

$$y = \sum_{i=1}^k a_i x_i$$

Therefore

$$\langle x, y \rangle = \left\langle x, \sum_{i=1}^k a_i x_i \right\rangle = \sum_{i=1}^k a_i \langle x, x_i \rangle = 0$$

Thus $\langle x, y \rangle = 0$ for all $y \in \mathcal{H}$. Now let $\{u_i\}_{i=1}^n$ be an orthonormal basis for $\mathcal{H}$. Then $\langle x, u_i \rangle = 0$ for all $i = 1, \ldots, n$. Next notice

$$x = \sum_{i=1}^n \langle x, u_i \rangle u_i$$

Thus

$$\langle x, x \rangle = \left\langle x, \sum_{i=1}^n \langle x, u_i \rangle u_i \right\rangle = \sum_{i=1}^n \langle x, u_i \rangle \langle x, u_i \rangle = 0$$

Hence $x = 0$ and therefore $\Theta_{\mathbb{X}}$ is injective.

The operator $\Theta_{\mathbb{X}}^* : \mathbb{F}^n \to \mathcal{H}$, the adjoint of the analysis operator of $\mathbb{X}$, is called the *synthesis operator* of $\mathbb{X}$. Hence the matrix representation of this operator is

$$\begin{bmatrix} | & & | \\ x_1 & \dots & x_k \\ | & & | \end{bmatrix}$$

and therefore for any $a = \{a_i\}_{i=1}^k \in \mathbb{F}^k$ one has

$$\Theta_{\mathbb{X}}^* a = \sum_{i=1}^k a_i x_i$$

Also notice $\Theta_{\mathbb{X}}^*$ is surjective since $\Theta_{\mathbb{X}}$ is injective.

Next, the operator $S_{\mathbb{X}} : \mathcal{H} \to \mathcal{H}$ defined as $S_{\mathbb{X}} := \Theta_{\mathbb{X}}^* \Theta_{\mathbb{X}}$ is called the *frame operator* of $\mathbb{X}$. Notice since $\Theta_{\mathbb{X}}$ is injective it follows from Proposition (2.1.1) that $S_{\mathbb{X}}$ is bijective. Furthermore notice $S_{\mathbb{X}}$ is self-adjoint.

The following result will be needed to construct the reconstruction formula for general frames. The proof of the result makes heavy use of the operators defined above.

**Proposition 3.3.2. [Proposition 3.19 [17]]** *Let $\mathbb{X} = \{x_i\}_{i=1}^k$ be a sequence in a finite-dimensional Hilbert space $\mathcal{H}$ and suppose there exists a sequence $\{y_i\}_{i=1}^k \subseteq \mathcal{H}$ such that for all $x \in \mathcal{H}$,*

$$x = \sum_{i=1}^k \langle x, y_i \rangle \, x_i$$

*Then, in addition,*

$$x = \sum_{i=1}^k \langle x, x_i \rangle \, y_i$$

*for all $x \in \mathcal{H}$ and $\mathbb{Y}$ is a frame for $\mathcal{H}$.*

**Proof.**

Let $\mathbb{Y}$ denote the sequence $\{y_i\}_{i=1}^k$. Notice then, by assumption, for all $x \in \mathcal{H}$,

$$\Theta_{\mathbb{X}}^* \Theta_{\mathbb{Y}} x = \Theta_{\mathbb{X}} \{\langle x, y_i \rangle\}_{i=1}^k = \sum_{i=1}^k \langle x, y_i \rangle \, x_i = x$$

Thus $\Theta_{\mathbb{X}}^* \Theta_{\mathbb{Y}} = I$ the identity operator. Hence $I = I^* = (\Theta_{\mathbb{X}}^* \Theta_{\mathbb{Y}})^* = \Theta_{\mathbb{Y}}^* \Theta_{\mathbb{X}}$. That is for all $x \in \mathcal{H}$,

$$x = \Theta_{\mathbb{Y}}^* \Theta_{\mathbb{X}} x = \Theta_{\mathbb{Y}}^* \{\langle x, x_i \rangle\}_{i=1}^k = \sum_{i=1}^k \langle x, x_i \rangle \, y_i$$

The fact that $\mathbb{Y}$ is a frame for $\mathcal{H}$ follows from Proposition (3.3.1) since, by assumption, $\mathbb{X}$ is a spanning set for $\mathcal{H}$.

□

The next result provides a reconstruction formula for a general frame.

**Proposition 3.3.3.** *Let $\mathbb{X} = \{x_i\}_{i=1}^k$ be a frame for a finite-dimensional Hilbert space $\mathcal{H}$ and let $y_i = S_{\mathbb{X}}^{-1} x_i$ for all $i = 1, \ldots, k$. Then for all $x \in \mathcal{H}$,*

$$x = \sum_{i=1}^{k} \langle x, y_i \rangle \, x_i = \sum_{i=1}^{k} \langle x, x_i \rangle \, y_i$$

**Proof.**

Let $\mathbb{Y}$ denote the sequence $\{y_i\}_{i=1}^k$. As noted above, $S_{\mathbb{X}}$ is bijective and thus $\mathbb{Y}$ is well-defined. Next for any $x \in \mathcal{H}$,

$$\begin{aligned}
x &= S_{\mathbb{X}}^{-1} S_{\mathbb{X}} x \\
&= S_{\mathbb{X}}^{-1} \Theta_{\mathbb{X}}^* \Theta_{\mathbb{X}} x \\
&= S_{\mathbb{X}}^{-1} \Theta_{\mathbb{X}}^* \{\langle x, x_i \rangle\}_{i=1}^k \\
&= S_{\mathbb{X}}^{-1} \sum_{i=1}^{k} \langle x, x_i \rangle \, x_i \\
&= \sum_{i=1}^{k} \langle x, x_i \rangle \, S_{\mathbb{X}}^{-1} x_i \\
&= \sum_{i=1}^{k} \langle x, x_i \rangle \, y_i
\end{aligned}$$

Therefore, since $S_{\mathbb{X}}$ is self-adjoint,

$$\begin{aligned}
S_{\mathbb{X}}^{-1} x &= \sum_{i=1}^{k} \left\langle S_{\mathbb{X}}^{-1} x, x_i \right\rangle y_i \\
&= \sum_{i=1}^{k} \left\langle x, (S_{\mathbb{X}}^{-1})^* x_i \right\rangle y_i \\
&= \sum_{i=1}^{k} \left\langle x, (S_{\mathbb{X}}^*)^{-1} x_i \right\rangle y_i \\
&= \sum_{i=1}^{k} \left\langle x, S_{\mathbb{X}}^{-1} x_i \right\rangle y_i
\end{aligned}$$

$$= \sum_{i=1}^{k} \langle x, y_i \rangle \, y_i$$
$$= \sum_{i=1}^{k} \langle x, y_i \rangle \, S_{\mathbb{X}}^{-1} x_i$$

Thus

$$\begin{aligned} x &= S_{\mathbb{X}} S_{\mathbb{X}}^{-1} x \\ &= S_{\mathbb{X}} \sum_{i=1}^{k} \langle x, y_i \rangle \, S_{\mathbb{X}}^{-1} x_i \\ &= \sum_{i=1}^{k} \langle x, y_i \rangle \, S_{\mathbb{X}} S_{\mathbb{X}}^{-1} x_i \\ &= \sum_{i=1}^{k} \langle x, y_i \rangle \, x_i \end{aligned}$$

□

The above two propositions show that, in a finite-dimensional Hilbert space $\mathcal{H}$, a sequence $\mathbb{X} = \{x_i\}_{i=1}^{k} \subseteq \mathcal{H}$ is a frame for $\mathcal{H}$ if and only if there exists a sequence $\mathbb{Y} = \{y_i\}_{i=1}^{k} \subseteq \mathcal{H}$ such that for all $x \in \mathcal{H}$,

$$x = \sum_{i=1}^{k} \langle x, y_i \rangle \, x_i = \sum_{i=1}^{k} \langle x, x_i \rangle \, y_i$$

Any such sequence $\mathbb{Y}$ is called a *dual frame* of $\mathbb{X}$. The fact that such a sequence $\mathbb{Y}$ is a frame follows from Proposition (3.3.2). Furthermore, the above proposition shows that every frame does, in fact, have a dual frame. That is, the sequence $\left\{S_{\mathbb{X}}^{-1} x_i\right\}_{i=1}^{k}$ in Proposition (3.3.3) is a dual frame of $\mathbb{X}$ called the *canonical dual frame* of $\mathbb{X}$. Any dual frame for a frame that is not the canonical dual is called an *alternate dual*. Notice the proof of Proposition (3.3.2) shows that $\mathbb{Y}$ is a dual frame of $\mathbb{X}$ if and only if $\Theta_{\mathbb{Y}}^{*}$ is a left inverse of $\Theta_{\mathbb{X}}$, or equivalently, $\Theta_{\mathbb{X}}^{*}$ is a left inverse of $\Theta_{\mathbb{Y}}$.

Next notice by Proposition (3.3.1), in a finite-dimensional Hilbert space, a basis is necessarily a frame. Furthermore, a frame has a unique dual frame if and only if it is a basis. Otherwise it has infinitely many dual frames. This is made precise in the next two propositions. In the case when a sequence $\mathbb{X}$ is a basis for a finite-dimensional Hilbert space, its unique dual frame is called the *dual basis* of $\mathbb{X}$.

**Proposition 3.3.4. [Proposition 6.3 [17]]** *A frame $\{x_i\}_{i=1}^k$ for a finite-dimensional Hilbert space over $\mathbb{F}$ (with $\mathbb{F}$ either $\mathbb{R}$ or $\mathbb{C}$) has a unique dual frame if and only if only if it is a basis.*

**Proof.**

Let $\mathbb{X} = \{x_i\}_{i=1}^k$ denote the frame in question. Then the above two propositions show there exists a sequence $\mathbb{Y}$ that is a dual of $\mathbb{X}$. Thus $\Theta_{\mathbb{Y}}^*\Theta_{\mathbb{X}} = I$.

Now set $n = \dim \mathcal{H}$ and let $\{a_i\}_{i=1}^n \subseteq \mathbb{F}^k$ such that $a_i \in \ker \Theta_{\mathbb{X}}^*$ for all $i = 1, \ldots, n$. Next let $A \in \mathbb{F}^{k\times n}$ be such that the $i$th column of $A$ is $a_i$. Then $\Theta_{\mathbb{X}}^* A = 0$ and hence

$$(\Theta_{\mathbb{Y}}^* + A^*)\Theta_{\mathbb{X}} = \Theta_{\mathbb{Y}}^*\Theta_{\mathbb{X}} + A^*\Theta_{\mathbb{X}} = I$$

Thus the columns of $\Theta_{\mathbb{Y}}^* + A^*$ are a dual of $\mathbb{X}$. Conversely suppose $\mathbb{W} = \{w_i\}_{i=1}^k$ is a dual of $\mathbb{X}$. Then $\Theta_{\mathbb{W}}^*\Theta_{\mathbb{X}} = I$. Thus

$$(\Theta_{\mathbb{W}}^* - \Theta_{\mathbb{Y}}^*)\Theta_{\mathbb{X}} = \Theta_{\mathbb{W}}^*\Theta_{\mathbb{X}} - \Theta_{\mathbb{Y}}^*\Theta_{\mathbb{X}} = I - I = 0$$

Thus $\Theta_{\mathbb{W}} = \Theta_{\mathbb{Y}} + (\Theta_{\mathbb{W}} - \Theta_{\mathbb{Y}})$ where $\Theta_{\mathbb{W}} - \Theta_{\mathbb{Y}} \in \ker \Theta_{\mathbb{X}}^*$.

Therefore the columns of a matrix $U$ form a dual of $\mathbb{X}$ if and only if $U = \Theta_{\mathbb{Y}}^* + V$ such that $\Theta_X^* V = 0$. Thus if $\mathbb{X}$ is a basis then $\ker \Theta_X^* = 0$ and therefore $\mathbb{X}$ has only one dual, namely $\mathbb{Y}$. Otherwise, if $\mathbb{X}$ is not a basis then because $\mathbb{X}$ is a spanning set $\Theta_{\mathbb{X}}^*$ has infinitely many elements. Hence $\mathbb{X}$ has infinitely many duals.

□

The next result characterizes all duals of a frame.

**Proposition 3.3.5. [Proposition 6.4 [17]]** *Let $\mathbb{X} = \{x_i\}_{i=1}^k$ be a frame for a finite-dimensional Hilbert space. Then $\{y_i\}_{i=1}^k$ is a dual frame of $\mathbb{X}$ if and only if $y_i = S_{\mathbb{X}}^{-1}x_i + z_i$ for $i = 1, \ldots, k$ for some sequence $\mathbb{Z} = \{z_i\}_{i=1}^k$ such that $\Theta_{\mathbb{X}}^*\Theta_{\mathbb{Z}} = 0$.*

**Proof.**

For the forward direction suppose $\mathbb{Y} = \{y_i\}_{i=1}^k$ is a dual frame of $\mathbb{X}$ and define $\mathbb{Z} = \{z_i\}_{i=1}^k$ as $z_i = y_i - S_{\mathbb{X}}^{-1}x_i$ for $i = 1, \ldots, k$. Notice then $y_i = S_{\mathbb{X}}^{-1}x_i + z_i$ for $i = 1, \ldots, k$. That is, $\Theta_{\mathbb{Y}}^* = S_{\mathbb{X}}^{-1}\Theta_{\mathbb{X}}^* + \Theta_{\mathbb{Z}}^*$ and thus $\Theta_{\mathbb{Z}}^* = \Theta_{\mathbb{Y}}^* - S_{\mathbb{X}}^{-1}\Theta_{\mathbb{X}}^*$. Therefore since $\Theta_{\mathbb{X}}^*\Theta_{\mathbb{Y}} = I$, because $\mathbb{Y}$ is a

dual frame of $\mathbb{X}$, it follows that

$$\begin{aligned}
\Theta_{\mathbb{X}}^*\Theta_{\mathbb{Z}} &= \Theta_{\mathbb{X}}^*(\Theta_{\mathbb{Y}}^* - S_{\mathbb{X}}^{-1}\Theta_{\mathbb{X}}^*)^* \\
&= \Theta_{\mathbb{X}}^*(\Theta_{\mathbb{Y}} - \Theta_{\mathbb{X}}S_{\mathbb{X}}^{-1}) \\
&= \Theta_{\mathbb{X}}^*\Theta_{\mathbb{Y}} - \Theta_{\mathbb{X}}^*\Theta_{\mathbb{X}}S_{\mathbb{X}}^{-1} \\
&= \Theta_{\mathbb{X}}^*\Theta_{\mathbb{Y}} - S_{\mathbb{X}}S_{\mathbb{X}}^{-1} \\
&= I - I \\
&= 0
\end{aligned}$$

For the reverse direction suppose let $\mathbb{Z} = \{z_i\}_{i=1}^k$ be such that $\Theta_{\mathbb{X}}^*\Theta_{\mathbb{Z}} = 0$ and define $\mathbb{Y} = \{y_i\}_{i=1}^k$ as $y_i = S_{\mathbb{X}}^{-1}x_i + z_i$ for $i = 1, \ldots, k$. Notice then $\Theta_{\mathbb{Y}}^* = S_{\mathbb{X}}^{-1}\Theta_{\mathbb{X}}^* + \Theta_{\mathbb{Z}}^*$. Therefore

$$\begin{aligned}
\Theta_{\mathbb{X}}^*\Theta_{\mathbb{Y}} &= \Theta_{\mathbb{X}}^*(S_{\mathbb{X}}^{-1}\Theta_{\mathbb{X}}^* + \Theta_{\mathbb{Z}}^*)^* \\
&= \Theta_{\mathbb{X}}^*(\Theta_{\mathbb{X}}S_{\mathbb{X}}^{-1} + \Theta_{\mathbb{Z}}) \\
&= \Theta_{\mathbb{X}}^*\Theta_{\mathbb{X}}S_{\mathbb{X}}^{-1} + \Theta_{\mathbb{X}}^*\Theta_{\mathbb{Z}} \\
&= S_{\mathbb{X}}S_{\mathbb{X}}^{-1} + \Theta_{\mathbb{X}}^*\Theta_{\mathbb{Z}} \\
&= I
\end{aligned}$$

Thus $\mathbb{Y}$ is a dual frame of $\mathbb{X}$.

□

## 3.4 A Problem in Frame Theory

Given a countable index set $\mathcal{I}$, a sequence $\{x_i\}_{i\in\mathcal{I}}$ in Hilbert space $\mathcal{H}$ is called a *Riesz basic sequence* for $\mathcal{H}$ if there exists constants $0 < A \leq B < \infty$ such that

$$A\left(\sum_{i\in\mathcal{J}}|a_i|^2\right)^{1/2} \leq \left\|\sum_{i\in\mathcal{J}}a_ix_i\right\| \leq B\left(\sum_{i\in\mathcal{J}}|a_i|^2\right)^{1/2} \tag{3.2}$$

for every finite scalar sequence $\{a_i\}_{i\in\mathcal{J}}$ with $\mathcal{J} \subseteq \mathcal{I}$. If the sequence also satisfies $\overline{\text{span}}\,\{x_i\}_{i\in\mathcal{I}} = \mathcal{H}$, then the sequence is called a *Riesz basis*.

If a sequence $\{x_i\}_{i\in\mathcal{I}}$ is a Riesz basis then there exist optimal constants $A$ and $B$ that satisfy inequality (3.2). That is, if $A' > A$ or $B' < B$ then $A'$ and $B'$ cannot serve as the constants in

inequality (3.2). Such optimal constants $A$ and $B$ are called the *lower and upper basis bounds* of $\{x_i\}_{i\in\mathcal{I}}$ respectively. Next, the ratio $B/A$ is called the *condition number* of the sequence $\{x_i\}_{i\in\mathcal{I}}$.

Notice since $A \leq B$ the condition number of a Riesz basis is never less than one. In the case when the condition number of a Riesz basis is one, the Riesz basis is, in fact, an orthonormal basis. As such, Riesz bases with small condition numbers are often favorable to Riesz bases with large condition numbers since the former are more akin to orthonormal basis.

Also note that in a finite-dimensional Hilbert space, a sequence is a Riesz basis if and only if it is a basis.

Now in 1959, after studying the work of Dirac, Kadison and Singer introduced a question now known as the Kadison-Singer problem, which has important connections to quantum mechanics. Since that time, no one has been able to prove or disprove the Kadison-Singer Problem. However, progress has been made showing that the Kadison-Singer Problem is equivalent to a wide collection of other conjectures. One such conjecture is the Feichtinger Conjecture. Note that in the statement of this conjecture below, a frame is called a unit norm frame if all the elements of the frame are of unit length.

**Feichtinger Conjecture.** *Every unit norm frame for an infinite-dimensional Hilbert space can be written as a finite union of Riesz basic sequences.*

Thus the Feichtinger Conjecture is a question dealing with the decomposition of unit norm frames. Now consider the following problem.

**Problem 3.4.1.** *Given a frame $\mathbb{X} = \{x_i\}_{i=1}^k$ for $\mathbb{R}^n$, find the subset of $\mathbb{X}$ that is a Riesz basis with the smallest condition number.*

Now the Feichtinger Conjecture has a finite-dimensional counterpart that is of the same flavor as the Feichtinger Conjecture detailed above. The "hope" is that the proposed optimization problem discussed in the introduction of this work will solve the above problem in general. Then, given a frame $\mathbb{X}$, by using the optimization problem, the subset of $\mathbb{X}$ that is the Riesz basis with the smallest condition number can be identified and removed from $\mathbb{X}$. This

process could then be repeated to identify and remove the Riesz basis with the second smallest condition number and continue until the only elements that haven't been removed from $\mathbb{X}$ do not even form a basis. If one could characterize these elements, one could possibly gain insight the Feichtinger Conjecture.

Thus the Feichtinger Conjecture and Problem (3.4.1) are the main inspiration for the work in this document. For more information about Riesz bases see [11] and for more information about the Kadison-Singer Problem or the Feichtinger Conjecture see [19], [7], [8], and [9].

# CHAPTER 4. Convex Analysis

Convex optimization theory is a very diverse field with a wide collection of tools and techniques that can be viewed from many perspectives. This chapter covers only those techniques and results that will be needed in the remainder of this thesis and summarizes the theory described in [25], [3], [21], and [22].

## 4.1 Convex Sets

A subset $\mathcal{C}$ of $\mathbb{R}^n$ is a *convex set* if $(1-\lambda)x+\lambda y \in \mathcal{C}$ for all $x, y \in \mathcal{C}$ and $0 \leq \lambda \leq 1$. That is, a set is convex if it contains the line segment between any two points in the set. The following shows that convexity is preserved under arbitrary intersections.

**Proposition 4.1.1.** *Let $\mathcal{I}$ be an arbitrary index set and suppose $\mathcal{C}_i \subseteq \mathbb{R}^n$ is a convex set for all $i \in \mathcal{I}$. Then $\bigcap_{i\in\mathcal{I}} \mathcal{C}_i$ is convex.*

**Proof.**

Set $\mathcal{C} = \bigcap_{i\in\mathcal{I}} \mathcal{C}_i$ and let $x, y \in \mathcal{C}$ and $0 \leq \lambda \leq 1$. Next let $i \in \mathcal{I}$. Notice then $x, y \in \mathcal{C}_i$ and since $\mathcal{C}_i$ is convex, $(1-\lambda)x + \lambda y \in \mathcal{C}_i$. Therefore because $i \in \mathcal{I}$ was arbitrary, $(1-\lambda)x + \lambda y \in \mathcal{C}$. Hence $\mathcal{C}$ is convex.

□

Next, given a subset $\mathcal{S}$ of $\mathbb{R}^n$ the *convex hull* of $\mathcal{S}$, denoted $\operatorname{conv}\mathcal{S}$, is defined as the smallest convex set containing $\mathcal{S}$. That is

$$\operatorname{conv}\mathcal{S} := \bigcap \{X \supseteq \mathcal{S} \,:\, X \text{ convex}\}$$

Notice $\mathbb{R}^n$ is convex and $\mathcal{S} \subseteq \mathbb{R}^n$. Therefore $\operatorname{conv} \mathcal{S}$ is necessarily nonempty. Next, by the above proposition, $\operatorname{conv} \mathcal{S}$ is convex. The following proposition provides an alternate way to characterize the convex hull of a set.

**Proposition 4.1.2.** *Let $\mathcal{S} \subseteq \mathbb{R}^n$. Then*

$$\operatorname{conv} \mathcal{S} = \left\{ \sum_{i=1}^{m} \lambda_i x_i \,:\, x_1, \ldots, x_m \in \mathcal{S},\ 0 \leq \lambda_1, \ldots, \lambda_m \leq 1,\ \sum_{i=1}^{m} \lambda_i = 1,\ \textit{and } m = 1, 2, \ldots \right\}$$

**Proof.**

Set

$$\mathcal{C} = \left\{ \sum_{i=1}^{m} \lambda_i x_i \,:\, x_1, \ldots, x_m \in \mathcal{S},\ 0 \leq \lambda_1, \ldots, \lambda_m \leq 1,\ \sum_{i=1}^{m} \lambda_i = 1,\ \text{and } m = 1, 2, \ldots \right\}$$

Now let $m_1$ and $m_2$ be positive integers $x_1, \ldots, x_{m_1}, y_1, \ldots, y_{m_2} \in \mathcal{S}$, and $0 \leq \lambda_1, \ldots, \lambda_{m_1} \leq 1$ and $0 \leq \mu_1, \ldots, \mu_{m_2} \leq 1$ such that $\sum_{i=1}^{m_1} \lambda_i = 1$ and $\sum_{i=1}^{m_2} \mu_i = 1$. Further let $0 \leq \kappa \leq 1$. Then

$$z = (1-\kappa) \sum_{i=1}^{m_1} \lambda_i x_i + \kappa \sum_{i=1}^{m_2} \mu_i y_i = \sum_{i=1}^{m_1} (1-\kappa) \lambda_i x_i + \sum_{i=1}^{m_2} \kappa \mu_i y_i$$

where

$$\sum_{i=1}^{m_1} (1-\kappa) \lambda_i + \sum_{i=1}^{m_2} \kappa \mu_i = (1-\kappa) \sum_{i=1}^{m_1} \lambda_i + \kappa \sum_{i=1}^{m_2} \mu_i = (1-\kappa) + \kappa = 1$$

Thus $z \in \mathcal{C}$ and hence $\mathcal{C}$ is convex. Therefore, $\operatorname{conv} \mathcal{S} \subseteq \mathcal{C}$ by the definition of $\operatorname{conv} \mathcal{S}$.

Before establishing the converse notice if $X$ is a convex set notice if $x_1, \ldots, x_n \in X$ and $0 \leq \lambda_1, \ldots, \lambda_n \leq 1$ for some $n$ such that

$$\sum_{i=1}^{n} \lambda_i = 1$$

then

$$\sum_{i=1}^{n} \lambda_i x_i \in X$$

This follows from induction. The case for $n = 2$ is the definition of convexity. Next, if the result holds for $n$, consider the $n+1$ case and let $x_1, \ldots, x_{n+1} \in X$ and $0 \leq \lambda_1, \ldots, \lambda_{n+1} \leq 1$ such that

$$\sum_{i=1}^{n+1} \lambda_i = 1$$

Then,

$$\begin{aligned}\sum_{i=1}^{n+1}\lambda_i x_i &= \sum_{i=1}^{n}\lambda_i x_i + \lambda_{n+1}x_{n+1} \\ &= \left(\sum_{i=1}^{n}\lambda_i\right)\sum_{i=1}^{n}\frac{\lambda_i}{\sum_{i=1}^{n}\lambda_i}x_i + \lambda_{n+1}x_{n+1} \in X\end{aligned}$$

since first,

$$\sum_{i=1}^{n}\frac{\lambda_i}{\sum_{i=1}^{n}\lambda_i}x_i \in X$$

by the induction hypothesis since

$$\sum_{i=1}^{n}\frac{\lambda_i}{\sum_{i=1}^{n}\lambda_i} = 1$$

and $x_1, \ldots, x_n \in X$, and second

$$\sum_{i=1}^{n}\lambda_i + \lambda_{n+1} = 1$$

Now to establish the converse, let $x \in \mathcal{C}$. Then

$$x = \sum_{i=1}^{n}\lambda_i x_i$$

for some $x_1, \ldots, x_n \in \mathcal{S}$ and $0 \leq \lambda_1, \ldots, \lambda_n \leq 1$ for some $n$ such that

$$\sum_{i=1}^{n}\lambda_i = 1$$

Next suppose $X \supseteq \mathcal{S}$ such that $X$ is convex. Then $x_1, \ldots, x_n \in X$ and therefore $x \in X$ by the convexity of $X$. Thus as $X$ was an arbitrary convex superset of $\mathcal{S}$, it follows that $x \in \operatorname{conv}\mathcal{S}$ establishing equality.

□

An element $z$ of a convex set $\mathcal{C}$ is an *extreme point* of $\mathcal{C}$ if being written as $z = (1-\lambda)x + \lambda y$ for some $x, y \in \mathcal{C}$ such that $0 < \lambda < 1$ then necessarily either $z = x$ or $z = y$. The extreme points of a compact convex subset of $\mathbb{R}^n$ have the nice property that any element of the set can be represented in terms of the extreme points of the set as specified in the important Krein-

Milman theorem. The following statement of the Krein-Milman theorem is based on the one given in [27].

**Krein-Milman Theorem.** *The set of extreme points of a nonempty, compact, convex subset $\mathcal{C}$ of $\mathbb{R}^n$ is nonempty, and $\mathcal{C}$ is the closed convex hull of those extreme points.*

The next proposition shows, by the Krein-Milman Theorem, that if a compact, convex subset of $\mathbb{R}^n$ has a finite number of extreme points, then it is simply the convex hull of those extreme points.

**Proposition 4.1.3.** *Let $\mathcal{P}$ be a finite subset of $\mathbb{R}^n$. Then $\overline{\text{conv}}(\mathcal{P}) = \text{conv}(\mathcal{P})$.*

**Proof.**

The result will be established by showing $\text{conv}(\mathcal{P})$ is closed. To do so, suppose $x = \lim_n x_n$ for some $\{x_n\}_n \subseteq \text{conv}(\mathcal{P})$. Next write $\mathcal{P} = \{p_1, \ldots, p_N\}$. Then as $x_n \in \text{conv}(\mathcal{P})$ for each $n$ there exists $\{\lambda_i^n\}_{i=1}^N \subseteq [0,1]$ such that

$$x_n = \sum_{i=1}^N \lambda_i^n p_i \qquad \text{where} \qquad \sum_{i=1}^N \lambda_i^n = 1$$

Next since the sequence $\{\lambda_1^n\}_{n=1}^\infty$ is bounded, the Bolzano-Weierstrass Theorem asserts it has a convergent subsequence. Let $\{\lambda_1^{n_{k_1}}\}_{k_1=1}^\infty$ denote this subsequence. Now consider the sequence $\{\lambda_2^{n_{k_1}}\}_{k_1=1}^\infty$. Again this sequence is bounded and hence has a convergent subsequence by the Bolzano-Weierstrass Theorem. Let $\{\lambda_2^{(n_{k_1})_{k_2}}\}_{k_2=1}^\infty$ denote this subsequence. This process can be repeated for each of the finitely many $i = 1, \ldots, N$ resulting, in the end with the convergent sequence, $\{\lambda_N^{(((n_{k_1})_{k_2})\cdots)_{k_N}}\}_{k_N=1}^\infty$. Now let $\ell_j = ((((n_{k_1})_{k_2})\cdots)_{k_{N-1}})_j$. Notice then, by construction, the sequence $\{\lambda_i^{\ell_j}\}_{j=1}^\infty$ is convergent for all $i = 1, \ldots, N$. Thus let $\lim_j \lambda_i^{\ell_j} = \lambda_i$ for all $i = 1, \ldots, N$. Then

$$x = \lim_n x_n = \lim_j x_{\ell_j} = \lim_j \sum_{i=1}^N \lambda_i^{\ell_j} p_i = \sum_{i=1}^N \lim_j \lambda_i^{\ell_j} p_i = \sum_{i=1}^N \left( \lim_j \lambda_i^{\ell_j} \right) p_i = \sum_{i=1}^N \lambda_i p_i$$

Next,

$$\sum_{i=1}^N \lambda_i = \sum_{i=1}^N \lim_j \lambda_i^{\ell_j} = \lim_j \sum_{i=1}^N \lambda_i^{\ell_j} = \lim_j 1 = 1$$

since $\sum_{i=1}^{N} \lambda_i^n = 1$ for all $n$ and therefore in particular for $n = \ell_j$ for each $j$. Hence $x \in \text{conv}(\mathcal{P})$ and therefore $\text{conv}(\mathcal{P})$ is closed.

□

## 4.2 Convex Functions

If $\mathcal{C} \subseteq \mathbb{R}^n$ is a convex set then a function $f : \mathcal{C} \to \mathbb{R}$ is *convex* if

$$f((1-\lambda)x + \lambda y) \leq (1-\lambda)f(x) + \lambda f(y)$$

for every $x, y \in \mathcal{C}$ and $0 \leq \lambda \leq 1$. That is, $f$ lies below the line segment connecting $(x, f(x))$ and $(y, f(y))$. Convex functions have many nice properties. The first deals with minimizing a convex function over a convex set.

If $\mathcal{C}$ is a convex set, $f : \mathcal{C} \to \mathbb{R}$ is a convex function, and $f$ achieves its minimum at some point in $\mathcal{C}$, the minimizing set of $f$ over $\mathcal{C}$ is defined as

$$\arg\min_{x \in \mathcal{C}} f(x) := \left\{ y \in \mathcal{C} \ : \ f(y) = \min_{x \in \mathcal{C}} f(x) \right\}$$

In fact, the above set is convex as shown by the next result.

**Proposition 4.2.1.** *Let $\mathcal{C}$ be a convex subset of $\mathbb{R}^n$ and $f : \mathcal{C} \to \mathbb{R}$ a convex function such that $f$ achieves its minimum at some point in $\mathcal{C}$. Then* $\arg\min_{x \in \mathcal{C}} f(x)$ *is convex.*

**Proof.**

Let

$$\mathcal{M} = \arg\min_{x \in \mathcal{C}} f(x)$$

Since $f$ achieves its minimum at some point in $\mathcal{C}$, the set $\mathcal{M}$ is well-defined and nonempty. Now let $x, y \in \mathcal{M}$ and $0 \leq \lambda \leq 1$. Then

$$f(x) = f(y) = \mu = \min_{x \in \mathcal{C}} f(x)$$

Thus, by the convexity of $f$, and the construction of $\mu$,

$$\mu \leq f((1-\lambda)x + \lambda y) \leq (1-\lambda)f(x) + \lambda f(y) = (1-\lambda)\mu + \lambda\mu = \mu$$

Here the fact that $(1-\lambda)x+\lambda y \in \mathcal{C}$ from the convexity of $\mathcal{C}$ is used. Thus $f((1-\lambda)x+\lambda y) = \mu$ and therefore $(1-\lambda)x + \lambda y \in \mathcal{M}$ establishing the convexity of $\mathcal{M}$.

□

Not only is the minimizing set of a convex function convex, but convex functions also have the extremely useful property that local minima are necessarily global minima. That is, $x_0 \in \mathbb{R}^n$ is said to be a *local minimizer* of $f : \mathbb{R}^n \to \mathbb{R}$ if there exists an open set $\mathcal{O}$ containing $x_0$ such that $f(x) \geq f(x_0)$ for all $x \in \mathcal{O}$. In this case, $f(x_0)$ is said to be a *local minima* of $f$. Otherwise if $x_1 \in \mathbb{R}^n$ such that $f(x) \geq f(x_1)$ for all $x \in \mathbb{R}^n$ then $x_1$ is said to be a *global minimizer* of $f$ and $f(x_1)$ is said to be a *global minima*.

**Proposition 4.2.2.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function. Then $x_0$ is a local minimizer of $f$ if and only if it is a global minimizer of $f$.*

**Proof.**

If a point is a global minimizer then it is clearly a local minimizer. Conversely, suppose $x_0$ is a local minimizer of $f$. Then there exists an open neighborhood $\mathcal{O}$ of $x_0$ such that $f(x) \geq f(x_0)$ for all $x \in \mathcal{O}$. Now, to reach a contradiction, suppose there exists $x_1 \in \mathbb{R}^n$ such that $f(x_1) < f(x_0)$. Next let $0 \leq \lambda \leq 1$ and set $y_\lambda = (1-\lambda)x_0 + \lambda x_1$. If $\lambda > 0$ then by the convexity of $f$,

$$\begin{aligned} f(y_\lambda) &= f((1-\lambda)x_0 + \lambda x_1) \\ &= (1-\lambda)f(x_0) + \lambda f(x_1) \\ &< (1-\lambda)f(x_0) + \lambda f(x_0) \\ &= f(x_0) \end{aligned}$$

The assertion that $\lambda > 0$ is needed since $f(x_1) < f(x_0)$ implies $\lambda f(x_1) < \lambda f(x_0)$ provided $\lambda > 0$. Last since $\mathcal{O}$ is an open neighborhood of $x_0$ there exists $\lambda > 0$ such that $y_\lambda \in \mathcal{O}$. However, by the work above, then $f(y_\lambda) < f(x_0)$ contradicting the assumption that $x_0$ is a local minimizer of $f$. Therefore $x_0$ is a global minimizer of $f$.

□

Convex functions also share a property, known as the *Maximum Principle*, which dictates where the function may achieve its maximum value over a convex set, if it, in fact, achieves its maximum on the set. The Maximum Principle can be expressed in many different forms. The following form will be the only form needed in the rest of this work.

**Theorem 4.2.3. (Maximum Principle)** *Let $\mathcal{C}$ be a nonempty, compact, convex subset of $\mathbb{R}^n$ with a finite number of extreme points $\{e_i\}_{i=1}^N$. If $f : \mathcal{C} \to \mathbb{R}$ is convex then*

$$\max_{x\in\mathcal{C}} f(x) = \max_{i=1,\dots,N} f(e_i)$$

**Proof.**

By the Krein-Milman Theorem and Proposition (4.1.3),

$$\mathcal{C} = \overline{\text{conv}}\,\{e_i \,:\, i = 1,\dots,N\} = \text{conv}\,\{e_i \,:\, i = 1,\dots,N\}$$

Therefore for any $x \in \mathcal{C}$ there exists $\{\lambda_i\}_{i=1}^N \subseteq [0,1]$ such that $\sum_{i=1}^N \lambda_i = 1$ and $x = \sum_{i=1}^N \lambda_i e_i$. Now let

$$\mu = \max_{i=1,\dots,N} f(e_i)$$

The maximum exists and is finite since there are only finitely many extreme points $e_i$. Next by the convexity of $f$,

$$f(x) = f\left(\sum_{i=1}^N \lambda_i e_i\right) \le \sum_{i=1}^N \lambda_i f(e_i) \le \sum_{i=1}^N \lambda_i \mu = \mu \sum_{i=1}^N \lambda_i = \mu$$

Therefore for every $x \in \mathcal{C}$ it follows that $f(x) \le \mu$. Hence

$$\sup_{x\in\mathcal{C}} f(x) \le \mu$$

Next since $\{e_i \,:\, i = 1,\dots,N\} \subseteq \mathcal{C}$

$$\mu \le \sup_{x\in\mathcal{C}} f(x)$$

Last since $e_i \in \mathcal{C}$ for $i = 1,\dots,N$, the supremum above is attained and

$$\max_{x\in\mathcal{C}} f(x) = \mu$$

□

## 4.3 Minimizing Convex Functions

Now let $\mathcal{A} \subseteq \mathbb{R}^n$ and $f : \mathcal{A} \to \mathbb{R}$ and consider the optimization problem

$$\text{minimize } f(x) \quad \text{subject to} \quad x \in \mathcal{A}$$

If $\mathcal{A} = \{x \in \mathbb{R}^n : g(x) = 0\}$ for some differentiable function $g : \mathbb{R}^n \to \mathbb{R}^n$ and $f$ is differentiable on $\mathcal{A}$, the method of Lagrange multiplies can be used to solve such an optimization problem. However, if the set $\mathcal{A}$ is the solution set of a system of inequalities and equalities, more work may be needed to solve the problem. The Karush-Kuhn-Tucker conditions are an extension of the method of Lagrange multipliers and provide conditions for a point to be a local minimizer of a function. These conditions have a general form that can analyze a very wide variety of optimization problems. However, since optimization problems can often be manipulated to simplify their form, this work will only need to consider minimizing a convex function over all of $\mathbb{R}^n$. A more thorough description of the method of Lagrange multipliers and the Karush-Kuhn-Tucker conditions can be found in [2], [3], and [23].

Since the main function that will be the focus of minimization in this document is not differentiable on $\mathbb{R}^n$, the next result cannot be directly applied. However, it serves as a starting point for the techniques that can be used. Recall $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable on $\mathbb{R}^n$ if each of its partial derivatives with respect to each of its variables is continuous on $\mathbb{R}^n$. In this case, the *gradient* of $f$ at $x$ is defined as

$$\nabla f(x) := \begin{bmatrix} \frac{\partial f}{\partial x_1}(x) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x) \end{bmatrix}$$

In the case that $f$ is differentiable on $\mathbb{R}^n$, the following result provides another characterization of convexity in terms of the gradient of $f$. The proof below is based on the one given in [3].

**Proposition 4.3.1.** *A differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if and only if*

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$$

*for all $x, y \in \mathbb{R}^n$.*

**Proof.**

First consider the case when $n = 1$. For the forward direction, suppose $f$ is convex on $\mathbb{R}^n$ and let $x, y \in \mathbb{R}$. If $x = y$ then trivially $f(y) \geq f(x) + f'(x)(y - x)$. Thus suppose $x \neq y$ and let $0 < \lambda \leq 1$. Then, by the convexity of $f$,

$$\begin{aligned} f(x + \lambda(y - x)) &= f((1 - \lambda)x + \lambda y) \\ &\leq (1 - \lambda)f(x) + \lambda f(y) \\ &= f(x) + \lambda(f(y) - f(x)) \end{aligned}$$

Hence since $\lambda > 0$,

$$\frac{f(x + \lambda(y - x)) - f(x)}{\lambda} \leq f(y) - f(x)$$

and therefore

$$f(y) \geq f(x) + \frac{f(x + \lambda(y - x)) - f(x)}{\lambda} \tag{4.1}$$

Next notice since $f$ is differentiable at $x$,

$$\begin{aligned} \lim_{\lambda \to 0^+} \frac{f(x + \lambda(y - x)) - f(x)}{\lambda} &= (y - x) \lim_{\lambda \to 0^+} \frac{f(x + \lambda(y - x)) - f(x)}{\lambda(y - x)} \\ &= (y - x) \lim_{\lambda \to 0} \frac{f(x + \lambda(y - x)) - f(x)}{\lambda(y - x)} \\ &= (y - x) \lim_{t \to 0} \frac{f(x + t) - f(x)}{t} \\ &= (y - x) f'(x) \end{aligned}$$

Thus since inequality (4.1) holds for all $0 < \lambda \leq 1$, letting $\lambda \to 0$ with $\lambda > 0$ yields

$$\begin{aligned} f(y) &\geq f(x) + \lim_{\lambda \to 0^+} \frac{f(x + \lambda(y - x)) - f(x)}{\lambda} \\ &= f(x) + f'(x)(y - x) \end{aligned}$$

For the reverse direction, suppose

$$f(y) \geq f(x) + f'(x)(y - x) \tag{4.2}$$

for all $x, y \in \mathbb{R}$. Now let $0 \leq \lambda \leq 1$ and set $z = \lambda x + (1 - \lambda)y$. Then applying (4.2) to $x$ yields

$$f(x) \geq f(z) + f'(z)(x - z)$$

and to $y$ yields

$$f(y) \geq f(z) + f'(z)(y - z).$$

Hence

$$\begin{aligned}\lambda f(x) + (1-\lambda)f(y) &\geq f(z) + \lambda f'(z)(x - z) + (1-\lambda)f'(z)(y - z) \\ &= f(z) + f'(z)(\lambda x + (1-\lambda)y - z) \\ &= f(\lambda x + (1-\lambda)y)\end{aligned}$$

establishes that $f$ is convex.

Now to consider the general case, define, for fixed $x, y \in \mathbb{R}^n$, the function $g_{x,y} : \mathbb{R} \to \mathbb{R}$ as $g_{x,y}(\lambda) = f(\lambda x + (1-\lambda)y)$. Then for fixed $\lambda \in \mathbb{R}$ let $z = \lambda x + (1-\lambda)y$ and let $\Delta\lambda \in \mathbb{R}$. Then the Taylor expansion of $f$ around $z$ is

$$f(z + (x - y)\Delta\lambda) = f(z) + \langle \nabla f(z), (x - y)\Delta\lambda \rangle + \varepsilon(z, (x - y)\Delta\lambda)$$

where $\varepsilon$ is a function such that

$$\lim_{\Delta\lambda \to 0} \frac{\varepsilon(z, (x - y)\Delta\lambda)}{\Delta\lambda} = 0.$$

Thus

$$\begin{aligned}g'_{x,y}(\lambda) &= \lim_{\Delta\lambda \to 0} \frac{f((\lambda + \Delta\lambda)x + (1 - \lambda - \Delta\lambda)y) - f(\lambda x + (1-\lambda)y)}{\Delta\lambda} \\ &= \lim_{\Delta\lambda \to 0} \frac{f(z + (x - y)\Delta\lambda) - f(z)}{\Delta\lambda} \\ &= \lim_{\Delta\lambda \to 0} \frac{\langle \nabla f(z), (x - y)\Delta\lambda \rangle + \varepsilon(z, (y - x)\Delta\lambda)}{\Delta\lambda} \\ &= \langle \nabla f(z), x - y \rangle \\ &= \langle \nabla f(\lambda x + (1-\lambda)y), x - y \rangle\end{aligned}$$

Now suppose $f$ is convex. Then for any $x, y \in \mathbb{R}$ the function $g_{x,y}$ is convex and therefore, as established in the one-dimensional case, $g(0) \geq g(1) + g'(1)(0 - 1) = g(1) - g'(1)$. Thus

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$$

Conversely suppose the above inequality holds for all $x, y \in \mathbb{R}^n$. Fix $x, y \in \mathbb{R}^n$ and let $\lambda_1, \lambda_2 \in \mathbb{R}$ and set $z_i = \lambda_i x + (1 - \lambda_i)y$ for $i = 1, 2$. Notice then

$$\begin{aligned} z_2 - z_1 &= \lambda_2 x + (1 - \lambda_2)y - (\lambda_1 x + (1 - \lambda_1)y) \\ &= \lambda_2 x - \lambda_2 y - \lambda_1 x + \lambda_1 y \\ &= (\lambda_2 - \lambda_1)(x - y) \end{aligned}$$

Therefore by inequality (4.3),

$$\begin{aligned} g_{x,y}(\lambda_2) &= f(z_2) \\ &\geq f(z_1) + \langle \nabla f(z_1), z_2 - z_1 \rangle \\ &\geq f(z_1) + \langle \nabla f(z_1), (\lambda_2 - \lambda_1)(x - y) \rangle \\ &= f(z_1) + (\lambda_2 - \lambda_1) \langle \nabla f(z_1), x - y \rangle \\ &= g_{x,y}(\lambda_1) + (\lambda_2 - \lambda_1) g'_{x,y}(\lambda_1) \end{aligned}$$

Thus, by the results above for the one-dimensional case, since $\lambda_1$ and $\lambda_2$ were arbitrary $g_{x,y}$ is convex and since $x$ and $y$ were arbitrary, $f$ is convex.

□

The result above shows if $x \in \mathbb{R}^n$ such that $\nabla f(x) = 0$ then necessarily $x$ is a global minimizer of $f$. This is formally stated next. Note that for a general differentiable function $g$, the condition that $\nabla g(y) = 0$ for some $y$ is not enough to ensure $y$ is a global minimizer of $y$.

**Proposition 4.3.2.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex differentiable function on $\mathbb{R}^n$ and suppose $x_0 \in \mathbb{R}^n$ is such that $\nabla f(x_0) = 0$. Then $x_0$ is a global minimizer of $f$.*

**Proof.**

By the above proposition, for any $x \in \mathbb{R}^n$

$$f(x) \geq f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle = f(x_0)$$

and therefore $x_0$ is a global minimizer of $f$.

□

Thus if the convex function $f$ is differentiable on $\mathbb{R}^n$, global minimizers of $f$ can be identified by solving $\nabla f(x) = 0$. If $f$ is convex but not differentiable, this technique cannot be applied. However, the technique can be generalized to develop other techniques to determine if a point is a global minimizer of $f$. Note that without loss of generality, focus can be given to show that zero is a global minimizer of $f$.

The first technique considers, for a fixed $x \in \mathbb{R}^n$, the function in a single variable $g_x : [0,1] \to \mathbb{R}$ defined as $g_x(t) := f(tx)$. For a particular $x \in \mathbb{R}^n$, the function $g_x$ follows $f$ along the vector $x$ radiating away from the origin. The idea is if $g_x$ is nondecreasing for all $x \in \mathbb{R}^n$ then as one moves away from the origin, along the vector $x$, the value of $f$ cannot decrease. Thus the origin is a global minimizer of $f$. The next two propositions show that this is in fact the case. Notice since local minimizers of a convex function are necessarily global minimizers, one only needs to consider $x$ in a neighborhood of the origin.

**Proposition 4.3.3.** *Let $\|\cdot\|$ be any norm on $\mathbb{R}^n$. Next let $f : \mathbb{R}^n \to \mathbb{R}$ be continuous on $\overline{B_r(0)}$ for some $r > 0$ and define, for fixed $x \in \overline{B_r(0)} \setminus \{0\}$, $g_x : [0,1] \to \mathbb{R}$ by*

$$g_x(t) := f(tx)$$

*Then for any $x \in \overline{B_r(0)}$, $g_x$ is continuous on $[0,1]$.*

**Proof.**

Let $t_0 \in [0,1]$ and notice $t_0 x \in \overline{B_r(0)}$ since

$$\|t_0 x\| = t_0 \|x\| \leq \|x\| \leq R$$

Since $f$ is continuous at $t_0 x$, given $\varepsilon > 0$ there exists $\delta > 0$ such that if $y \in \overline{B_r(0)}$ and $\|t_0 x - y\| < \delta$ then $\|f(t_0 x) - f(y)\| < \varepsilon$. Thus

$$\begin{aligned} |t_0 - t| < \frac{\delta}{|x|} \quad &\implies \quad |t_0 - t||x| < \delta \\ &\implies \quad |t_0 x - tx| < \delta \\ &\implies \quad |f(t_0 x) - f(tx)| < \varepsilon \\ &\implies \quad |g_x(t_0) - g_x(t)| < \varepsilon \end{aligned}$$

Thus $g_x$ is continuous at $t_0$.

□

**Proposition 4.3.4.** *Let $\|\cdot\|$ be any norm on $\mathbb{R}^n$. Next let $f : \mathbb{R}^n \to \mathbb{R}$ be continuous on $\overline{B_r(0)}$ for some $r > 0$ and define, for fixed $x \in \overline{B_r(0)} \setminus \{0\}$, $g_x : [0,1] \to \mathbb{R}$ by*

$$g_x(t) := f(tx)$$

*Suppose for every $x \in B_r(0)$ the function $g_x$ is differentiable on $(0,1)$ and $g_x'(t) \geq 0$ for $0 < t < 1$. Then*

$$\min_{x \in B_r(0)} f(x) = f(0)$$

*Furthermore if $f$ is convex on $\overline{B_r(0)}$ and $g_x'(t) \geq 0$ for all $x \in B_r(0)$ and $0 < t < 1$, then*

$$\arg\min_{x \in B_r(0)} f(x) = \{0\}$$

*if and only if $g_x'(t) > 0$ for all $x \in B_r(0)$ and $0 < t < 1$.*

**Proof.**

The proposition claims $f(0) \leq f(x)$ for any $x \in B_r(0)$. For sake of contradiction, suppose there exists $x_0 \in B_r(0)$ such that $f(x_0) < f(0)$.

Since, by the above lemma, $g_{x_0}$ is continuous on $[0,1]$ there exists, by the Mean Value Theorem, $t_0 \in (0,1)$ such that

$$g_{x_0}'(t_0) = \frac{g_{x_0}(1) - g_{x_0}(0)}{1 - 0} = f(x_0) - f(0) < 0$$

However, this contradicts the fact that $g_x'(t) \geq 0$ for any $x \in B_r(0)$ and any $t \in (0,1)$. Thus $f(0) \leq f(x)$ for all $x \in B_r(0)$, proving the first claim.

To prove the second claim notice if $f$ is convex on $\overline{B_r(0)}$ then I claim $g_x$ is convex on $[0,1]$ for any $x \in \overline{B_r(0)}$. To see why let $x \in \overline{B_r(0)}$, $s, t \in [0,1]$, and $\lambda \in [0,1]$. Then

$$\begin{aligned} g_x(\lambda s + (1-\lambda)t) &= f((\lambda s + (1-\lambda)t)x) \\ &= f(\lambda sx + (1-\lambda)tx) \\ &\leq \lambda f(sx) + (1-\lambda) f(tx) \\ &= \lambda g_x(s) + (1-\lambda) g_x(t) \end{aligned}$$

Hence $g_x$ is convex on $[0,1]$. Therefore since, by assumption, $g_x$ is differentiable on $(0,1)$ it follows, by Proposition (4.3.1), that for any $t \in (0,1)$ and $s \in [0,1]$,

$$g_x(s) \geq g_x(t) + g_x'(t)(s-t)$$

Next since, by assumption, $g_x'(t) \geq 0$ for all $x \in B_r(0)$ and $0 < t < 1$ it has been established from the first claim that zero is a minimizer of $f$ on $B_r(0)$. Now to complete the proof of the second claim, the contrapositive of each direction will be established.

To show the contrapositive of the forward direction suppose there exists $x_0 \in B_r(0)$ with $x_0 \neq 0$ and $t_0 \in (0,1)$ such that $g_{x_0}'(t_0) \leq 0$. Then since, by assumption, $g_x'(t) \geq 0$ for all $x \in B_r(0)$ and $0 < t < 1$, it follows that $g_{x_0}'(t_0) \geq 0$ and hence $g_{x_0}'(t_0) = 0$. Therefore by the above inequality

$$f(0) = g_{x_0}(0) \geq g_{x_0}(t_0) + g_{x_0}'(t_0)(0 - t_0) = f(t_0 x_0)$$

Hence $f(t_0 x_0) \leq f(0)$ and since zero is a minimizer of $f$ on $B_r(0)$ it follows that $f(0) \leq f(t_0 x_0)$ and therefore $f(t_0 x_0) = f(0)$. Thus $t_0 x_0$ is also a minimizer of $f$ on $B_r(0)$, establishing the contrapositive of the forward direction since $t_0 x_0 \neq 0$.

To prove the contrapositive of the reverse direction suppose there exists $x_0 \neq 0$ with $x_0 \in B_r(0)$ such that

$$x_0 \in \arg\min_{x \in B_r(0)} f(x)$$

Then $f(0) = f(x_0)$. Now consider the function $g_{x_0}$ and notice since this function is convex on $[0,1]$

$$\begin{aligned}
f(\frac{1}{2}x_0) &= g_{x_0}(1/2) \\
&= g_{x_0}(1/2 \cdot 0 + 1/2 \cdot 1) \\
&\leq \frac{1}{2} g_{x_0}(0) + \frac{1}{2} g_{x_0}(1) \\
&= \frac{1}{2} f(0) + \frac{1}{2} f(x_0) \\
&= \frac{1}{2} f(0) + \frac{1}{2} f(0) \\
&= f(0)
\end{aligned}$$

However, zero is a minimizer of $f$ on $B_r(0)$ and hence $f(0) \leq f(\frac{1}{2}x_0)$. Therefore $f(\frac{1}{2}x_0) = f(0)$ and thus by the above inequality,

$$\begin{aligned} f(0) &= f(x_0) \\ &= g_{x_0}(1) \\ &\geq g_{x_0}(1/2) + g'_{x_0}(1/2)(1 - 1/2) \\ &= f(\frac{1}{2}x_0) + \frac{1}{2}g'_{x_0}(1/2) \\ &= f(0) + \frac{1}{2}g'_{x_0}(1/2) \end{aligned}$$

Hence $g'_{x_0}(1/2) \leq 0$, but $g'_{x_0}(1/2) \geq 0$ since, by assumption, $g_x(t) \geq 0$ for all $x \in B_r(0)$ and $t \in (0, 1)$. Thus $g'_{x_0}(1/2) = 0$ establishing the contrapositive of the reverse direction.

□

As discussed in [25], a second way to identify global minimizers of a nondifferentiable convex function is to generalize the concept of the derivative. For a convex function $f : \mathbb{R}^n \to \mathbb{R}$, a vector $\gamma \in \mathbb{R}^n$ is said to be a *subgradient* of $f$ at $x_0$ if $f(x) \geq f(x_0) + \langle \gamma, x - x_0 \rangle$ for all $x \in \mathbb{R}^n$. The *subdifferential* of $f$ at $x_0$ is the set of all subgradients of $f$ at $x_0$ and is denoted $\partial f(x_0)$.

Notice $g_\gamma(x) := f(x_0) + \langle \gamma, x - x_0 \rangle$ is a linear function touching $f$ at $(x_0, f(x_0))$. Then $\gamma$ is a subgradient of $f$ at $x_0$ if $f(x) \geq g_\gamma(x)$ for all $x \in \mathbb{R}^n$. That is, $g_\gamma$ is a global underestimator of $f$.

Also notice if $f$ is convex and is differentiable at $x_0$ then $g_{\nabla f(x_0)}(x)$ is the only global linear underestimator of $f$ that passes through $(x_0, f(x_0))$. Otherwise if $f$ is not differentiable at $x_0$ there are possibly more than one linear underestimator of $f$ at $x_0$ that passes through $(x_0, f(x_0))$. This is illustrated in Figure (4.3).

In fact since $f$ is a convex function, a linear function $g_\gamma$ is a global underestimator of $f$ that passes through $(x_0, f(x_0))$ if and only if $g_\gamma$ is a local underestimator of $f$ that passes through $(x_0, f(x_0))$ as shown in the next result.

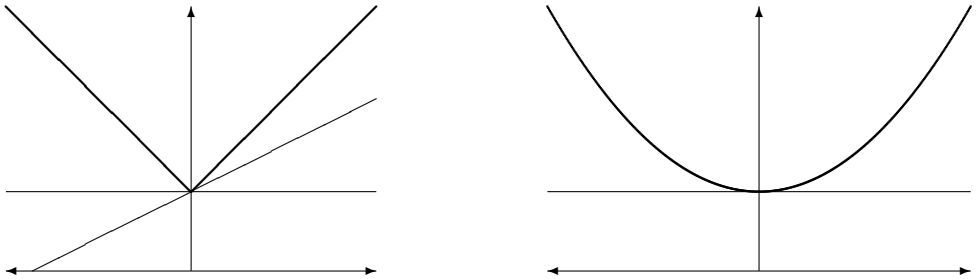

Figure 4.1 The function on the left is nondifferentiable at zero and thus there are more than one linear function touching the function at zero (two are shown) that underestimates the function. The function on the right is differentiable at zero and there is only one linear function touching the function at zero that underestimates the function.

**Proposition 4.3.5.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function on $\mathbb{R}^n$ and let $\mathcal{O}$ be an open neighborhood of $x_0$. Then*

$$f(x) \geq f(x_0) + \langle \gamma, x - x_0 \rangle \qquad \textit{for all } x \in \mathbb{R}^n$$

*if and only if*

$$f(x) \geq f(x_0) + \langle \gamma, x - x_0 \rangle \qquad \textit{for all } x \in \mathcal{O}$$

**Proof.**

The forward direction is obvious. To prove the reverse direction suppose

$$f(x) \geq f(x_0) + \langle \gamma, x - x_0 \rangle \qquad \text{for all } x \in \mathcal{O}$$

but assume, for the sake of reaching a contradiction, there exists $y \in \mathbb{R}^n$ such that

$$f(y) < f(x_0) + \langle \gamma, y - x_0 \rangle$$

Then since $\mathcal{O}$ is open there exists $\lambda \in (0, 1)$ such that

$$z := x_0 + \lambda(y - x_0) = \lambda y + (1 - \lambda)x_0 \in \mathcal{O}$$

Specifically since $\mathcal{O}$ is open there exists $r > 0$ such that

$$\{x \in \mathbb{R}^n \,:\, \|x - x_0\| < r\} \subseteq \mathcal{O}$$

where $\|\cdot\|$ is a norm on $\mathbb{R}^n$. Then

$$\begin{aligned}\|z - x_0\| &= \|x_0 + \lambda(y - x_0) - x_0\| \\ &= \|\lambda(y - x_0)\| \\ &= \lambda \|y - x_0\| \\ &< r\end{aligned}$$

if $\lambda < r/\|y - x_0\|$ since $\|y - x_0\| \neq 0$ because necessarily $y \neq x_0$. Thus because $f$ is convex on $\mathbb{R}^n$ and $z \in \mathcal{O}$

$$\begin{aligned}f(x_0) + \langle \gamma, z - x_0 \rangle &\leq f(z) \\ &= f(\lambda y + (1 - \lambda)x_0) \\ &\leq \lambda f(y) + (1 - \lambda) f(x_0) \\ &< \lambda(f(x_0) + \langle \gamma, y - x_0 \rangle) + (1 - \lambda) f(x_0) \\ &= \lambda f(x_0) + \lambda \langle \gamma, y - x_0 \rangle + f(x_0) - \lambda f(x_0) \\ &= f(x_0) + \lambda \langle \gamma, y - x_0 \rangle\end{aligned}$$

That is

$$\begin{aligned}f(x_0) + \lambda \langle \gamma, y - x_0 \rangle &> f(x_0) + \langle \gamma, z - x_0 \rangle \\ &= f(x_0) + \langle \gamma, x_0 + \lambda(y - x_0) - x_0 \rangle \\ &= f(x_0) + \lambda \langle \gamma, y - x_0 \rangle\end{aligned}$$

a contradiction, completing the proof.

□

The following result is fundamental to showing that a point is a global minimizer of a convex function, and it generalizes Proposition (4.3.2).

**Proposition 4.3.6.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function. Then $x_0$ is a global minimizer of $f$ if and only if $0 \in \partial f(x_0)$.*

**Proof.**

Notice $0 \in \partial f(x_0)$ if and only if $f(x) \geq f(x_0) + \langle 0, x - x_0 \rangle$ for all $x \in \mathbb{R}^n$ if and only if $f(x) \geq f(x_0)$ for all $x \in \mathbb{R}^n$ if and only if $x_0$ is a global minimizer of $f$.

□

# CHAPTER 5. Subdifferential Analysis

Recall the function $\|\cdot\|_1 : \mathbb{R}^{m\times n} \to \mathbb{R}$ from Definition (2.2.1) is defined as

$$\|X\|_1 = \sum_{j=1}^{n}\sum_{i=1}^{m} |X_{i,j}|$$

for any $X \in \mathbb{R}^{m\times n}$. Next notice this function is convex on $\mathbb{R}^{m\times n}$ since for any $X, Y \in \mathbb{R}^{m\times n}$ and $0 \leq \lambda \leq 1$,

$$\begin{aligned}
\|(1-\lambda)X + \lambda Y\|_1 &= \sum_{j=1}^{n}\sum_{i=1}^{m} |(1-\lambda)X_{i,j} + \lambda Y_{i,j}| \\
&\leq \sum_{j=1}^{n}\sum_{i=1}^{m} ((1-\lambda)\,|X_{i,j}| + \lambda\,|Y_{i,j}|) \\
&= (1-\lambda)\sum_{j=1}^{n}\sum_{i=1}^{m} |X_{i,j}| + \lambda\sum_{j=1}^{n}\sum_{i=1}^{m} |Y_{i,j}| \\
&= (1-\lambda)\,\|X\|_1 + \lambda\,\|Y\|_1
\end{aligned}$$

The goal of this chapter is to develop the tools of subdifferential analysis that will be needed in the rest of this work. In particular, given $A \in \mathbb{R}^{n\times m}$, $B \in \mathbb{R}^{N\times M}$, and $C \in \mathbb{R}^{n\times M}$ the ultimate goal is to completely characterize the subdifferential of the function $\|AXB + C\|_1$ at the origin.

Theorem 23.8 of [25] will be paramount at developing this characterization. Before stating this theorem, the Minkowski sum of two sets $A, B \subseteq \mathbb{R}^n$ is defined as

$$A + B := \{a + b \,:\, a \in A \text{ and } b \in B\}$$

Since the subdifferential of a convex function at a point is a set, if $f_1, f_2 : \mathbb{R}^n \to \mathbb{R}$ are two convex functions on $\mathbb{R}^n$ and $x \in \mathbb{R}^n$ the sum $\partial f_1(x) + \partial f_2(x)$ refers to the Minkowski sum.

**Theorem 23.8, [25].** *Let $f_1, \ldots, f_k : \mathbb{R}^n \to \mathbb{R}$ be convex functions such that $f_i(x)$ is finite for*

*all* $x \in \mathbb{R}^n$ *and all* $i = 1, \ldots, k$. *Then*

$$\partial(f_1 + \cdots + f_k)(x) = \partial f_1(x) + \cdots + \partial f_k(x)$$

More general forms of the above theorem exist, but the above statement is the only one that will be needed here. In particular, if the convex functions in the above theorem are allowed to attain the value $\infty$ or $-\infty$ then the functions may need to satisfy more conditions to guarantee the theorem holds.

In this chapter, the sign patterns of vectors and matrices will play a major role. The sign of a real number $x \in \mathbb{R}$ is defined as

$$\operatorname{sgn}(x) := \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0 \end{cases}$$

and the sign of a matrix (or vector) $X \in \mathbb{R}^{m \times n}$ is defined as $\operatorname{Sgn}(X)_{i,j} := (\operatorname{sgn}(X_{i,j}))_{i,j}$ for $i = 1, \ldots, m$ and $j = 1, \ldots, n$.

## 5.1 Preliminary Work

First given $a \in \mathbb{R}^n$ and $\beta \in \mathbb{R}$, the following lemma will be used to calculate the subdifferential of the function $f(x) := |\langle a, x\rangle + \beta|$ at the origin.

**Lemma 5.1.1.** *Let* $a \in \mathbb{R}^n$ *and* $\beta \in \mathbb{R}$. *If* $x$ *is such that* $|\langle a, x\rangle| \leq |\beta|$ *then*

$$|\langle a, x\rangle + \beta| = \operatorname{sgn}(\beta)\, \langle a, x\rangle + |\beta|$$

**Proof.**

First if $\beta = 0$ then $|\langle a, x\rangle| \leq |\beta| = 0$ implies $\langle a, x\rangle = 0$. Therefore

$$|\, \langle a, x\rangle + \beta| = 0 = \operatorname{sgn}(\beta)\, \langle a, x\rangle + |\beta|$$

Next notice $|\langle a, x\rangle| \leq |\beta|$ implies $\langle a, x\rangle \geq -|\beta|$. Hence $\langle a, x\rangle + |\beta| \geq 0$. Therefore if $\beta > 0$ then

$$\begin{aligned} |\, \langle a, x\rangle + \beta| &= |\, \langle a, x\rangle + |\beta|| \\ &= \langle a, x\rangle + |\beta| \\ &= \operatorname{sgn}(\beta)\, \langle a, x\rangle + |\beta| \end{aligned}$$

since $\operatorname{sgn}(\beta) = 1$.

Similarly, $|\langle a, x\rangle| \leq |\beta|$ implies $-\langle a, x\rangle \geq -|\beta|$. Hence $-\langle a, x\rangle + |\beta| \geq 0$. Thus if $\beta < 0$ then

$$\begin{aligned} |\langle a, x\rangle + \beta| &= |-\langle a, x\rangle - \beta| \\ &= |-\langle a, x\rangle + |\beta|| \\ &= -\langle a, x\rangle + |\beta| \\ &= \operatorname{sgn}(\beta)\langle a, x\rangle + |\beta| \end{aligned}$$

since $\operatorname{sgn}(\beta) = -1$.

□

Now for $a \in \mathbb{R}^n$ and $\beta \neq 0$, the subdifferential of the function $f(x) = |\langle a, x\rangle + \beta|$ at the origin can be evaluated.

**Proposition 5.1.2.** *For fixed $a \in \mathbb{R}^n$ and $\beta \in \mathbb{R}$ with $\beta \neq 0$ define $f : \mathbb{R}^n \to \mathbb{R}$ as*

$$f(x) = |\langle a, x\rangle + \beta|$$

*Then*

$$\partial f(0) = \{\operatorname{sgn}(\beta)a\}$$

**Proof.**

If $a = 0$ then $f(x) = |\beta|$ and thus is differentiable at zero. Therefore,

$$\partial f(0) = \{\nabla f(0)\} = \{0\} = \{\operatorname{sgn}(\beta)a\}$$

To consider the case where $a \neq 0$ notice by Proposition (4.3.5), $\gamma \in \partial f(0)$ if and only if $f(x) \geq f(0) + \langle \gamma, x\rangle$ for all $x \in \mathcal{O}$ where $\mathcal{O}$ is an open neighborhood of zero. Thus let $\mathcal{O} = \{x \in \mathbb{R}^n : |\langle a, x\rangle| < |\beta|\}$. The fact that $\beta \neq 0$ implies $0 \in \mathcal{O}$ and thus $\mathcal{O}$ is nonempty. Furthermore note that $\mathcal{O}$ is open. To see why let $\|\cdot\|$ denote the norm induced by the inner product on $\mathbb{R}^n$. Then for a fixed $x \in \mathcal{O}$ notice $|\beta| - |\langle a, x\rangle| > 0$. Furthermore $\|a\| \neq 0$ since $a \neq 0$. Thus if $y \in \mathbb{R}^n$ such that

$$\|y - x\| < \frac{|\beta| - |\langle a, x\rangle|}{\|a\|}$$

then using the Cauchy-Schwarz inequality

$$\begin{aligned}
|\langle a, y\rangle| &= |\langle a, y - x + x\rangle| \\
&= |\langle a, y - x\rangle + \langle a, x\rangle| \\
&\leq |\langle a, y - x\rangle| + |\langle a, x\rangle| \\
&\leq \|a\| \cdot \|y - x\| + |\langle a, x\rangle| \\
&< \|a\| \left( \frac{|\beta| - |\langle a, x\rangle|}{\|a\|} \right) + |\langle a, x\rangle| \\
&= |\beta|
\end{aligned}$$

Hence $y \in \mathcal{O}$ and therefore

$$\left\{ y \in \mathbb{R}^n \,:\, \|y - x\| < \frac{|\beta| - |\langle a, x\rangle|}{\|a\|} \right\} \subseteq \mathcal{O}$$

which shows $\mathcal{O}$ is an open neighborhood of zero.

Now by the previous result $f(x) = \operatorname{sgn}(\beta)\langle a, x\rangle + |\beta|$ for $x \in \mathcal{O}$. Therefore $f$ is differentiable at zero and hence $\partial f(0) = \{\nabla f(0)\}$. Last,

$$\begin{aligned}
(\nabla f(x))_i &= \frac{\partial}{\partial x_i}(\operatorname{sgn}(\beta)\langle a, x\rangle + |\beta|) \\
&= \operatorname{sgn}(\beta)\frac{\partial}{\partial x_i}\langle a, x\rangle \\
&= \operatorname{sgn}(\beta)\frac{\partial}{\partial x_i}\sum_{j=1}^{n} a_j x_j \\
&= \operatorname{sgn}(\beta) a_i
\end{aligned}$$

Hence $\nabla f(x) = \operatorname{sgn}(\beta)a$ and therefore $\nabla f(0) = \operatorname{sgn}(\beta)a$, proving the result.

□

Now if $\beta = 0$ then the calculation of the subdifferential of $f(x) = |\langle a, x\rangle|$ at the origin involves more work since, in this case, $f$ is not differentiable there. To calculate this subdifferential, a sequence of specialized cases will be considered.

**Proposition 5.1.3.** *If $f : \mathbb{R}^n \to \mathbb{R}$ is defined as $f(y) := |y_1 + \cdots + y_n|$ then*

$$\partial f(0) = \left\{ (\lambda, \cdots, \lambda)^T \in \mathbb{R}^n \,:\, \lambda \in [-1, 1] \right\}$$

**Proof.**

By direct calculation

$$\begin{aligned}
\gamma \in \partial f(0) &\iff f(y) \geq f(0) + \langle \gamma, y - 0 \rangle && \text{for all } y \in \mathbb{R}^n \\
&\iff f(y) \geq \langle \gamma, y \rangle && \text{for all } y \in \mathbb{R}^n \\
&\iff \langle \gamma, y \rangle \leq |y_1 + \cdots + y_n| && \text{for all } y \in \mathbb{R}^n
\end{aligned}$$

Next notice $\gamma \in \partial f(0)$ if and only if $\langle \gamma, y \rangle \leq |y_1 + \cdots + y_n|$ for all $y$ if and only if $\langle -\gamma, -y \rangle \leq |-y_1 + \cdots + -y_n|$ for all $y$ if and only if $\langle -\gamma, y \rangle \leq |y_1 + \cdots + y_n|$ for all $y$ if and only if $-\gamma \in \partial f(0)$. This symmetry will be useful later.

Now if $n = 1$ then $\gamma \in \partial f(0)$ if and only if $\gamma_1 y_1 \leq |y_1|$ if and only if $|\gamma_1| \leq 1$ establishing the result for the $n = 1$ case. Therefore assume $n \geq 2$.

Then I claim that $\gamma$ cannot be in $\partial f(0)$ if $\gamma_p > 0$ for some $p$ and $\gamma_q < 0$ for some $q$. To see why notice if $\gamma$ satisfies this condition then setting $y = (y_1, \cdots, y_n)^T$ where $y_i = \operatorname{sgn}(\gamma_i)$ for $i = p$ or $i = q$ and $y_i = 0$ otherwise, one has

$$\langle \gamma, y \rangle = \gamma_p \operatorname{sgn}(\gamma_p) + \gamma_q \operatorname{sgn}(\gamma_q) = |\gamma_p| + |\gamma_q| > 0$$

However,

$$|y_1 + \cdots + y_n| = \operatorname{sgn}(\gamma_p) + \operatorname{sgn}(\gamma_q) = 0$$

Thus $\gamma \notin \partial f(0)$. Thus if $\gamma \in \partial f(0)$ then it must be that $\gamma_i \geq 0$ for all $i$ or $\gamma_i \leq 0$ for all $i$.

Next I claim if $\gamma$ is such that $\gamma_p > \gamma_q > 0$ for some $p$ and $q$ then $\gamma \notin \partial f(0)$. To see why define $y \in \mathbb{R}^n$ such that $y_p = 1$, $y_q = -1$, and $y_i = 0$ otherwise. Then

$$\langle \gamma, y \rangle = \gamma_p - \gamma_q > 0$$

However,

$$|y_1 + \cdots + y_n| = 0$$

Thus $\gamma \notin \partial f(0)$. Thus if $\gamma \in \partial f(0)$ and if $\gamma_i, \gamma_j > 0$ for some $i \neq j$ then $\gamma_i = \gamma_j$. However, since $\gamma \in \partial f(0)$ if and only if $-\gamma \in \partial f(0)$, one has if $\gamma \in \partial f(0)$ and if $\gamma_i, \gamma_j \neq 0$ for some $i \neq j$ then $\gamma_i = \gamma_j$.

Now I claim if $\gamma$ is such that $\gamma_p = 0$ for some $p$ and $\gamma_q > 0$ for some $q$ then $\gamma \notin \partial f(0)$. To see why define $y \in \mathbb{R}^n$ such that $y_p = -1$, $y_q = 1$, and $y_i = 0$ otherwise. Then

$$\langle \gamma, y \rangle = \gamma_p y_p + \gamma_q y_q = 0 \cdot (-1) + \gamma_q \cdot 1 = \gamma_q > 0$$

However,

$$|y_1 + \cdots + y_n| = 0$$

Thus $\gamma \notin \partial f(0)$.

Further if $\gamma'$ is such that $\gamma'_{p'} = 0$ for some $p'$ and $\gamma'_{q'} < 0$ for some $q'$ then, by the above calculation, one has that $-\gamma' \notin \partial f(0)$ and hence $\gamma' \notin \partial f(0)$ since $\gamma' \in \partial f(0)$ if and only if $-\gamma' \in \partial f(0)$.

Therefore, if $\gamma \in \partial f(0)$ then it must be that $\gamma = 0$ or $\gamma_i \neq 0$ for all $i$.

Now clearly $0 \in \partial f(0)$. Next, if $\gamma_i \neq 0$ for all $i$ then, by the above results, there exists $\lambda \neq 0$ such that $\gamma_i = \lambda$ for all $i$. Then $\gamma \in \partial f(0)$ if and only if

$$\langle \gamma, y \rangle = \lambda(y_1 + \cdots + y_n) \leq |y_1 + \cdots + y_n|$$

if and only if $|\lambda| \leq 1$.

Thus, collecting all our results, if $\gamma \in \partial f(0)$ then $\gamma \in \left\{(\lambda, \cdots, \lambda)^T \in \mathbb{R}^n \,:\, \lambda \in [-1, 1]\right\}$.

Last if $\gamma \in \left\{(\lambda, \cdots, \lambda)^T \in \mathbb{R}^n \,:\, \lambda \in [-1, 1]\right\}$ then for $y \in \mathbb{R}^n$

$$\langle \gamma, y \rangle = \lambda(y_1 + \cdots + y_n) \leq |y_1 + \cdots + y_n|$$

and hence $\gamma \in \partial f(0)$ establishing the equality and proving the result.

□

The next result generalizes the previous result.

**Proposition 5.1.4.** *If $f : \mathbb{R}^n \to \mathbb{R}$ is defined as $f(x) := |\langle \delta, x \rangle|$ where $\delta \in \mathbb{R}^n$ such that $\delta_i \in \{0, 1\}$ for each $i = 1, \ldots, n$ then*

$$\partial f(0) = \{\lambda \delta \,:\, \lambda \in [-1, 1]\}$$

**Proof.**

By direct calculation,

$$\begin{aligned}\gamma \in \partial f(0) &\iff f(x) \geq f(0) + \langle \gamma, x - 0 \rangle \text{ for all } x \\ &\iff f(x) \geq \langle \gamma, x \rangle \text{ for all } x \\ &\iff \gamma_1 x_1 + \cdots + \gamma_n x_n \leq |\delta_1 x_1 + \cdots + \delta_n x_n| \text{ for all } x\end{aligned}$$

Now let

$$\mathcal{I} = \{i = 1, \ldots, n \,:\, \delta_i = 1\}$$

Notice if $\mathcal{I} = \{1, \ldots, n\}$ then by Proposition (5.1.3)

$$\begin{aligned}\partial f(0) &= \left\{(\lambda, \ldots, \lambda)^T \,:\, \lambda \in [-1, 1]\right\} \\ &= \left\{\lambda(1, \ldots, 1)^T \,:\, \lambda \in [-1, 1]\right\} \\ &= \{\lambda\delta \,:\, \lambda \in [-1, 1]\}\end{aligned}$$

Next if $\mathcal{I} = \emptyset$ then

$$\gamma \in \partial f(0) \iff \gamma_1 x_1 + \cdots + \gamma_n x_n \leq 0 \text{ for all } x \in \mathbb{R}^n$$

Hence if $\gamma \in \partial f(0)$ then $\gamma_1 \operatorname{sgn}(\gamma_1) + \cdots + \gamma_n \operatorname{sgn}(\gamma_n) = \|\gamma\|_1 \leq 0$ and therefore $\gamma = 0$. Conversely if $\gamma = 0$ then $\gamma_1 x_1 + \cdots + \gamma_n x_n = 0 \leq 0$ for all $x \in \mathbb{R}^n$. Thus because $\delta = 0$ if $\mathcal{I} = \emptyset$

$$\partial f(0) = \{0\} = \{\delta\} = \{\lambda\delta \,:\, \lambda \in [-1, 1]\}$$

Now suppose $\mathcal{I} \neq \emptyset$ and $\mathcal{I} \neq \{1, \ldots, n\}$. Then if $\gamma \neq 0$ then there exists $p$ such that $\gamma_p \neq 0$. Thus defining $y \in \mathbb{R}^n$ by $y_p = \operatorname{sgn}(\gamma_p)$ and $y_i = 0$ otherwise, one has

$$\gamma_1 y_1 + \cdots + \gamma_n y_n = \gamma_p y_p = \gamma_p \operatorname{sgn}(\gamma_p) = |\gamma_p| > 0$$

Hence $\gamma \notin \partial f(0)$. Thus if $\mathcal{I} = \emptyset$ then $\partial f(0) = \{0\}$.

Now suppose $\mathcal{I} \neq \emptyset$ and $\mathcal{I} \neq \{1, \ldots, n\}$. Now I claim if $\gamma \in \partial f(0)$ then $\gamma_i = 0$ for all $i \notin \mathcal{I}$. To prove this let $\gamma \in \mathbb{R}^n$ such that $\gamma_p \neq 0$ for some $p \notin \mathcal{I}$. Now define $y \in \mathbb{R}^n$ such that

$y_p = \operatorname{sgn}(\gamma_p)$ and $y_i = 0$ otherwise. Then $p \notin \mathcal{I}$ implies $\delta_p = 0$. Hence

$$\begin{aligned} |\delta_1 y_1 + \cdots + \delta_n y_n| &= \delta_p y_p \qquad \text{since } y_i = 0 \text{ for } i \neq p \\ &= 0 \cdot \operatorname{sgn}(\gamma_p) \\ &= 0 \end{aligned}$$

However,

$$\begin{aligned} \gamma_1 y_1 + \cdots + \gamma_n y_n &= \gamma_p y_p \qquad \text{since } y_i = 0 \text{ for } i \neq p \\ &= \gamma_p \operatorname{sgn}(\gamma_p) \\ &= |\gamma_p| \\ &> 0 \end{aligned}$$

Hence $\gamma \notin \partial f(0)$.

Thus if $\gamma \in \partial f(0)$ then $\gamma_i = 0$ for all $i \notin \mathcal{I}$. Hence

$$\begin{aligned} \gamma \in \partial f(0) &\iff \gamma_1 x_1 + \cdots + \gamma_n x_n \leq |\delta_1 x_1 + \cdots + \delta_n x_n| \text{ for all } x \\ &\iff \sum_{i \in \mathcal{I}} \gamma_i x_i \leq \left| \sum_{i \in \mathcal{I}} \delta_i x_i \right| \text{ for all } x \\ &\iff \sum_{i \in \mathcal{I}} \gamma_i x_i \leq \left| \sum_{i \in \mathcal{I}} x_i \right| \text{ for all } x \\ &\iff \text{there exists } -1 \leq \lambda \leq 1 \text{ such that } \gamma_i = \lambda \text{ for all } i \in \mathcal{I} \end{aligned}$$

from the results of Proposition (5.1.3). Thus

$$\begin{aligned} \gamma \in \partial f(0) &\iff \gamma \in \{\gamma \in \mathbb{R}^n \,:\, \exists\, \lambda \in [-1,1] \text{ such that } \gamma_i = \lambda,\, \forall\, i \in \mathcal{I} \text{ and } \gamma_i = 0,\, \forall\, i \notin \mathcal{I}\} \\ &\iff \gamma_i = \lambda \delta_i \qquad \forall\, i = 1, \ldots, n \\ &\iff \gamma = \lambda \delta \end{aligned}$$

□

Given $a \in \mathbb{R}^n$, the next result provides a complete characterization of the subdifferential of the function $f(x) = |\langle a, x \rangle|$ at the origin.

**Proposition 5.1.5.** *If $f : \mathbb{R}^n \to \mathbb{R}$ is defined as $f(x) = |\langle a, x\rangle|$ where $a \in \mathbb{R}^n$, then*

$$\partial f(0) = \{\lambda a \,:\, \lambda \in [-1, 1]\}$$

**Proof.**

Referring to Proposition (5.1.4), define $\tau : \mathbb{R} \to \mathbb{R}$ by

$$\tau(x) := \begin{cases} x & \text{if } x \neq 0 \\ 1 & \text{if } x = 0 \end{cases}$$

and define $\delta \in \mathbb{R}^n$ by

$$\delta_i = \begin{cases} 0 & \text{if } a_i = 0 \\ 1 & \text{otherwise} \end{cases}$$

Then set $y \in \mathbb{R}^n$ by $y_i = \tau(a_i)x_i$ and notice

$$\langle a, x\rangle = \sum_{i=1}^{n} a_i x_i = \sum_{i=1}^{n} \delta_i a_i x_i = \sum_{i=1}^{n} \delta_i \tau(a_i) x_i = \sum_{i=1}^{n} \delta_i y_i = \langle \delta, y\rangle$$

Furthermore if $\gamma \in \mathbb{R}^n$ then

$$\langle \gamma, x\rangle = \sum_{i=1}^{n} \gamma_i x_i = \sum_{i=1}^{n} \frac{\gamma_i}{\tau(a_i)} \tau(a_i) x_i = \sum_{i=1}^{n} \xi_i y_i = \langle \xi, y\rangle$$

where $\xi \in \mathbb{R}^n$ is defined by

$$\xi_i = \frac{\gamma_i}{\tau(a_i)}$$

for $i = 1, \ldots, n$.

Next notice given $z \in \mathbb{R}^n$ there exists $x \in \mathbb{R}^n$ such that $z_i = \tau(a_i)x_i$ for all $i = 1, \ldots, n$ because $\tau(a_i) \neq 0$ for all $i$. Particularly set $x_i = z_i/\tau(a_i)$ for each $i$. Therefore setting $g(x) := |\langle \delta, y\rangle|$ and using the previous result,

$$\begin{aligned}
\gamma \in \partial f(0) &\iff f(x) \geq f(0) + \langle \gamma, x\rangle && \text{for all } x \in \mathbb{R}^n \\
&\iff |\langle a, x\rangle| \geq \langle \gamma, x\rangle && \text{for all } x \in \mathbb{R}^n \\
&\iff |\langle \delta, y\rangle| \geq \langle \xi, y\rangle && \text{for all } y \in \mathbb{R}^n \\
&\iff g(y) \geq g(0) + \langle \xi, y - 0\rangle && \text{for all } y \in \mathbb{R}^n \\
&\iff \xi \in \partial g(0) \\
&\iff \xi = \lambda\delta && \text{for some } \lambda \in [-1, 1]
\end{aligned}$$

Further for $\lambda \in [-1,1]$ notice $\xi = \lambda\delta$ if and only if

$$\frac{\gamma_i}{\tau(a_i)} = \lambda\delta_i$$

for all $i = 1, \ldots, n$ if and only if $\gamma_i = \lambda\delta_i\tau(a_i) = \lambda a_i$ for all $i = 1, \ldots, n$ if and only if $\gamma = \lambda a$. Thus

$$\partial f(0) = \{\lambda a \,:\, \lambda \in [-1,1]\}$$

□

## 5.2 The Main Result

The above results, which describe the subdifferential of the function $f(x) = |\langle a, x\rangle + \beta|$ at the origin, can be extended to a similar function on $\mathbb{R}^{m\times n}$. This is the goal of the next proposition and is the main result of this chapter.

**Proposition 5.2.1.** *Fix $A \in \mathbb{R}^{n\times m}$, $B \in \mathbb{R}^{N\times M}$, and $C \in \mathbb{R}^{n\times M}$ and define $f : \mathbb{R}^{m\times N} \to \mathbb{R}$ by*

$$f(X) := \|AXB + C\|_1$$

*Then*

$$\partial f(0) = \left\{A^T(\Lambda + \mathrm{Sgn}(C))B^T \,:\, \Lambda \in \mathbb{R}^{n\times M} \text{ where } \|\Lambda\|_\infty \leq 1 \text{ and if } C_{i,j} \neq 0 \text{ then } \Lambda_{i,j} = 0\right\}$$

**Proof.**

Notice for any $X \in \mathbb{R}^{m\times N}$

$$\begin{aligned}
f(X) &= \|AXB + C\|_1 \\
&= \sum_{j=1}^{M}\sum_{i=1}^{n} |(AXB)_{i,j} + C_{i,j}| \\
&= \sum_{j=1}^{M}\sum_{i=1}^{n} \left|\sum_{\ell=1}^{N}\sum_{k=1}^{m} A_{i,k}X_{k,\ell}B_{\ell,j} + C_{i,j}\right| \\
&= \sum_{j=1}^{M}\sum_{i=1}^{n} \left|\sum_{\ell=1}^{N}\sum_{k=1}^{m} A_{i,k}B_{\ell,j}X_{k,\ell} + C_{i,j}\right| \\
&= \sum_{j=1}^{M}\sum_{i=1}^{n} f_{i,j}(X)
\end{aligned}$$

where for $i = 1, \ldots, n$ and $j = 1, \ldots, M$ the function $f_{i,j} : \mathbb{R}^{m \times N} \to \mathbb{R}$ is defined by

$$f_{i,j}(X) := \left| \sum_{\ell=1}^{N} \sum_{k=1}^{m} D_{i,j,k,\ell} X_{k,\ell} + C_{i,j} \right|$$

where $D_{i,j,k,\ell} = A_{i,k} B_{\ell,j}$. Then by Theorem 23.8, [25],

$$\partial f(0) = \sum_{j=1}^{M} \sum_{i=1}^{n} \partial f_{i,j}(0)$$

Now for any positive integer $p$ set $\mathbb{Z}_p := \{1, \ldots, p\}$ and let $\sigma : \mathbb{Z}_{mN} \to \mathbb{Z}_m \times \mathbb{Z}_N$ be a bijection. Such a bijection exists since $|\mathbb{Z}_m \times \mathbb{Z}_N| = |\mathbb{Z}_{mN}| < \infty$. Then for any $i = 1, \ldots, n$ and $j = 1, \ldots, M$

$$\begin{aligned} f_{i,j}(X) &= \left| \sum_{\ell=1}^{N} \sum_{k=1}^{m} D_{i,j,k,\ell} X_{k,\ell} + C_{i,j} \right| \\ &= \left| \sum_{t=1}^{mN} D_{i,j,\sigma(t)} X_{\sigma(t)} + C_{i,j} \right| \\ &= \left| \left\langle d^{(i,j)}, x \right\rangle + C_{i,j} \right| \end{aligned}$$

where $d^{(i,j)} \in \mathbb{R}^{mN}$ is defined by $d_k^{(i,j)} = D_{i,j,\sigma(k)}$ and $x \in \mathbb{R}^{mN}$ is defined by $x_k = X_{\sigma(k)}$. Based on this for any $i = 1, \ldots, n$ and $j = 1, \ldots, M$ define $g_{i,j} : \mathbb{R}^{mN} \to \mathbb{R}$ by

$$g_{i,j}(x) := \left| \left\langle d^{(i,j)}, x \right\rangle + C_{i,j} \right|$$

Then because $\sigma$ is bijective, for $i = 1, \ldots, n$ and $j = 1, \ldots, M$ if $\Gamma^{(i,j)} \in \mathbb{R}^{m \times N}$ and $\gamma^{(i,j)} \in \mathbb{R}^{mN}$ is defined by $\gamma_k^{(i,j)} = \Gamma_{\sigma(k)}^{(i,j)}$ for $k = 1, \ldots, mN$ then

$$\begin{aligned} \Gamma^{(i,j)} \in \partial f_{i,j}(0) &\iff f_{i,j}(X) \geq f_{i,j}(0) + \left\langle \Gamma^{(i,j)}, X \right\rangle && \text{for all } X \in \mathbb{R}^{m \times N} \\ &\iff f_{i,j}(X) \geq f_{i,j}(0) + \sum_{\ell=1}^{N} \sum_{k=1}^{m} \Gamma_{k,\ell}^{(i,j)} X_{k,\ell} && \text{for all } X \in \mathbb{R}^{m \times N} \\ &\iff f_{i,j}(X) \geq f_{i,j}(0) + \sum_{t=1}^{mN} \Gamma_{\sigma(t)}^{(i,j)} X_{\sigma(t)} && \text{for all } X \in \mathbb{R}^{m \times N} \\ &\iff g_{i,j}(x) \geq |C_{i,j}| + \left\langle \gamma^{(i,j)}, x \right\rangle && \text{for all } x \in \mathbb{R}^{mN} \\ &\iff \gamma^{(i,j)} \in \partial g_{i,j}(0) \end{aligned}$$

Next given $i = 1, \ldots, n$ and $j = 1, \ldots, M$ if $C_{i,j} = 0$ then $\delta(C_{i,j}) = 1$ and hence

$$\begin{aligned}\partial g_{i,j}(0) &= \left\{\lambda_{i,j} d^{(i,j)} \ : \ \lambda_{i,j} \in [-1,1]\right\} \\ &= \left\{(\mathrm{Sgn}(C_{i,j}) + \lambda_{i,j}\delta(C_{i,j}))d^{(i,j)} \ : \ \lambda_{i,j} \in [-1,1]\right\}\end{aligned}$$

Furthermore if $C_{i,j} \neq 0$ then $\delta(C_{i,j}) = 0$ and hence

$$\begin{aligned}\partial g_{i,j}(0) &= \left\{\mathrm{sgn}(C_{i,j}) d^{(i,j)}\right\} \\ &= \left\{(\mathrm{sgn}(C_{i,j}) + \lambda_{i,j}\delta(C_{i,j}))d^{(i,j)} \ : \ \lambda_{i,j} \in [-1,1]\right\}\end{aligned}$$

Therefore for any $i = 1, \ldots, n$ and $j = 1, \ldots, M$

$$\partial g_{i,j}(0) = \left\{(\mathrm{sgn}(C_{i,j}) + \lambda_{i,j}\delta(C_{i,j}))d^{(i,j)} \ : \ \lambda_{i,j} \in [-1,1]\right\}$$

regardless of whether or not $C_{i,j}$ is nonzero. Thus for $\Gamma \in \mathbb{R}^{m\times N}$ define $\gamma \in \mathbb{R}^{mN}$ by $\gamma_k = \Gamma_{\sigma(k)}$. Then $\Gamma \in \partial f(0)$ if and only if

$$\gamma \in \sum_{j=1}^{M}\sum_{i=1}^{n} \left\{(\mathrm{sgn}(C_{i,j}) + \lambda_{i,j}\delta(C_{i,j}))d^{(i,j)} \ : \ \lambda_{i,j} \in [-1,1]\right\}$$

if and only if

$$\gamma = \sum_{j=1}^{M}\sum_{i=1}^{n} (\mathrm{sgn}(C_{i,j}) + \lambda_{i,j}\delta(C_{i,j}))d^{(i,j)}$$

for some $\lambda_{i,j}$ such that $\lambda_{i,j} \in [-1,1]$ for all $i = 1, \ldots, n$ and $j = 1, \ldots, M$. Then given $(k,\ell) \in \mathbb{Z}_m \times \mathbb{Z}_N$ let $t \in \mathbb{Z}_{mN}$ be such that $\sigma(t) = (k,\ell)$. Then

$$\begin{aligned}\Gamma_{k,\ell} &= \Gamma_{\sigma(t)} \\ &= \gamma_t \\ &= \sum_{j=1}^{M}\sum_{i=1}^{n} (\mathrm{sgn}(C_{i,j}) + \lambda_{i,j}\delta(C_{i,j}))d_t^{(i,j)} \\ &= \sum_{j=1}^{M}\sum_{i=1}^{n} (\mathrm{sgn}(C_{i,j}) + \lambda_{i,j}\delta(C_{i,j}))D_{i,j,\sigma(t)} \\ &= \sum_{j=1}^{M}\sum_{i=1}^{n} (\mathrm{sgn}(C_{i,j}) + \lambda_{i,j}\delta(C_{i,j}))D_{i,j,k,\ell} \\ &= \sum_{j=1}^{M}\sum_{i=1}^{n} (\mathrm{sgn}(C_{i,j}) + \lambda_{i,j}\delta(C_{i,j}))A_{i,k}B_{\ell,j}\end{aligned}$$

$$= \sum_{j=1}^{M} \sum_{i=1}^{n} A_{k,i}^T \operatorname{sgn}(C_{i,j}) B_{j,\ell}^T + \sum_{j=1}^{N} \sum_{i=1}^{m} A_{k,i}^T \lambda_{i,j} \delta(C_{i,j}) B_{j,\ell}^T$$

$$= (A^T \operatorname{Sgn}(C) B^T)_{k,\ell} + (A^T \Lambda B^T)_{k,\ell}$$

where $\Lambda \in \mathbb{R}^{n \times M}$ is such that $\Lambda_{i,j} \in [-1, 1]$ for all $i = 1, \ldots, n$ and $j = 1, \ldots, M$ and if $C_{i,j} \neq 0$ then $\Lambda_{i,j} = 0$. Therefore

$$\Gamma = A^T(\Lambda + \operatorname{Sgn}(C))B^T$$

□

# CHAPTER 6. Introducing the Method

## 6.1 An Introduction to Compressed Sensing

The area of compressed sensing focuses on finding sparse solutions to underdetermined system of linear equations. Specifically, let $\|x\|_0$ denote the the $\ell_0$ pseudo-norm of $x$ which counts the number of nonzero entries of $x$. Note $\|\cdot\|_0$ is not a true norm since, for example, $\|\alpha x\|_0 = \|x\|_0$ for all $\alpha \neq 0$. A vector $x \in \mathbb{R}^n$ then is said to be *s-sparse* if $\|x\|_0 \leq s$. Compressed sensing focuses on addressing for which $s$ an $s$-sparse solution to an underdetermined system is necessarily the sparsest solution. This section has a brief introduction to compressed sensing. For a more detailed introduction see [1], [5], and [4].

That is, given $A \in \mathbb{R}^{m \times n}$ with $m \ll n$ and $b \in \mathbb{R}^n$ consider the optimization problem

$$\text{minimize} \|x\|_0 \quad \text{subject to} \quad Ax = b \tag{$P_0$}$$

As shown in [24], solving problem ($P_0$) is NP-hard. A common way to remedy this obstacle, as described in [10], is to instead consider the following convex relaxation of problem ($P_0$).

$$\text{minimize} \|x\|_1 \quad \text{subject to} \quad Ax = b \tag{$P_1$}$$

Specifically since $\|\cdot\|_1$ is convex, problem ($P_1$) is a convex optimization problem, and many methods, such as the simplex method [12] and interior-point methods [3], exist to solve such problems. Furthermore, if the solution to problem ($P_1$) is sparse enough, it is also a solution of problem ($P_0$). Thus, sometimes, problem ($P_1$) can be used to solve problem ($P_0$).

There are many conditions analyzed in the compressed sensing literature which guarantee a solution to problem ($P_1$) solves problem ($P_0$). Focus will be given to the condition developed in [15] which analyzes the case when

$$A = [A_1 \mid A_2]$$

where $A_1, A_2 \in \mathbb{R}^{n\times n}$ are real orthogonal matrices. To analyze this case, the authors defined the *mutual incoherence* between $A_1$ and $A_2$ as

$$M(A_1, A_2) := \max_{1\le i,j\le n} |\langle (A_1)_i, (A_2)_j\rangle| = \left\|A_1 A_2^T\right\|_\infty \tag{6.1}$$

where $(A_1)_i$ denotes the $i$th row of $A_1$. They then established the following theorem.

**Theorem 1, [15].** *Let $A_1, A_2 \in \mathbb{R}^{n\times n}$ be real orthogonal matrices and $b \in \mathbb{R}^n$. If $x$ solves $[A_1 \mid A_2]x = b$ and $\|x\|_0 < 1/M(A_1, A_2)$ then $x$ is necessarily the sparsest solution.*

Furthermore, the following theorem is shown in [16].

**Theorem 3, [16].** *Let $A_1, A_2 \in \mathbb{R}^{n\times n}$ be real orthogonal matrices and $b \in \mathbb{R}^n$. If there exists any $x$ satisfying $[A_1 \mid A_2]x = b$ such that*

$$\|x\|_0 < \left(\sqrt{2} - \frac{1}{2}\right)\frac{1}{M(A_1, A_2)}$$

*then it is necessarily the unique solution to problem ($P_1$).*

Therefore if the system $[A_1 \mid A_2]x = b$ has a solution $x$ that is significantly sparse, problem ($P_1$) will find it, and if $\|x\|_0 < 1/M(A_1, A_2)$ then $x$ is also a solution to problem ($P_0$).

## 6.2 A Prototype Basis Identification Method

Notice if $x$ is a solution to problem ($P_1$) and $\mathcal{I} = \operatorname{supp} x = \{i : x_i \neq 0\}$ then $b \in \operatorname{span}\{\widetilde{a}_i\}_{i\in\mathcal{I}}$ where $\widetilde{a}_i$ denotes the $i$th column of $A$. It has been shown in [10] that if $x$ is the sparsest solution to problem ($P_1$) then in fact $\{\widetilde{a}_i\}_{i\in\mathcal{I}}$ are linearly independent. Thus $\{\widetilde{a}_i\}_{i\in\mathcal{I}}$ is a basis for a subspace that contains $b$. However, as shown in [15], if $x$ is the sparsest solution then $\|x\|_0 \le \sqrt{n}$ and therefore set $\{\widetilde{a}_i\}_{i\in\mathcal{I}}$ does not contain enough elements to be a basis for $\mathbb{R}^n$. Instead it is a basis for a subspace of dimension at most $\sqrt{n}$. Also notice the selection of $\{\widetilde{a}_i\}_{i\in\mathcal{I}}$ is dependent on $b$.

Now given an injective matrix $X \in \mathbb{R}^{m\times n}$ consider the problem

$$\text{minimize } \|Y\|_1 \qquad \text{subject to} \qquad YX = I \tag{$\widetilde{P}_1$}$$

Notice if $X$ is injective then the rows of $X$, denoted $\mathbb{X}$, form a spanning set for $\mathbb{R}^n$ and are therefore a frame for $\mathbb{R}^n$. Thus by Proposition (3.3.4), if $\mathbb{X}$ is not a basis, that is if $X$ is not bijective, there exists infinitely many dual frames of $\mathbb{X}$.

Let $\mathbb{Y}$ be one of these dual frames. Recall then $\Theta_{\mathbb{Y}}^*\Theta_{\mathbb{X}} = I$. However, $\Theta_{\mathbb{X}} = X$ since the rows of the analysis operator are the elements of $\mathbb{X}$ since $\mathbb{X} \subseteq \mathbb{R}^n$. Therefore $\Theta_{\mathbb{Y}}^*$ is a matrix such that $\Theta_{\mathbb{Y}}^* X = I$. Conversely, if $Y$ is such that $YX = I$ then the columns of $Y$ form a dual of $\mathbb{X}$. Therefore $Y$ is such that $YX = I$ if and only if the columns of $Y$ form a dual to the rows of $X$.

Now suppose the first $n$ rows of $X$ are a basis for $\mathbb{R}^n$ and write

$$X = \left[\frac{X_1}{X_2}\right] \tag{6.2}$$

which denotes the $2 \times 1$ block matrix with blocks $X_1$ and $X_2$.

Notice then $X_1$ is invertible. Now let $Z = [X_1^{-1} \mid 0]$, the $1 \times 2$ block matrix with blocks $X_1^{-1}$ and 0. Then $ZX = I$ and therefore the columns of $Z$ form a dual frame of $X$. Next $X_1 \in \mathbb{R}^{n\times n}$ and therefore $Z$ has the nice property that $m - n$ of the columns of $Z$ are zero and the nonzero columns of $Z$ correspond to the basis for $\mathbb{R}^n$ consisting of the rows of $X_1$. That is the submatrix of $Z$ obtained by removing the zero columns of $Z$ is $X_1^{-1}$ and the inverse of this matrix is $X_1$. Last the rows of $X_1$ are a basis for $\mathbb{R}^n$. This idea can be extended to construct a dual frame that identifies, in the same way, any subset of the rows of $X$ that form a basis for $\mathbb{R}^n$. These special types of dual frames are said to *correspond to a basis*.

**Definition 6.2.1.** *Given an injective matrix $A \in \mathbb{R}^{m\times n}$, a matrix $Y \in \mathbb{R}^{n\times m}$ satisfying $YX = I$ is said to* correspond to a basis *if $m - n$ columns of $Y$ are columns of zeros.*

If $X \in \mathbb{R}^{m\times n}$ is an injective matrix and there exists a matrix $Y \in \mathbb{R}^{n\times m}$ that is a solution to problem ($\widetilde{P}_1$) such that $Y$ corresponds to a basis, it will be said that problem ($\widetilde{P}_1$) *identifies a basis in $X$*.

Now problem ($\widetilde{P}_1$) is inspired from problem ($P_1$) where $b$ is replaced with $I$, $A$ is replaced with $X$, and $x$ is replaced with $Y$. With this form, the constraint $YX = I$ identifies all matrices $Y$ that are synthesis operators of duals of the frame formed from the rows of $X$. Then, just

as problem ($P_1$) identifies linearly independent columns of $A$, the hope is that the solution of problem ($\widetilde{P}_1$) will identify a subset of the rows of $X$ that are linearly independent. Furthermore, since a solution to problem ($\widetilde{P}_1$) must be a left inverse of $X$, perhaps the subset of rows selected from $X$ are in fact a basis. This section will address when this is in fact true.

First, the above description of problem ($\widetilde{P}_1$) assumes the minimum value of the problem is finite and is actually achieved for some matrix $Y$. The next result shows that this is in fact true.

**Proposition 6.2.1.** *Let $X \in \mathbb{R}^{m\times n}$ be an injective matrix. Then there exists $Y_0 \in \mathbb{R}^{n\times m}$ such that*

$$\|Y_0\|_1 = \inf_{YX=I} \|Y\|_1$$

**Proof.**

Since $X$ is injective, $X^\dagger$, the Moore-Penrose inverse of $X$, satisfies $X^\dagger X = I$ by Proposition (2.1.2). Now consider the set

$$\mathcal{A} = \left\{ Y \in \mathbb{R}^{n\times m} \,:\, YX = I \text{ and } \|Y\|_1 \leq \|X^\dagger\|_1 \right\}.$$

Notice $\mathcal{A}$ is bounded by construction and if $\{Y_k\}_{k=1}^\infty \subseteq \mathcal{A}$ such that $Y_k \to Y$ for some $Y \in \mathbb{R}^{n\times m}$ then $YX = (\lim_k Y_k)X = \lim_k (Y_k X) = I$ and

$$\|Y\|_1 = \left\| \lim_k Y_k \right\|_1 = \lim_k \|Y_k\|_1 \leq \|X^\dagger\|_1.$$

Thus $\mathcal{A}$ is compact by the Heine-Borel Theorem.

Next notice since $X^\dagger X = I$ if $Z \in \mathbb{R}^{n\times m}$ is such that $ZX = I$ satisfies $\|Z\|_1 > \|X^\dagger\|_1$ then $Z$ cannot be a minimizer of $\|Y\|_1$ subject to $YX = I$. Thus

$$\inf_{YX=I} \|Y\|_1 = \inf_{Y\in\mathcal{A}} \|Y\|_1$$

and because $\|\cdot\|_1$ is continuous, since it is a norm, and $\mathcal{A}$ is compact, there exists, $Y_0 \in \mathcal{A}$ where the infimum $\inf_{Y\in\mathcal{A}} \|Y\|_1$ is attained.

□

The following result also shows that reordering the rows of $X$ simply reorders the columns of the solutions of problem ($\widetilde{P}_1$). To precisely state the result, given a matrix $A \in \mathbb{R}^{m\times n}$ and a permutation $\sigma : \{1,\dots,m\} \to \{1,\dots,m\}$ define $A_\sigma \in \mathbb{R}^{m\times n}$ as

$$(A_\sigma)_{i,j} = A_{\sigma(i),j}$$

Next if $\psi : \{1,\dots,n\} \to \{1,\dots,n\}$ is a permutation, define $X^\psi \in \mathbb{R}^{m\times n}$ as

$$(A^\psi)_{i,j} = A_{i,\psi(j)}$$

**Proposition 6.2.2.** *Let $X \in \mathbb{R}^{m\times n}$ be a real orthogonal matrix and $\sigma : \{1,\dots,m\} \to \{1,\dots,m\}$ a permutation. Then $Y$ is a solution of*

$$\text{minimize } \|Y\|_1 \qquad \text{subject to} \qquad YX = I$$

*if and only if $Y^\sigma$ is a solution of*

$$\text{minimize } \|Y\|_1 \qquad \text{subject to} \qquad YX_\sigma = I$$

**Proof.**

Let $i, j = 1,\dots,n$ and notice since $\sigma$ is bijective,

$$\begin{aligned}(YX)_{i,j} &= \sum_{k=1}^{m} Y_{i,k}X_{k,j} \\ &= \sum_{k=1}^{m} Y_{i,\sigma(k)}X_{\sigma(k),j} \\ &= \sum_{k=1}^{m} (Y^\sigma)_{i,k}(X_\sigma)_{k,j} \\ &= (Y^\sigma X_\sigma)_{i,j}\end{aligned}$$

Thus $YX = Y^\sigma X_\sigma$. Furthermore

$$\begin{aligned}\|Y^\sigma\|_1 &= \sum_{j=1}^{m}\sum_{i=1}^{n} |(Y^\sigma)_{i,j}| \\ &= \sum_{j=1}^{m}\sum_{i=1}^{n} \left|Y_{i,\sigma(j)}\right| \\ &= \sum_{j=1}^{m}\sum_{i=1}^{n} |Y_{i,j}| \\ &= \|Y\|_1\end{aligned}$$

Therefore

$$\inf_{YX=I} \|Y\|_1 = \inf_{YX=I} \|Y^\sigma\|_1 = \inf_{Y^\sigma X_\sigma = I} \|Y^\sigma\| = \inf_{YX_\sigma = I} \|Y\|_1$$

which completes the proof since the infima above must be attained for some matrices by the previous proposition.

□

The following result provides necessary and sufficient conditions for a particular subset of the rows of an injective matrix, that form a basis, to be identified by problem ($\widetilde{P}_1$).

**Proposition 6.2.3.** *Consider*

$$X = \left[\frac{X_1}{X_2}\right]$$

*where $X_1 \in \mathbb{R}^{n\times n}$ is invertible and $X_2 \in \mathbb{R}^{m\times n}$. Then $[X_1^{-1} \,|\, 0]$ is a minimizer of problem ($\widetilde{P}_1$) if and only if there exists $\Lambda_1 \in \mathbb{R}^{n\times m}$ and $\Lambda_2 \in \mathbb{R}^{n\times n}$ such that*

$$\Lambda_1 + \Lambda_2(X_2X_1^{-1})^T = \mathrm{Sgn}(X_1^{-1})(X_2X_1^{-1})^T$$

*where $\|\Lambda_1\|_\infty, \|\Lambda_2\|_\infty \leq 1$ and $(\Lambda_2)_{i,j} = 0$ if $(X_1)^{-1}_{i,j} \neq 0$.*

**Proof.**

Notice

$$I = [Y_1 \,|\, Y_2]\left[\frac{X_1}{X_2}\right] = Y_1X_1 + Y_2X_2$$

if and only if $Y_1 = X_1^{-1} - Y_2X_2X_1^{-1}$. Therefore $[X_1^{-1} \,|\, 0]$ is a minimizer of problem ($\widetilde{P}_1$) if and only if the zero matrix is a minimizer of

$$\text{minimize } \|Y\|_1 + \left\|X_1^{-1} - YX_2X_1^{-1}\right\|_1 \quad \text{subject to} \quad Y \in \mathbb{R}^{n\times m}$$

Next let $f(Y) = \|Y\|_1 + \left\|X_1^{-1} - YX_2X_1^{-1}\right\|_1$. Then, by Proposition (5.2.1), zero is a minimizer of the $f$ if and only if

$$0 \in \partial f(0) = \{\Lambda_1 + (\Lambda_2 - \mathrm{Sgn}(X_1^{-1}))(X_2X_1^{-1})^T \,:\, \|\Lambda_1\|_\infty, \|\Lambda_2\|_\infty \leq 1 \text{ and if } (X_1^{-1})_{i,j} \neq 0 \text{ then } (\Lambda_2)_{i,j} = 0\}$$

if and only if

$$\Lambda_1 + \Lambda_2(X_2X_1^{-1})^T = \operatorname{Sgn}(X_1^{-1})(X_2X_1^{-1})^T$$

for some $\Lambda_1 \in \mathbb{R}^{n\times m}$ and $\Lambda_2 \in \mathbb{R}^{n\times n}$ such that $\|\Lambda_k\|_\infty \leq 1$ for $k = 1, 2$ and $(\Lambda_2)_{i,j} = 0$ if $(X_1^{-1})_{i,j} \neq 0$.

□

**Corollary 6.2.4.** *Consider*

$$X = \left[\frac{X_1}{X_2}\right]$$

*where $X_1 \in \mathbb{R}^{n\times n}$ is invertible, $X_2 \in \mathbb{R}^{m\times n}$, and $(X_1^{-1})_{i,j} \neq 0$ for all $i$ and $j$. Then $[X_1^{-1} \mid 0]$ is a minimizer of*

$$\text{minimize } \|Y\|_1 \qquad \text{subject to} \qquad YX = I$$

*if and only if*

$$\left\|\operatorname{Sgn}(X_1^{-1})(X_2X_1^{-1})^T\right\|_\infty \leq 1$$

**Proof.**

This result follows directly from the previous proposition.

□

The above propositions provide necessary and sufficient conditions for problem ($\widetilde{P}_1$) to identify a dual that corresponds to a basis. However, consider the injective matrix

$$X = \begin{bmatrix} \cos(\pi/3) & -\sin(\pi/3) \\ \sin(\pi/3) & \cos(\pi/3) \\ \cos(\pi/6) & -\sin(\pi/6) \\ \sin(\pi/6) & \cos(\pi/6) \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & -\frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & \frac{1}{2} \\ \frac{\sqrt{3}}{2} & -\frac{1}{2} \\ \frac{1}{2} & \frac{\sqrt{3}}{2} \end{bmatrix}$$

Table 6.2 lists all the duals of $X$ that correspond to a basis and $\|Y\|_1$ for each of these matrices.

Notice the rows of $X$ is a concatenation of two orthonormal bases. Next since $\binom{4}{2} = 6$ and a basis for $\mathbb{R}^2$ has two elements, there are six duals that correspond to a basis. Therefore since the first six matrices in the above table are left inverses of $X$, they are the complete list of all duals of $X$ that correspond to a basis. Now let $Y_i$ for $i = 1, \ldots, 6$ denote these duals. Next

| Dual $Y$ | $\|Y\|_1$ |
|---|---|
| $\begin{bmatrix} \frac{1}{2} & \frac{\sqrt{3}}{2} & 0 & 0 \\ -\frac{\sqrt{3}}{2} & \frac{1}{2} & 0 & 0 \end{bmatrix}$ | $1+\sqrt{3} \approx 2.73205$ |
| $\begin{bmatrix} -1 & 0 & \sqrt{3} & 0 \\ -\sqrt{3} & 0 & 1 & 0 \end{bmatrix}$ | $2+2\sqrt{3} \approx 5.4641$ |
| $\begin{bmatrix} 1 & 0 & 0 & 1 \\ -\frac{1}{\sqrt{3}} & 0 & 0 & \frac{1}{\sqrt{3}} \end{bmatrix}$ | $2+\frac{2}{\sqrt{3}} \approx 3.1547$ |
| $\begin{bmatrix} 0 & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & 0 \\ 0 & 1 & -1 & 0 \end{bmatrix}$ | $2+\frac{2}{\sqrt{3}} \approx 3.1547$ |
| $\begin{bmatrix} 0 & \sqrt{3} & 0 & -1 \\ 0 & -1 & 0 & \sqrt{3} \end{bmatrix}$ | $2+2\sqrt{3} \approx 5.4641$ |
| $\begin{bmatrix} 0 & 0 & \frac{\sqrt{3}}{2} & \frac{1}{2} \\ 0 & 0 & -\frac{1}{2} & \frac{\sqrt{3}}{2} \end{bmatrix}$ | $1+\sqrt{3} \approx 2.73205$ |
| $\begin{bmatrix} 0 & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & 0 \\ -\frac{1}{\sqrt{3}} & 0 & 0 & \frac{1}{\sqrt{3}} \end{bmatrix}$ | $\frac{4}{\sqrt{3}} \approx 2.3094$ |

Table 6.1 An example where Problem ($\widetilde{P}_1$) does not identify a basis.

let $Y_7$ denote the last dual which does not correspond to a basis. Notice $\|Y_7\|_1 < \|Y_i\|_1$ for $i = 1, \ldots, 6$. Therefore $Y_i$, for $i = 1, \ldots, 6$, cannot be a minimizer of $\|Y\|_1$ subject to $YX = I$.

Hence the above injective matrix $X$ shows that, in general, problem ($\widetilde{P}_1$) does not always have a solution that corresponds to a basis. Under some circumstances, precisely described by Proposition (6.2.3), problem ($\widetilde{P}_1$) does identify a basis. However, since problem ($\widetilde{P}_1$) does not always identify a basis perhaps problem ($\widetilde{P}_1$) can be modified to always identify a basis. This is the topic of the next section.

## 6.3 A Novel Basis Identification Method

As shown in the previous section, given an injective matrix $X \in \mathbb{R}^{m\times n}$ problem ($\widetilde{P}_1$) does not always identify a subset of the rows of $X$ that form a basis for $\mathbb{R}^n$. This section introduces another optimization problem, also inspired by problem ($P_1$), which is conjectured to always identify a basis in the rows of the injective matrix $X$. Code was written to try to find a counterexample to the conjecture. That is, the code attempted to find an injective matrix that did not have a basis identified by the new optimization problem introduced in this section. However, unlike problem ($\widetilde{P}_1$), after analyzing millions of randomly generated injective matrices, no counterexamples were found.

To describe this new optimization problem, given a matrix $A \in \mathbb{R}^{m\times n}$ let $\|A\|_0$ denote the number of nonzero entries in the matrix $A$. Then, given injective matrix $X \in \mathbb{R}^{m\times n}$, consider the following optimization problem.

$$\text{minimize } \|XY\|_0 \quad \text{subject to} \quad YX = I \tag{$P_0'$}$$

Just as with problem ($\widetilde{P}_1$), if $Y$ is a matrix such that $YX = I$ then the columns of $Y$ form a dual of the rows of $X$. Thus the columns of $Y$ are a frame for $\mathbb{R}^n$ and thus span $\mathbb{R}^n$. Hence there can be at most $m - n$ columns of zeros in $Y$. Otherwise, there would not be enough nonzero vectors for the columns of $Y$ to span $\mathbb{R}^n$.

Next notice if $x_i$ denotes the $i$th row of $X$ and $y_i$ denotes the $i$th column of $Y$ then $(XY)_{i,j} = \langle x_i, y_j\rangle$. Therefore $\|XY\|_0$ counts the number of nonzero inner products between the rows of $X$ and the columns of $Y$. Now if $y_j = 0$ for some $j$ then $\langle x_i, y_j\rangle = 0$ for all $i = 1, \ldots, m$. Therefore

to minimize $\|XY\|_0$ subject to $YX = I$ it would seem, heuristically, that one should select matrices $Y$ such that $y_j = 0$ for as many indices $j$ as possible. These are precisely the matrices that correspond to a basis. As such, heuristically, it would seem that a solution to problem $(P_0')$ should correspond to a basis and thus, as described above, can be used to identify a basis from the rows of $X$.

As with problem $(P_0)$, problem $(P_0')$ is difficult to solve since $\|\cdot\|_0$ is not even continuous. Thus, as was done in the compressed sensing literature, problem $(P_0')$ will be relaxed by considering the new problem,

$$\text{minimize} \|XY\|_1 \quad \text{subject to} \quad YX = I \tag{$P_1'$}$$

Problem $(P_1')$, just as problem $(P_1)$, is a convex optimization problem since $\|\cdot\|_1$ is a norm as shown in Proposition (2.2.3). However, the matrix-matrix operations in problem $(P_1')$ as compared to the matrix-vector operations in problem $(P_1)$ makes problem $(P_1')$ fundamentally different from problem $(P_1)$.

As with problem $(P_1)$, the hope is that, under the right conditions, a solution to problem $(P_1')$ is a solution to problem $(P_0')$. In particular, since heuristically problem $(P_0')$ identifies duals that correspond to a basis, the hope is that there exists at least one solution to problem $(P_1')$ that correspond to a basis. In particular, the hope is there exists an extreme point of the set of all minimizers of problem $(P_1')$ that corresponds to a basis. As such, solving problem $(P_1')$ for a particular injective matrix $X \in \mathbb{R}^{m\times n}$ would allow one to identify a subset of the rows of $X$ that is a basis for $\mathbb{R}^n$. This is made precise in the next conjecture.

**Conjecture 6.3.1.** *Given an injective matrix $X$, whose rows are of unit length, each extreme point of*

$$\arg\min_{YX=I} \|XY\|_1$$

*corresponds to a basis.*

In particular, following the work of Donoho and Elad [15], chapter seven of this work will focus on the case where the rows of $X$, in the conjecture above, are the concatenation of two orthonormal bases. For such an $X$, this work will focus on the following conjecture.

**Conjecture 6.3.2.** *Given an injective matrix $X$, whose rows are a concatenation of two orthonormal bases, each left inverse of $X$ that corresponds to an orthonormal basis is an extreme point of*

$$\arg\min_{YX=I} \|XY\|_1$$

Again problem ($P_1'$) has been solved numerically for millions of randomly selected injective matrices. For each such matrix, problem ($P_1'$) always had a solution that corresponded to a basis. However, notice that although problem ($P_1'$) is a convex optimization problem, the function $\|\cdot\|_1$ is not differentiable. As such, perhaps one could modify problem ($P_1'$) further to construct a new problem where the function to minimize is differentiable, but a solution to the new problem identifies a basis. Thus consider the following optimization problem.

$$\text{minimize } \|XY\|_2^2 \quad \text{subject to} \quad YX = I \tag{$P_2'$}$$

where recall from Definition (2.2.1), $\|\cdot\|_2 : \mathbb{R}^{m\times n} \to \mathbb{R}$ is defined as

$$\|X\|_2 = \left(\sum_{j=1}^{n}\sum_{i=1}^{m} X_{i,j}^2\right)^{1/2}$$

Notice the function $\|\cdot\|_2^2$ is differentiable on all of $\mathbb{R}^{m\times n}$. Due to this, problem ($P_2'$) can be more easily solved as compared to problem ($P_1'$). However, as shown in the next result, if $X$ is an injective matrix then problem ($P_2'$) has a unique solution which is the matrix representation of the synthesis operator of the canonical dual of $X$. Moreover, this dual cannot correspond to a basis since, in general, the entries of the canonical dual are all nonzero.

**Proposition 6.3.1.** *Let $X \in \mathbb{R}^{m\times n}$. Then $X^\dagger$ is the unique minimizer of $\|XY\|_2^2$ subject to $YX = I$. In addition, $X^\dagger$ is also the unique minimizer of $\|Y\|_2^2$ subject to $YX = I$.*

**Proof.**

Notice $X^\dagger$ satisfies $X^\dagger X = I$ and thus any $Y$ that satisfies $YX = I$ is of the form $Y = X^\dagger + Z$ where $ZX = 0$. Next $\|\cdot\|_2$ as it is defined in Definition (2.2.1) satisfies the Pythagorean Theorem since it extends the $\ell_2$ vector norm which satisfies the Pythagorean Theorem. Furthermore,

from this extension, the inner product of two matrices $A$ and $B$ is

$$\begin{aligned}\langle A, B\rangle &= \sum_{i,j} A_{i,j}B_{i,j} \\ &= \sum_{i,j} A_{i,j}B^T_{j,i} \\ &= \sum_{i} (AB^T)_{i,i} \\ &= \mathrm{Tr}(AB^T)\end{aligned}$$

Therefore

$$\begin{aligned}\left\langle XX^\dagger, XZ\right\rangle &= \mathrm{Tr}(XX^\dagger(XZ)^T) \\ &= \mathrm{Tr}(XX^\dagger Z^T X^T) \\ &= \mathrm{Tr}(X(X^TX)^{-1}X^TZ^TX^T) \\ &= \mathrm{Tr}(X^TX(X^TX)^{-1}X^TZ^T) \\ &= \mathrm{Tr}(X^TZ^T) \\ &= \mathrm{Tr}((ZX)^T) \\ &= 0\end{aligned}$$

Thus by the Pythagorean Theorem,

$$\left\|X(X^\dagger + Z)\right\|_2^2 = \left\|XX^\dagger + XZ\right\|_2^2 = \left\|XX^\dagger\right\|_2^2 + \|XZ\|_2^2$$

Next $\|XZ\|_2^2 \geq 0$ with $\|XZ\|_2 = 0$ if and only if $XZ = 0$ if and only if $Z = 0$ since $X$ is injective. Using a similar analysis of the second problem establishes the second result.

□

Based on the previous result, the rest of this document will focus on developing conditions for problem ($P_1'$) to identify a basis, since numerical evidence suggests problem ($P_1'$) can always identify a basis, and problems ($\widetilde{P}_1$) and ($P_2'$) have been shown to, at least sometimes, fail to identify a basis.

Now it will turn out that given an injective matrix $X$, the set of minimizers of problem ($P_1'$) is nonempty, compact, and convex. Furthermore, in general, this solution set will contain

infinitely many elements. However, since it is nonempty, compact, and convex, the Krein-Milman Theorem asserts that the solution set is the closed convex hull of its extreme points. Recall the solution set is a set of matrices whose columns are dual frames of the rows of the matrix $X$. The hope is that at least some of the extreme points of the solution set of problem $(P_1')$ are duals that correspond to bases.

The problem, however, is the duals corresponding to which bases that are subsets of the rows of $X$ are minimizers of problem $(P_1')$. Looking at examples of injective matrices quickly show that all duals that correspond to a basis cannot be minimizers. The hope is that if the rows of $X$ contain an orthonormal basis, then the dual corresponding to that basis is a minimizer of problem $(P_1')$.

To that extent, and following the work of Donoho and Elad in [15] who analyzed problem $(P_1)$ for frames that were a concatenation of two orthonormal bases, we will analyze problem $(P_1')$ when $X$ is an injective matrix whose rows are a concatenation of two orthonormal bases. That is, $X$ is of the form (6.2) where $X_1$ and $X_2$ are real orthogonal matrices. In this situation, the goal is to show that the duals that correspond to bases that are the rows of $X_1$ and $X_2$ are solutions of problem of $(P_1')$. That is, the matrices $[X_1^T \,|\, 0]$ and $[0 \,|\, X_2^T]$ are solutions to problem $(P_1')$. Moreover, the goal is to show that these solutions are extreme points of the solution set of problem $(P_1')$. If satisfied, these goals would establish that if a frame is a concatenation of two orthonormal bases, then constructing the matrix $X$ whose rows are the elements of that frame, and identifying the extreme points of the solution set of problem $(P_1')$ for that $X$ would allow one to identify the two orthonormal bases. Note that although this document focuses on the case when $X$ is a concatenation of two orthonormal bases, many of the results below apply to, and are proven for more general situations.

To derive what is needed to analyze the above claims, suppose $X$ is of the form (6.2). It will be shown below that, in this case, the dual that corresponds to the basis consisting of the rows of $X_1$, the matrix $[X_1^{-1} \,|\, 0]$, is a minimizer of problem $(P_1')$ if and only if 0 is a minimizer of $G_{X_2X_1^{-1}}$ as defined below. Furthermore, it will be shown below that permuting the rows of $X$ just permutes the columns of the minimizers of problem $(P_1')$. Therefore, to analyze problem $(P_1')$ it is enough to analyze the function $G_B$ defined next.

**Definition 6.3.1.** *For fixed $B \in \mathbb{R}^{m\times n}$ define $G_B : \mathbb{R}^{n\times m} \to \mathbb{R}$ as*

$$G_B(Y) := \|Y\|_1 + \|YB - I\|_1 + \|BY\|_1 + \|BYB - B\|_1$$

The following definition will be needed to see the importance of the above function.

**Definition 6.3.2.** *If $X \in \mathbb{R}^{m\times n}$ is injective define*

$$\mathcal{D}_X := \left\{Y \in \mathbb{R}^{n\times m} \,:\, YX = I\right\}$$

Again since $X$ is injective its rows $\mathbb{X} = \{x_i\}_{i=1}^m$ form a frame for $\mathbb{R}^n$. Then $\mathcal{D}_X$ is just the set of matrix representations of the synthesis operators of the dual frames to $\mathbb{X}$. Thus since $\mathbb{X}$ is a frame it has at least one dual frame by Proposition (3.3.3). Let $Y$ denote the matrix representation of the synthesis operator of this dual frame. Then $Y \in \mathcal{D}_X$ and therefore $\mathcal{D}_X$ is nonempty.

The following proposition shows how to describe the set $\mathcal{D}_X$ for $X$ of the form (6.2).

**Proposition 6.3.2.** *If*

$$X = \left[\frac{X_1}{X_2}\right]$$

*where $X_1 \in \mathbb{R}^{n\times n}$ is invertible and $X_2 \in \mathbb{R}^{m\times n}$ then*

$$\mathcal{D}_X = \left\{[(I - YX_2)X_1^{-1} \,|\, Y] \,:\, Y \in \mathbb{R}^{n\times m}\right\}$$

**Proof.**

Notice $[Y_1 \,|\, Y] \in \mathcal{D}_X$ where $Y_1 \in \mathbb{R}^{n\times n}$ if and only if

$$I = [Y_1 \,|\, Y]\left[\frac{X_1}{X_2}\right] = Y_1X_1 + YX_2$$

if and only if $Y_1 = (I - YX_2)X_1^{-1}$ and hence

$$\mathcal{D}_X = \left\{[(I - YX_2)X_1^{-1} \,|\, Y] \,:\, Y \in \mathbb{R}^{n\times m}\right\}$$

□

The next result explicitly shows that the objective function in problem ($P_1'$) with $X$ of the form (6.2) can be analyzed in terms of the function $G_B$ in definition (6.3.1) for a particular matrix $B$ derived from $X$.

**Proposition 6.3.3.** *If*

$$X = \left[\frac{X_1}{X_2}\right]$$

*where* $X_1 \in \mathbb{R}^{n\times n}$ *is invertible and* $X_2 \in \mathbb{R}^{m\times n}$ *and* $\widetilde{Y} = [(I - YX_2)X_1^{-1} \mid Y] \in \mathcal{D}_X$ *then*

$$\|X\widetilde{Y}\|_1 = G_{X_2X_1^{-1}}(X_1Y)$$

**Proof.**

For convenience, write $\widetilde{Y} = [Y_1 \mid Y]$ where $Y_1 = (I - YX_2)X_1^{-1}$. Next let $Z = X_1Y$ and $B = X_2X_1^{-1}$. Then

$$\begin{aligned}
\|X\widetilde{Y}\|_1 &= \left\| \left[\frac{X_1}{X_2}\right] [Y_1 \mid Y] \right\|_1 \\
&= \left\| \begin{bmatrix} X_1Y_1 & X_1Y \\ X_2Y_1 & X_2Y \end{bmatrix} \right\| \\
&= \|X_1Y_1\|_1 + \|X_1Y\|_1 + \|X_2Y_1\|_1 + \|X_2Y\|_1 \\
&= \left\|X_1(I - YX_2)X_1^{-1}\right\|_1 + \|X_1Y\|_1 + \left\|X_2(I - YX_2)X_1^{-1}\right\|_1 + \|X_2Y\|_1 \\
&= \left\|X_1X_1^{-1} - X_1YX_2X_1^{-1}\right\|_1 + \|X_1Y\|_1 + \left\|X_2X_1^{-1} - X_2YX_2X_1^{-1}\right\|_1 + \|X_2Y\|_1 \\
&= \left\|I - X_1YX_2X_1^{-1}\right\|_1 + \|X_1Y\|_1 + \left\|X_2X_1^{-1} - X_2YX_2X_1^{-1}\right\|_1 + \|X_2Y\|_1 \\
&= \left\|I - X_1YX_2X_1^{-1}\right\|_1 + \|X_1Y\|_1 + \left\|X_2X_1^{-1} - X_2YX_2X_1^{-1}\right\|_1 + \left\|X_2X_1^{-1}X_1Y\right\|_1 \\
&= \left\|I - (X_1Y)(X_2X_1^{-1})\right\|_1 + \|X_1Y\|_1 + \\
&\qquad \left\|(X_2X_1^{-1}) - (X_2X_1^{-1})(X_1Y)(X_2X_1^{-1})\right\|_1 + \left\|(X_2X_1^{-1})(X_1Y)\right\|_1 \\
&= \|I - ZB\|_1 + \|Z\|_1 + \|B - BZB\|_1 + \|BZ\|_1 \\
&= G_B(Z) \\
&= G_{X_2X_1^{-1}}(X_1Y)
\end{aligned}$$

□

The next results will establish that permuting the rows of $X$ does not fundamentally change the minimizers of problem ($P_1'$). That is, if $Y$ is a minimizer of problem ($P_1'$) with the matrix $X$ and if the rows of $X$ are permuted, then $Y$, with its corresponding columns permuted, is a solution to problem ($P_1'$) with the permuted $X$. Again recall given $A \in \mathbb{R}^{m\times n}$ with

$\sigma : \{1, \ldots, m\} \to \{1, \ldots, m\}$ a permutation, $A_\sigma \in \mathbb{R}^{m \times n}$ is defined as

$$(A_\sigma)_{i,j} = A_{\sigma(i),j}$$

Next if $\psi : \{1, \ldots, n\} \to \{1, \ldots, n\}$ a permutation, $X^\psi \in \mathbb{R}^{m \times n}$ is defined as

$$(A^\psi)_{i,j} = A_{i,\psi(j)}$$

That is, $A_\sigma$ is the matrix formed by permuting the rows of $A$ according to the permutation $\sigma$ and $A^\psi$ is the matrix formed by permuting the columns of $A$ according to the permutation $\psi$. The next result makes the above description of the affects of permuting the rows of $X$ precise.

**Proposition 6.3.4.** *Let $X \in \mathbb{R}^{m \times n}$ and $Y \in \mathbb{R}^{n \times m}$ with $\sigma : \{1, \ldots, m\} \to \{1, \ldots, m\}$ a permutation. Then*

$$Z \in \underset{YX=I}{\arg\min} \|XY\|_1 \qquad \textit{if and only if} \qquad Z^\sigma \in \underset{YX_\sigma=I}{\arg\min} \|X_\sigma Y\|_1$$

**Proof.**

Notice for any $i$ and $j$,

$$(XY)_{\sigma(i),\sigma(j)} = \sum_{k=1}^{n} X_{\sigma(i),k} Y_{k,\sigma(j)} = \sum_{k=1}^{n} (X_\sigma)_{i,k} (Y^\sigma)_{k,j} = (X_\sigma Y^\sigma)_{i,j}$$

Therefore since $\sigma$ is bijective,

$$\|XY\|_1 = \sum_{i,j=1}^{m} |(XY)_{i,j}| = \sum_{i,j=1}^{m} \left|(XY)_{\sigma(i),\sigma(j)}\right| = \sum_{i,j=1}^{m} |(X_\sigma Y^\sigma)_{i,j}| = \|X_\sigma Y^\sigma\|_1$$

Next, for any $i$ and $j$,

$$(YX)_{i,j} = \sum_{k=1}^{m} Y_{i,k} X_{k,j} = \sum_{k=1}^{m} Y_{i,\sigma(k)} X_{\sigma(k),j} = \sum_{k=1}^{m} (Y^\sigma)_{i,k} (X_\sigma)_{k,j} = (Y^\sigma X_\sigma)_{i,j}$$

and therefore $YX = Y^\sigma X_\sigma$. Thus $Z$ is a minimizer of $\|XY\|_1$ subject to $YX = I$ if and only if $\|XZ\|_1 \leq \|XY\|_1$ for all $Y$ such that $YX = I$ if and only if $\|X_\sigma Z^\sigma\|_1 \leq \|X_\sigma Y^\sigma\|_1$ for all $Y$ such that $I = YX = Y^\sigma X_\sigma$ if and only if $Z^\sigma$ is a minimizer of $\|X_\sigma Y\|_1$ subject to $YX_\sigma = I$.

□

Again $X \in \mathbb{R}^{m\times n}$ is injective, then its rows form a frame for $\mathbb{R}^n$ and there exists a subset of the rows of $X$ that are a basis for $\mathbb{R}^n$. Hence exists a permutation $\sigma$ of $\{1,\dots,m\}$ that permutes the rows of $X$ to generate $X_\sigma$ such that

$$X_\sigma = \left[\frac{X_1}{X_2}\right]$$

where $X_1$ is invertible. Therefore to analyze problem ($P_1'$) it is enough to analyze the problem for $X$ of the form (6.2). This is the subject of the next result.

**Proposition 6.3.5.** *If*

$$X = \left[\frac{X_1}{X_2}\right]$$

*where $X_1 \in \mathbb{R}^{n\times n}$ is invertible and $X_2 \in \mathbb{R}^{m\times n}$, then*

$$[X_1^{-1} \mid 0] \in \underset{YX=I}{\arg\min} \|XY\|_1 \qquad \textit{if and only if} \qquad 0 \in \underset{Y\in\mathbb{R}^{n\times m}}{\arg\min} G_{X_2X_1^{-1}}(Y)$$

**Proof.**

Recall from Proposition (6.3.3) if $\widetilde{Y} = [(I - YX_2)X_1^{-1} \mid Y] \in \mathcal{D}_X$ then

$$\|X\widetilde{Y}\|_1 = G_{X_2X_1^{-1}}(X_1Y)$$

Therefore

$$[X_1^{-1} \mid 0] \in \underset{YX=I}{\arg\min} \|XY\|_1 \qquad \text{if and only if} \qquad 0 \in \underset{Y\in\mathbb{R}^{n\times m}}{\arg\min} G_{X_2X_1^{-1}}(Y)$$

□

The above results assumed their corresponding minimization problems had a finite minimum value that was actually attained for some matrix. The following will show that this is necessarily the case. That is for fixed $B \in \mathbb{R}^{m\times n}$, notice $G_B(Y) \geq 0$ for any $Y \in \mathbb{R}^{n\times m}$ by the definition of $G_B$. Thus the set $\{G_B(Y) : Y \in \mathbb{R}^{n\times m}\}$ is bounded below. Hence by the completeness property of the real numbers the infimum

$$\inf_{Y\in\mathbb{R}^{n\times m}} G_B(Y) \tag{6.3}$$

exists. The following result, in fact, states that the infimum is obtained at a matrix in a compact subset of $\mathbb{R}^{n\times m}$.

**Proposition 6.3.6.** *Fix $B \in \mathbb{R}^{m\times n}$. Then for any $\alpha \geq n + \|B\|_1$ the set*

$$\mathcal{A}_B = \left\{Y \in \mathbb{R}^{n\times m} \,:\, \|Y\|_1 \leq \alpha\right\}$$

*is compact and*

$$\inf_{Y\in\mathbb{R}^{n\times m}} G_B(Y) = \min_{Y\in\mathcal{A}_B} G_B(Y)$$

**Proof.**

Notice

$$G_B(0) = \|0\|_1 + \|I_n - 0B\|_1 + \|B0\|_1 + \|B - B0B\|_1 = \|I_n\| + \|B\|_1 = n + \|B\|_1$$

However, if $Y \notin \mathcal{A}_B$ then $Y > \alpha \geq n + \|B\|_1$ and hence

$$G_B(Y) = \|Y\| + \|I_n - YB\|_1 + \|BY\|_1 + \|B - BYB\|_1 \geq \|Y\| > n + \|B\|_1$$

Therefore if $Y \notin \mathcal{A}_B$ then

$$G_B(0) = n + \|B\|_1 < G_B(Y)$$

and thus $Y$ cannot be a minimizer of the function $G_B$. Therefore

$$\inf_{Y\in\mathbb{R}^{n\times m}} G_B(Y) = \inf_{Y\in\mathcal{A}_B} G_B(Y)$$

Next, since $\mathbb{R}^{n\times m}$ is a finite dimensional normed linear space, all norms on the space are equivalent. Thus $\mathcal{A}_B$ is bounded by the definition of $\mathcal{A}_B$. Next, if $\{Z_n\}_{n=1}^{\infty} \subseteq \mathcal{A}_B$ such that $\lim Z_n = Z$ then given $\varepsilon > 0$ there exists $N$ such that $\|Z - Z_n\| < \varepsilon$ for all $n \geq N$. Therefore if $n \geq N$ then $Z_n \in \mathcal{A}_B$ implies

$$\begin{aligned}
\|Z\|_1 &= \|Z - Z_n + Z_n\|_1 \\
&\leq \|Z_n\|_1 + \|Z - Z_n\|_1 \\
&\leq \alpha + \|Z - Z_n\|_1 \\
&< \alpha + \varepsilon
\end{aligned}$$

and since $\varepsilon > 0$ was arbitrary it follows that $\|Z\|_1 \leq \alpha$ and hence $Z \in \mathcal{A}_B$. Therefore $\mathcal{A}_B$ is closed and since it is also bounded it is compact by the Heine-Borel theorem.

Next since $G_B$ is convex on $\mathbb{R}^{n\times m}$ it is continuous on the interior of $\mathbb{R}^{n\times m}$ which is again $\mathbb{R}^{n\times m}$. Therefore it is continuous on $\mathcal{A}_B$ and since $\mathcal{A}_B$ is compact, $G_B$ must attain its minimum there, completing the proof.

□

The result above, in fact, shows that the infimum in (6.3) is attained for some matrix and the set of matrices that attain the infimum is compact and convex. This is the topic of the next result. This result is very important since it shows that the set of matrices that attain the minimum in (6.3) is, by the Krein-Milman Theorem, the closed convex hull of its extreme points. This justifies the work of later sections that will analyze these extreme points.

**Theorem 6.3.7.** *For $B \in \mathbb{R}^{m\times n}$ let*

$$\mathcal{B} = \arg\min_{Y\in\mathbb{R}^{n\times m}} G_B(Y)$$

*Then $\mathcal{B}$ is nonempty, compact, and convex.*

**Proof.**

The previous result asserts $\mathcal{B}$ is nonempty. Next because $G_B$ is convex, the set $\mathcal{B}$ is convex. Furthermore the previous result asserts if $Y \in \mathcal{B}$ then $\|Y\|_1 \leq n + \|B\|_1$. Hence $\mathcal{B}$ is bounded since all norms on $\mathbb{R}^{n\times m}$ are equivalent since $\mathbb{R}^{n\times m}$ is finite-dimensional.

Now let $\alpha$ denote the common value of $G_B$ on $\mathcal{B}$. Then if $\{Y_n\}_{n=1}^{\infty} \subseteq \mathcal{B}$ such that $\lim Y_n = Y$ then since, by Proposition (2.2.6), $G_B$ is continuous on $\mathbb{R}^{n\times m}$

$$\begin{aligned} G_B(Y) &= G_B(\lim_{n\to\infty} Y_n) \\ &= \lim_{n\to\infty} G_B(Y_n) \\ &= \lim_{n\to\infty} \alpha \\ &= \alpha \end{aligned}$$

Hence $Y \in \mathcal{B}$ and therefore $\mathcal{B}$ is closed. Thus, by the Heine-Borel theorem, $\mathcal{B}$ is compact.

□

The previous result asserts for any matrix $B$ there exists at least one matrix $Y$ that is a global minimizer of $G_B$. Proposition (6.3.5) shows that to find necessary and sufficient conditions for the dual $[X_1^{-1} \,|\, 0]$ to be a minimizer of $\|XY\|_1$ subject to $YX = I$ where $X$ is of the form

$$X = \left[\frac{X_1}{X_2}\right]$$

where $X_1 \in \mathbb{R}^{n\times n}$ is invertible, one can equivalently identify necessary and sufficient conditions for the zero matrix to be a minimizer of $G_{X_2X_1^{-1}}$. The following result uses the tools from subdifferential analysis developed in the previous chapter to derive these conditions.

**Theorem 6.3.8.** *If $B \in \mathbb{R}^{m\times n}$ then*

$$0 \in \underset{Y\in\mathbb{R}^{n\times m}}{\arg\min}\, G_B(Y)$$

*if and only if*

$$B^T \operatorname{Sgn}(B) B^T + B^T = \Lambda_1 + \Lambda_2 B^T + B^T \Lambda_3 + B^T \Lambda_4 B^T$$

*for some $\Lambda_1 \in \mathbb{R}^{n\times m}$, $\Lambda_2 \in \mathbb{R}^{n\times n}$, $\Lambda_3 \in \mathbb{R}^{m\times m}$, $\Lambda_4 \in \mathbb{R}^{m\times n}$ such that*

1. $\|\Lambda_1\|_\infty, \|\Lambda_2\|_\infty, \|\Lambda_3\|_\infty, \|\Lambda_4\|_\infty \le 1$
2. $(\Lambda_2)_{i,i} = 0$ *for all* $i = 1, \dots, n$
3. $(\Lambda_4)_{i,j} = 0$ *for all* $i = 1, \dots, m$ *and* $j = 1, \dots, n$ *such that* $B_{i,j} \neq 0$

**Proof.**

Let $f_1, f_2, f_3, f_4 : \mathbb{R}^{n\times m} \to \mathbb{R}$ be defined as $f_1(Y) := \|Y\|_1$, $f_2(Y) := \|I_n - YB\|_1$, $f_3(Y) := \|BY\|_1$, and $f_4(Y) := \|B - BYB\|_1$. Notice then, by its definition, $G_B(Y) = f_1(Y) + f_2(Y) + f_3(Y) + f_4(Y)$. Thus using Theorem 23.8, [25],

$$\partial G_B(0) = \partial f_1(0) + \partial f_2(0) + \partial f_3(0) + \partial f_4(0)$$

Next, by Proposition (5.2.1),

$$\partial f_1(0) = \left\{\Lambda_1 \,:\, \Lambda_1 \in \mathbb{R}^{n\times m} \text{ and } \|\Lambda_1\|_\infty \leq 1\right\}$$

$$\begin{aligned}\partial f_2(0) &= \left\{(\Lambda_2 - \operatorname{Sgn}(I_n))B^T \,:\, \Lambda_2 \in \mathbb{R}^{n\times n} \text{ and } \|\Lambda_2\|_\infty \leq 1 \text{ and } (\Lambda_2)_{i,i} = 0 \text{ for } i = 1,\dots,n\right\} \\ &= \left\{(\Lambda_2 - I_n)B^T \,:\, \Lambda_2 \in \mathbb{R}^{n\times n} \text{ and } \|\Lambda_2\|_\infty \leq 1 \text{ and } (\Lambda_2)_{i,i} = 0 \text{ for } i = 1,\dots,n\right\}\end{aligned}$$

$$\partial f_3(0) = \left\{B^T\Lambda_3 \,:\, \Lambda_3 \in \mathbb{R}^{m\times m} \text{ and } \|\Lambda_3\|_\infty \leq 1\right\}$$

$$\partial f_4(0) = \left\{B^T(\Lambda_4 - \operatorname{Sgn}(B))B^T \,:\, \Lambda_4 \in \mathbb{R}^{m\times n} \text{ and } \|\Lambda_4\|_\infty \leq 1 \text{ and if } B_{i,j} \neq 0 \text{ then } (\Lambda_4)_{i,j} = 0\right\}$$

Further, by Proposition (4.3.6),

$$0 \in \operatorname*{arg\,min}_{Y\in\mathbb{R}^{n\times m}} G_B(Y) \qquad \text{if and only if} \qquad 0 \in \partial G_B(0)$$

Thus, from the above calculations, $0 \in \partial G_B(0)$ if and only if

$$0 = \Lambda_1 + (\Lambda_2 - I_n)B^T + B^T\Lambda_3 + B^T(\Lambda_4 - \operatorname{Sgn}(B))B^T$$

for some $\Lambda_1 \in \mathbb{R}^{n\times m}$, $\Lambda_2 \in \mathbb{R}^{n\times n}$, $\Lambda_3 \in \mathbb{R}^{m\times m}$, $\Lambda_4 \in \mathbb{R}^{m\times n}$ such that

1. $\|\Lambda_1\|_\infty, \|\Lambda_2\|_\infty, \|\Lambda_3\|_\infty, \|\Lambda_4\|_\infty \leq 1$
2. $(\Lambda_2)_{i,i} = 0$ for all $i = 1,\dots,n$
3. $(\Lambda_4)_{i,j} = 0$ for all $i = 1,\dots,m$ and $j = 1,\dots,n$ such that $B_{i,j} \neq 0$

if and only if

$$B^T \operatorname{Sgn}(B)B^T + B^T = \Lambda_1 + \Lambda_2 B^T + B^T\Lambda_3 + B^T\Lambda_4 B^T$$

□

# CHAPTER 7. Special Cases

This chapter is a collection of results that show if an injective matrix $X \in \mathbb{R}^{m\times n}$ has one of several special forms then there exists a solution to problem ($P_1'$) that identifies a subset of the rows of $X \in \mathbb{R}^{n\times n}$ that forms a basis for $\mathbb{R}^n$.

**Theorem 7.0.9.** *Let $b \in \mathbb{R}^n$ such that $\|b\|_\infty \leq 1$. Then zero is a global minimizer of $G_{b^T}$.*

**Proof.**

Recall from Theorem (6.3.8)

$$b\,\mathrm{Sgn}(b)^T b + b = \Lambda_1 + \Lambda_2 b + b\Lambda_3 + b\Lambda_4^T b \tag{7.1}$$

for some $\Lambda_1 \in \mathbb{R}^n$, $\Lambda_2 \in \mathbb{R}^{n\times n}$, $\Lambda_3 \in \mathbb{R}$, $\Lambda_4 \in \mathbb{R}^n$ such that

1. $\|\Lambda_1\|_\infty, \|\Lambda_2\|_\infty, \|\Lambda_3\|_\infty, \|\Lambda_4\|_\infty \leq 1$
2. $(\Lambda_2)_{i,i} = 0$ for all $i = 1, \ldots, n$
3. $(\Lambda_4)_j = 0$ for all $j = 1, \ldots, n$ such that $b_j \neq 0$

Now let

$$\Lambda_2 = \begin{bmatrix} b_1\,\mathrm{sgn}(b_1) & b_1\,\mathrm{sgn}(b_2) & \cdots & b_1\,\mathrm{sgn}(b_n) \\ b_2\,\mathrm{sgn}(b_1) & b_2\,\mathrm{sgn}(b_2) & \cdots & b_2\,\mathrm{sgn}(b_n) \\ \vdots & \vdots & \ddots & \vdots \\ b_n\,\mathrm{sgn}(b_1) & b_n\,\mathrm{sgn}(b_2) & \cdots & b_n\,\mathrm{sgn}(b_n) \end{bmatrix} - \mathrm{diag}(|b_1|, \ldots, |b_n|)$$

Notice $\|\Lambda_2\|_\infty \leq 1$ since $\|b\|_\infty \leq 1$ and $(\Lambda_2)_{i,i} = 0$ for all $i = 1, \ldots, n$. Also

$$\Lambda_2 b = \|b\|_1 b - \mathrm{diag}(|b_1|, \ldots, |b_n|)b$$

Next let $\Lambda_1 = \mathrm{diag}(|b_1|, \ldots, |b_n|)b$ and $\Lambda_3 = 1$ and $\Lambda_4 = 0$. Notice $(\Lambda_1)_i = |b_i|\, b_i$ and therefore $\|\Lambda_1\|_\infty \leq 1$ since $\|b\|_\infty \leq 1$. Thus

$$\Lambda_1 + \Lambda_2 b + b\Lambda_3 + b\Lambda_4^T b = \|b\|_1 b + b$$

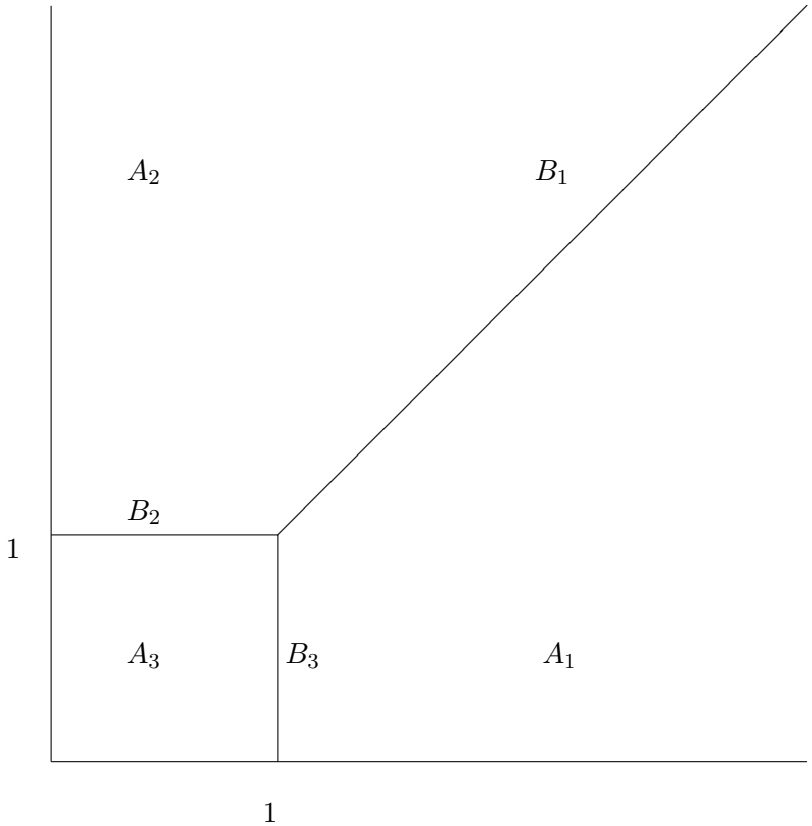


Figure 7.1 The sets defined in Theorem (7.0.11)

Therefore since

$$b\,\mathrm{Sgn}(b)^T b + b = \|b\|_1\, b + b$$

equation (7.1) is satisfied and therefore zero is a global minimizer of $G_{b^T}$.

□

In $\mathbb{R}^2$, the converse of the above proposition is also true.

**Proposition 7.0.10.** *Let $b \in \mathbb{R}^2$. Then zero is a global minimizer of $G_{b^T}$ if and only if $\|b\|_\infty \leq 1$.*

**Proof.**

The reverse direction follows from the previous result. Now suppose $\|b\|_\infty > 1$ and write $b = (b_1 \;\; b_2)^T$. Since $\|b\|_\infty > 1$ either $|b_1| > 1$ or $|b_2| > 1$. Without loss of generality suppose

$|b_1| > 1$. Then by Proposition (6.3.5), zero is a global minimizer of $G_{b^T}$ if and only if

$$Z = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

is a minimizer of $\|XY\|_1$ subject to $YX = I$ where

$$X = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ b_1 & b_2 \end{bmatrix}$$

Now let

$$W = \begin{bmatrix} 0 & -b_2/b_1 & -1/b_1 \\ 0 & 1 & 0 \end{bmatrix}$$

Notice both $ZX = I$ and $WX = I$. Also $\|XZ\|_1 = 2 + |b_1| + |b_2|$ and since

$$WZ = \begin{bmatrix} 0 & -b_2/b_1 & -1/b_1 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix}$$

it follows that $\|XW\|_1 = 2 + \frac{|b_2|}{|b_1|} + \frac{1}{|b_1|}$. Last notice $\|XW\|_1 < \|XZ\|_1$ if and only if $\frac{|b_2|}{|b_1|} + \frac{1}{|b_1|} < |b_1| + |b_2|$ if and only if $|b_2|\,(1 - |b_1|) < |b_1|^2 - 1$ if and only if (since $|b_1| > 1$ by assumption)

$$|b_2| > \frac{|b_1|^2 - 1}{1 - |b_1|} = -(|b_1| + 1)$$

Thus, because the last inequality holds, $Z$ cannot be a minimizer of $\|XY\|_1$ subject to $YX = I$ and therefore zero is not a global minimizer of $G_{b^T}$. Therefore if $\|b\|_\infty > 1$ then zero is not a global minimizer of $G_{b^T}$.

□

The next result is a very nice example of the claim that, given an injective matrix, there exists a solution to problem ($P_1'$) that identifies a basis in the rows of that injective matrix. In particular, given the injective matrix

$$X = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ x_1 & x_2 \end{bmatrix}$$

where $x_1$ and $x_2$ are any real numbers, the set of minimizers of problem ($P_1'$) is either a point, a line segment, or a triangle. In any case, the point, the endpoints of the line segment, or the vertices of the triangle are matrices $Y$ such that $YX = I$ and each $Y$ corresponds to a basis.

**Theorem 7.0.11.** *Let*

$$X = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ x_1 & x_2 \end{bmatrix}$$

*where* $x_1, x_2 \neq 0$ *and define the sets (see Figure (7)),*

$$\begin{aligned} A_1 &= \{(x,y) : x > 1 \text{ and } 0 < y < x\} \\ A_2 &= \{(x,y) : 0 < x < y \text{ and } y > 1\} \\ A_3 &= \{(x,y) : 0 < x < 1 \text{ and } 0 < y < 1\} \\ B_1 &= \{(x,y) : x > 1 \text{ and } x = y\} \\ B_2 &= \{(x,y) : 0 < x < 1 \text{ and } y = 1\} \\ B_3 &= \{(x,y) : x = 1 \text{ and } 0 < y < 1\} \\ C_1 &= \{(1,1)\} \end{aligned}$$

*Now consider the left inverses of* $X$*,*

$$Y_1 = \begin{bmatrix} 0 & -\frac{x_2}{x_1} & \frac{1}{x_1} \\ 0 & 1 & 0 \end{bmatrix}, \quad Y_2 = \begin{bmatrix} 1 & 0 & 0 \\ -\frac{x_1}{x_2} & 0 & \frac{1}{x_2} \end{bmatrix}, \quad \text{and} \quad Y_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

*Then*

$$\underset{YX=I}{\arg\min} \|XY\|_1 = \begin{cases} \{Y_1\} & \text{if } (|x_1|, |x_2|) \in A_1 \\ \{Y_2\} & \text{if } (|x_1|, |x_2|) \in A_2 \\ \{Y_3\} & \text{if } (|x_1|, |x_2|) \in A_3 \\ \operatorname{conv}\{Y_1, Y_2\} & \text{if } (|x_1|, |x_2|) \in B_1 \\ \operatorname{conv}\{Y_2, Y_3\} & \text{if } (|x_1|, |x_2|) \in B_2 \\ \operatorname{conv}\{Y_1, Y_3\} & \text{if } (|x_1|, |x_2|) \in B_3 \\ \operatorname{conv}\{Y_1, Y_2, Y_3\} & \text{if } (|x_1|, |x_2|) \in C_1 \end{cases}$$

**Proof.**

By Proposition (6.3.5), $Y_3$ is a minimizer of $\|XY\|_1$ subject to $YX = I$ if and only if the zero matrix is a minimizer of $G_{(x_1\ x_2)}$. Next, by the previous proposition, zero is a global minimizer of $G_{(x_1\ x_2)}$ if and only if $\max\{|x_1|, |x_2|\} \leq 1$ if and only if $|x_1| \leq 1$ and $|x_2| \leq 1$.

Next, by Proposition (6.3.4), $Y_2$ is a minimizer of $\|XY\|_1$ subject to $YX = I$ if and only if (by permuting the second and third column of $Y_2$)

$$\widetilde{Y}_2 = \begin{bmatrix} 1 & 0 & 0 \\ -\frac{x_1}{x_2} & \frac{1}{x_2} & 0 \end{bmatrix}$$

is a minimizer of $\|\widetilde{X}Y\|_1$ subject to $Y\widetilde{X} = I$ where

$$\widetilde{X} = \begin{bmatrix} 1 & 0 \\ x_1 & x_2 \\ 0 & 1 \end{bmatrix}$$

if and only if, by Proposition (6.3.5), zero is a global minimizer of $G_Z$ where

$$Z = \begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ x_1 & x_2 \end{bmatrix}^{-1} = \begin{bmatrix} -\frac{x_1}{x_2} & \frac{1}{x_2} \end{bmatrix}$$

if and only if, by the previous proposition, $\max\{|x_1|/|x_2|, 1/|x_2|\} \leq 1$ if and only if $|x_1| \leq |x_2|$ and $|x_2| \geq 1$.

Similarly, $Y_1$ is a minimizer of $\|XY\|_1$ subject to $YX = I$ if and only if $\max\{|x_2|/|x_1|, 1/|x_1|\} \leq 1$ if and only if $|x_2| \leq |x_1|$ and $|x_1| \geq 1$.

The result follows from collecting the above results and restating them in terms of the sets $C_1$ and $A_i$ and $B_i$ for $i = 1, 2, 3$.

□

The next result shows if the $\ell_1$ norm of the rows of a matrix $B \in \mathbb{R}^{m\times n}$ is no more than one, then the zero matrix is a global minimizer of $G_B$. That is, by Proposition (6.3.5), $[I \,|\, 0]$ is a minimizer of

$$\left\| \left[ \frac{I}{B} \right] Y \right\|_1$$

subject to

$$Y \left[ \frac{I}{B} \right] = I$$

where $Y \in \mathbb{R}^{n\times(m+n)}$.

**Lemma 7.0.12.** *Let $B \in \mathbb{R}^{m\times n}$ such that $\|b_i\|_1 \leq 1$ for all $i = 1, \ldots, m$ where $b_i$ denotes the $i$th row of $B$. Then zero is a global minimizer of $G_B$.*

**Proof.**

Let $\Lambda_1 = B^T$, $\Lambda_2 = \Lambda_4 = 0$, and $\Lambda_3 = \mathrm{Sgn}(B)B^T$. Now given $i = 1, \ldots, m$ notice

$$\|b_i\| = \sum_{j=1}^{n} |B_{i,j}| \leq 1$$

Hence $|B_{i,j}| \leq 1$ for all $j = 1, \ldots, n$. Therefore since $i$ was arbitrary, $\|\Lambda_1\|_\infty = \left\|B^T\right\|_\infty \leq 1$. Next notice for any $i, j = 1, \ldots, m$,

$$\begin{aligned}
|(\Lambda_3)_{i,j}| &= \left|(\operatorname{Sgn}(B)B^T)_{i,j}\right| \\
&= \left|\sum_{k=1}^{n} \operatorname{Sgn}(B)_{i,k} B^T_{k,j}\right| \\
&= \left|\sum_{k=1}^{n} \operatorname{sgn}(B_{i,k}) B_{j,k}\right| \\
&\leq \sum_{k=1}^{n} |\operatorname{sgn}(B_{i,k})| \cdot |B_{j,k}| \\
&\leq \sum_{k=1}^{n} |B_{j,k}| \\
&= \|b_j\|_1 \\
&\leq 1
\end{aligned}$$

Thus $\|\Lambda_3\|_\infty \leq 1$. Furthermore,

$$\Lambda_1 + \Lambda_2 B^T + B^T \Lambda_3 + B^T \Lambda_4 B^T = B^T + B^T \operatorname{Sgn}(B) B^T$$

Therefore by the previous result, zero is a global minimizer of $G_B$.

□

Now statement (6.1) can be modified to define the *mutual incoherence* between two matrices $X_1 \in \mathbb{R}^{m_1 \times n}$ and $X_2 \in \mathbb{R}^{m_2 \times n}$ as

$$M(X_1, X_2) := \max_{\substack{1 \leq i \leq m_1 \\ 1 \leq j \leq m_2}} |\langle (X_1)_i, (X_2)_j \rangle| = \left\|X_2 X_1^T\right\|_\infty \tag{7.2}$$

where $(X_i)_j$ denotes the $j$ row of matrix $X_i$ for $i = 1, 2$.

The next result shows if $\mathbb{X} = \{x_i\}_{i=1}^m$ is a frame such that $\{x_i\}_{i \in \mathcal{I}}$ is a basis for some index set $\mathcal{I} \subseteq \{1, \ldots, m\}$ then a dual corresponding to this basis a minimizers of problem ($P_1'$) provided

$$\max_{\substack{i \in \mathcal{I} \\ j \notin \mathcal{I}}} |\langle y_i, x_j \rangle| \leq \frac{1}{n}.$$

where $\{y_i\}_{i \in \mathcal{I}}$ is the (unique) dual basis of $\{x_i\}_{i \in \mathcal{I}}$.

That is, if the elements in the dual of the basis are not aligned with any of the other elements of the frame then the matrix $Z$ with columns $z_i$ where $z_i = y_i$ for $i \in \mathcal{I}$ and $z_i = 0$ for $i \notin \mathcal{I}$ is one minimizer of problem ($P_1'$). Note in general problem ($P_1'$) may have infinitely many minimizers.

Specifically, the next result involves the injective matrix

$$X = \left[\frac{X_1}{X_2}\right]$$

where $X_1 \in \mathbb{R}^{n\times n}$ is invertible. The result then involves the analysis of the the mutual incoherence $M((X_1^{-1})^T, X_2)$. To understand this value, let $\mathbb{X} = \{x_i\}_{i=1}^n$ denote the first $n$ rows of $X$. That is, $\mathbb{X}$ consists of all of the rows of $X_1$. Therefore since $X_1$ is invertible, $\mathbb{X}$ is a basis, and hence a frame.

Now let $\mathbb{Y} = \{y_i\}_{i=1}^n$ be the unique dual of $\mathbb{X}$. Since $X_1 X_1^{-1} = I$ notice that because $X_1$ is the analysis operator of $\mathbb{X}$ the matrix $X_1^{-1}$ is the synthesis operator of $\mathbb{Y}$. That is, $y_i$ is the $i$th column of $X_1^{-1}$ for $i = 1, \ldots, n$. Thus $y_i$ is the $i$th row of $(X_1^{-1})^T$ for $i = 1, \ldots, n$. Hence, by definition (7.2) the value $M((X_1^{-1})^T, X_2)$ is the largest absolute value of the inner products of the elements of $\mathbb{Y}$ with the rows of $X_2$.

**Theorem 7.0.13.** *Let $X \in \mathbb{R}^{m\times n}$ be an injective matrix with*

$$X = \left[\frac{X_1}{X_2}\right]$$

*where $X_1 \in \mathbb{R}^{n\times n}$ is an injective matrix. If $M((X_1^{-1})^T, X_2) \leq 1/n$, then*

$$[X_1^{-1} \mid 0] \in \underset{YX=I}{\arg\min} \|XY\|_1$$

**Proof.**

By Proposition (6.3.5)

$$[X_1^{-1} \mid 0] \in \underset{YX=I}{\arg\min} \|XY\|_1 \qquad \text{if and only if} \qquad 0 \in \underset{Y\in\mathbb{R}^{n\times(m-n)}}{\arg\min} G_B(Y)$$

where $B = X_2 X_1^{-1} \in \mathbb{R}^{(m-n)\times n}$. Notice if

$$M((X_1^{-1})^T, X_2) = \left\|X_2((X_1^{-1})^T)^T\right\|_\infty = \left\|X_2 X_1^{-1}\right\|_\infty \leq \frac{1}{n}$$

then $|b_{i,j}| \leq 1/n$ for all $i = 1, \ldots, m-n$ and $j = 1, \ldots, n$. Next for any $i = 1, \ldots, m-n$, if $b_i$ denotes the $i$th row of $B$ then

$$\|b_i\|_1 = \sum_{j=1}^{n} |b_{i,j}| \leq n\left(\frac{1}{n}\right) = 1$$

therefore by the previous proposition, zero is a global minimizer of $G_B$ completing the proof.

□

The last result deals with a particular class of real orthogonal matrices, the Hadamard matrices. That is, a matrix $H \in \mathbb{R}^{n\times n}$ is a *Hadamard matrix* if $H_{i,j} = \pm 1$ for all $i, j = 1, \ldots, n$ and $HH^T = nI$.

**Theorem 7.0.14.** *Let $X \in \mathbb{R}^{m\times n}$ be an injective matrix with*

$$X = \left[\frac{X_1}{X_2}\right]$$

*where $X_1 \in \mathbb{R}^{n\times n}$ is an injective matrix. Next let $B = X_2 X_1^{-1}$. If*

$$B = \frac{1}{\sqrt{n}} H$$

*where $H \in \mathbb{R}^{n\times n}$ is a Hadamard matrix then*

$$[X_1^{-1} \mid 0] \in \underset{YX=I}{\arg\min} \|XY\|_1$$

**Proof.**

By Proposition (6.3.5)

$$[X_1^{-1} \mid 0] \in \underset{YX=I}{\arg\min} \|XY\|_1 \qquad \text{if and only if} \qquad 0 \in \underset{Y\in\mathbb{R}^{n\times(m-n)}}{\arg\min} G_B(Y)$$

Next by Proposition (6.3.8) the zero matrix is a global minimizer of $G_B$ if and only if

$$B^T \operatorname{Sgn}(B) B^T + B^T = \Lambda_1 + \Lambda_2 B^T + B^T \Lambda_3 + B^T \Lambda_4 B^T \tag{7.3}$$

for some $\Lambda_1 \in \mathbb{R}^{n\times n}$, $\Lambda_2 \in \mathbb{R}^{n\times n}$, $\Lambda_3 \in \mathbb{R}^{n\times n}$, $\Lambda_4 \in \mathbb{R}^{n\times n}$ such that

1. $\|\Lambda_1\|_\infty, \|\Lambda_2\|_\infty, \|\Lambda_3\|_\infty, \|\Lambda_4\|_\infty \leq 1$
2. $(\Lambda_2)_{i,i} = 0$ for all $i = 1, \ldots, n$

3. $(\Lambda_4)_{i,j} = 0$ for all $i, j = 1, \ldots, n$ such that $B_{i,j} \neq 0$

Now notice $\operatorname{Sgn}(B) = H = \sqrt{n}B$. Therefore

$$B^T \operatorname{Sgn}(B) B^T + B^T = B^T(\sqrt{n}B)B^T + B^T = (1 + \sqrt{n})B^T$$

Thus set $\Lambda_1 = \sqrt{n}B^T$, $\Lambda_3 = I$, and $\Lambda_2 = \Lambda_4 = 0$, and notice these choices satisfy system (7.3). Next, notice

$$\left\|\sqrt{n}B^T\right\|_\infty = \left\|\sqrt{n}\left(\frac{1}{\sqrt{n}}H\right)\right\|_\infty = \|H\|_\infty = 1$$

Thus $\Lambda_i$ for $i = 1, \ldots, 4$ satisfy all the conditions imposed by Proposition (6.3.8) and therefore the zero matrix is a global minimizer of $G_B$.

□

# CHAPTER 8. The Real Orthogonal Case

Notice if $X \in \mathbb{R}^{2n\times n}$ is an injective matrix whose rows are a concatenation of two orthonormal bases then by Proposition (6.3.4) one can permute the rows of $X$ without fundamentally affecting the solution to problem ($P_1$). Therefore one can assume $X$ is of the form

$$X = \left[\frac{X_1}{X_2}\right] \tag{8.1}$$

where $X_1, X_2 \in \mathbb{R}^{n\times n}$ are real orthogonal matrices. Notice then for $X$ of the above form,

$$\left\| \left[\frac{X_1}{X_2}\right] [X_1^T \mid 0] \right\|_1 = \left\| \begin{bmatrix} I & 0 \\ X_2X_1^T & 0 \end{bmatrix} \right\|_1 = n + \left\| X_2X_1^T \right\|_1$$

and since $\left\|X_1X_2^T\right\|_1 = \left\|(X_1X_2^T)^T\right\|_1 = \left\|X_2X_1^T\right\|_1$ it follows that

$$\left\| \left[\frac{X_1}{X_2}\right] [0 \mid X_2^T] \right\|_1 = \left\| \begin{bmatrix} 0 & X_1X_2^T \\ 0 & I \end{bmatrix} \right\|_1 = n + \left\| X_2X_1^T \right\|_1$$

Next notice $[X_1^T \mid 0], [0 \mid X_2^T] \in \mathcal{D}_X$ and by the above calculations

$$[X_1^T \mid 0] \in \underset{YX=I}{\arg\min} \|XY\|_1 \qquad \text{if and only if} \qquad [0 \mid X_2^T] \in \underset{YX=I}{\arg\min} \|XY\|_1$$

Therefore if the solution to problem ($P_1$) identifies the orthonormal basis $X_1$ if and only if it identifies the orthonormal basis $X_2$. Next by Proposition (6.3.5)

$$[X_1^T \mid 0] \in \underset{YX=I}{\arg\min} \|XY\|_1 \qquad \text{if and only if} \qquad 0 \in \underset{Y\in\mathbb{R}^{n\times n}}{\arg\min} G_{X_2X_1^T}(Y)$$

Now let $B = X_2X_1^T$. Then the above work shows

$$[X_1^T \mid 0] \in \underset{YX=I}{\arg\min} \|XY\|_1 \qquad \text{if and only if} \qquad 0 \in \underset{Y\in\mathbb{R}^{n\times n}}{\arg\min} G_B(Y)$$

Notice $B$ is real orthogonal since $X_1$ and $X_2$ are real orthogonal. Thus in order for $[X_1^T \mid 0]$ and $[0 \mid X_2^T]$ to be solutions to problem ($P_1$) for $X$ of the form (8.1) for any real orthogonal

matrices $X_1$ and $X_2$, it must be that zero matrix is a global minimizer of $G_B$ for any real orthogonal matrix $B$.

Thus the work in this chapter will address, if $B$ is a real orthogonal matrix, then when is the zero matrix necessarily a global minimizer of $G_B$. The first section of this chapter shows if zero is, in fact, a global minimizer of $G_B$ then it is an extreme point of the set of global minimizers. The second section establishes that it is enough to consider the case when $B$ is a real orthogonal matrix such that $B_{i,j} \neq 0$ for all $i$ and $j$. The third section develops optimality conditions that guarantee the zero matrix is a global minimizer of $G_B$. The fourth section establishes conditions a real orthogonal matrix $B$ should satisfy provided it is a global minimizer of $G_B$. If one could find a real orthogonal matrix $A$ that does not satisfy these conditions, it is possible a that the zero matrix is not a global minimizer of $G_A$. The fifth section, however, establishes that all real orthogonal matrices satisfy the conditions derived in section three.

## 8.1 Results on Extreme Points of the Solution Set

Recall if $B \in \mathbb{R}^{m\times n}$ is any matrix then $G_B$ is defined as

$$G_B(Y) = \|Y\|_1 + \|I - YB\|_1 + \|BY\|_1 + \|B - BYB\|_1$$

Therefore if $B \in \mathbb{R}^{n\times n}$ is a real orthogonal matrix then the fact that $G_B(0) = n + \|B\|_1$ and

$$\begin{aligned} G_B(B^T) &= \left\|B^T\right\|_1 + \left\|I - (B^T)B\right\|_1 + \left\|BB^T\right\|_1 + \left\|B - B(B^T)B\right\| \\ &= \|I\|_1 + \left\|B^T\right\|_1 \\ &= n + \|B\|_1 \end{aligned}$$

shows that the zero matrix is a global minimizer of $G_B$ if and only if $B^T$ is a global minimizer.

This section will focus on showing that if the zero matrix and $B^T$ are global minimizers of $G_B$ for some real orthogonal matrix $B$ then both the zero matrix and $B^T$ are in fact extreme points of the set of global minimizers of $G_B$. The following result shows that if a point is the unique maximizer of the $\ell_1$ norm over a convex set then it is an extreme point of that set. Uniqueness is essential in the proposition. That is, consider the $\ell_1$ ball in $\mathbb{R}^2$

$$\mathcal{B}_1 = \left\{x \in \mathbb{R}^2 \,:\, \|x\|_1 \leq 1\right\}$$

Then any $x \in \mathcal{B}_1$ such that $\|x\|_1 = 1$ is a maximizer of $\|\cdot\|_1$ over $\mathcal{B}_1$. However, only $\pm(1\ 0)^T$ and $\pm(0\ 1)^T$ are extreme points of $\mathcal{B}_1$.

**Proposition 8.1.1.** *Consider a nonempty convex set $\mathcal{A} \subseteq \mathbb{R}^n$. If*

$$\arg\max_{x \in \mathcal{A}} \|x\|_1 = \{z\}$$

*then $z$ is an extreme point of $\mathcal{A}$.*

**Proof.**

Suppose, to reach a contradiction, that $z$ is not an extreme point of $\mathcal{A}$. Then there exists $\lambda \in (0,1)$ and $x, y \in \mathcal{A}$ with $x \neq z$ and $y \neq z$ such that $z = \lambda x + (1-\lambda)y$. Then because $x, y \in \mathcal{A}$ and $z$ is the unique maximizer of $\|\cdot\|_1$ on $\mathcal{A}$ it follows that $\|x\|_1, \|y\|_1 < \|z\|_1$. Therefore because $\|\cdot\|_1$ is a norm,

$$\begin{aligned}
\|z\|_1 &= \|\lambda x + (1-\lambda)y\|_1 \\
&\leq \|\lambda x\|_1 + \|(1-\lambda)y\|_1 \\
&= \lambda \|x\|_1 + (1-\lambda)\|y\|_1 \\
&< \lambda \|z\|_1 + (1-\lambda)\|z\|_1 \\
&= \|z\|_1
\end{aligned}$$

a contradiction. Thus $z$ is an extreme point of $\mathcal{A}$.

□

The next result shows that translating a convex set affects the extreme points of the set just as one would expect. That is, a point of the translated set is an extreme point of that set if and only if it is a translate of an extreme point of the original set.

**Proposition 8.1.2.** *Let $\mathcal{A} \subseteq \mathbb{R}^n$ be a convex set and $x \in \mathcal{A}$ an extreme point of $\mathcal{A}$. Then for any $y \in \mathbb{R}^n$ the point $x + y$ is an extreme point of $\mathcal{A} + y = \{a + y : a \in \mathcal{A}\}$.*

**Proof.**

Suppose $x+y$ is not an extreme point of $\mathcal{A}+y$. Then there exists $b_1, b_2 \in \mathcal{A}+y$ with $b_1, b_2 \neq x+y$

and $\lambda \in (0,1)$ such that $x + y = \lambda b_1 + (1-\lambda)b_2$. Next, since $b_i \in \mathcal{A} + y$ for $i = 1, 2$ there exist $a_1, a_2 \in \mathcal{A}$ such that $b_i = a_i + y$ for $i = 1, 2$. Then

$$x + y = \lambda b_1 + (1-\lambda)b_2$$

□

The fact that there is a unique minimizer to the function $g$ in the next proposition is essential to show that the zero matrix and $B^T$, for a real orthogonal matrix $B$, are extreme points of the set of global minimizers of $G_B$ provided they are themselves global minimizers of $G_B$.

**Proposition 8.1.3.** *Let $B \in \mathbb{R}^{n\times n}$ be a real orthogonal matrix and define*

$$g(Y) := \|BY\|_1 + \|I - YB\|_1 + \|B - BYB\|_1$$

*Then*

$$\arg\min_{Y\in\mathbb{R}^{n\times n}} g(Y) = \left\{B^T\right\}$$

**Proof.**

Define

$$\begin{aligned} f(Y) &:= g(Y + B^T) \\ &= \left\|B(Y + B^T)\right\|_1 + \left\|I - (Y + B^T)B\right\|_1 + \left\|B - B(Y + B^T)B\right\|_1 \\ &= \|BY + I\|_1 + \|I - (YB + I)\|_1 + \|B - (BYB + B)\|_1 \\ &= \|BY + I\|_1 + \|YB\|_1 + \|BYB\|_1 \end{aligned}$$

Notice $f$ is convex on $\mathbb{R}^{n\times n}$ and thus by Proposition (5.2.1)

$$\begin{aligned} \partial f(0) &= \left\{B^T(\Lambda_1 + \mathrm{Sgn}(I)) + \Lambda_2 B^T + B^T\Lambda_3 B^T \,:\, \|\Lambda_i\|_\infty \le 1 \text{ for } i = 1,2,3 \text{ and } (\Lambda_1)_{i,i} = 0 \text{ for all } i\right\} \\ &= \left\{B^T(\Lambda_1 + I) + \Lambda_2 B^T + B^T\Lambda_3 B^T \,:\, \|\Lambda_i\|_\infty \le 1 \text{ for } i = 1,2,3 \text{ and } (\Lambda_1)_{i,i} = 0 \text{ for all } i\right\} \end{aligned}$$

Now setting $\Lambda_1 = 0$, $\Lambda_2 = -I$, and $\Lambda_3 = 0$ notice

$$B^T(\Lambda_1 + I) + \Lambda_2 B^T + B^T\Lambda_3 B^T = 0$$

Therefore since $(\Lambda_1)_{i,i} = 0$ for all $i$ and $\|\Lambda_i\|_\infty \leq 1$ for $i = 1, 2, 3$ it has been established that $0 \in \partial f(0)$. Therefore by Proposition (8.1.2),

$$\begin{aligned}\min_{Y \in \mathbb{R}^{n\times n}} g(Y) &= \min_{Y \in \mathbb{R}^{n\times n}} g(Y + B^T) \\ &= \min_{Y \in \mathbb{R}^{n\times n}} f(Y) \\ &= f(0)\end{aligned}$$

In particular, $Y = 0$ is a global minimizer of $f(\cdot)$. That is, $f(Y) \geq f(0)$ for all $Y \in \mathbb{R}^{n\times n}$. Now for $0 < t < 1$

$$\begin{aligned}h_Y(t) &:= f(tY) \\ &= \|B(tY) + I\|_1 + \|(tY)B\|_1 + \|B(tY)B\|_1 \\ &= \|tBY + I\|_1 + t\|YB\|_1 + t\|BYB\|_1 \\ &= \sum_{i\neq j} |t(BY)_{i,j}| + \sum_i |t(BY)_{i,i} + 1| + t\|YB\|_1 + t\|BYB\|_1 \\ &= t\sum_{i\neq j} |(BY)_{i,j}| + \sum_i |t(BY)_{i,i} + 1| + t\|YB\|_1 + t\|BYB\|_1\end{aligned}$$

Now let

$$\begin{aligned}\mathcal{O}_1 &= \left\{Y \in \mathbb{R}^{n\times n} \,:\, \|BY\|_\infty < 1\right\} \\ \mathcal{O}_2 &= \left\{Y \in \mathbb{R}^{n\times n} \,:\, \|YB\|_\infty < 1\right\}\end{aligned}$$

Notice $0 \in \mathcal{O}_1$ and $0 \in \mathcal{O}_2$ and $\mathcal{O}_1$ and $\mathcal{O}_2$ are open. Thus setting $\mathcal{O} = \mathcal{O}_1 \cap \mathcal{O}_2$, notice $\mathcal{O}$ is open and contains zero. Then for all $Y \in \mathcal{O}$

$$\begin{aligned}h_Y(t) &= t\sum_{i\neq j} |(BY)_{i,j}| + \sum_i |t(BY)_{i,i} + 1| + t\|YB\|_1 + t\|BYB\|_1 \\ &= t\sum_{i\neq j} |(BY)_{i,j}| + \sum_i (1 + t(BY)_{i,i}) + t\|YB\|_1 + t\|BYB\|_1\end{aligned}$$

Therefore $h_Y$ is differentiable $t \in (0, 1)$ and any $Y \in \mathcal{O}$. Next if Tr denotes the trace of a square

matrix where for any $X \in \mathbb{R}^{n\times n}$ then

$$\begin{aligned}
h'_Y(t) &= \sum_{i\neq j} |(BY)_{i,j}| + \sum_i (BY)_{i,i} + \|YB\|_1 + \|BYB\|_1 \\
&= \sum_{i\neq j} |(BY)_{i,j}| + \mathrm{Tr}(BY) + \|YB\|_1 + \|BYB\|_1 \\
&= \sum_{i\neq j} |(BY)_{i,j}| + \mathrm{Tr}(YB) + \|YB\|_1 + \|BYB\|_1 \\
&= \sum_{i\neq j} |(BY)_{i,j}| + \sum_i (YB)_{i,i} + \sum_i |(YB)_{i,i}| + \sum_{i\neq j} |(YB)_{i,j}| + \|BYB\|_1 \\
&= \sum_{i\neq j} |(BY)_{i,j}| + \sum_i (|(YB)_{i,i}| - (YB)_{i,i}) + \sum_{i\neq j} |(YB)_{i,j}| + \|BYB\|_1
\end{aligned}$$

Then $\|YB\|_\infty < 1$ implies $|(YB)_{i,i}| \leq 1$ for each $i$ and thus $|(YB)_{i,i}| - (YB)_{i,i} \geq 0$ for all $i$. Therefore because $|(BY)_{i,j}|, |(YB)_{i,j}| \geq 0$ for all $i$ and $j$ and $\|BYB\|_1 \geq 0$ it follows that $h'_Y(t) \geq 0$. Next, since each term in the sum above is non-negative $h'_Y(t) = 0$ if and only if $\|BYB\|_1 = 0$ if and only if $BYB = 0$ if and only if $Y = 0$ since $B$ is invertible. Therefore $h'_Y(t) > 0$ for every $t \in (0,1)$ and every $Y \in \mathcal{O}$ with $Y \neq 0$. Therefore by Proposition (4.3.4),

$$\arg\min_{Y\in\mathbb{R}^{n\times n}} f(Y) = \{0\}$$

Hence by Proposition (8.1.2),

$$\arg\min_{Y\in\mathbb{R}^{n\times n}} g(Y) = \left\{B^T\right\}$$

□

Notice $G_B = \|Y\|_1 + g(Y)$ with $g$ as defined above. However, although $\|Y\|_1$ has the zero matrix as its unique global minimizer, and $g$ has $B^T$ as its unique global minimizer provided $B$ is a real orthogonal matrix, this is unfortunately not enough to ensure the zero matrix and $B^T$ are global minimizers of $G_B$.

**Proposition 8.1.4.** *Let $B \in \mathbb{R}^{n\times n}$ be a real orthogonal matrix. Then*

$$\arg\min_{Y\in\mathbb{R}^{n\times n}} G_B(Y) \subseteq \left\{Y \in \mathbb{R}^{n\times n} \,:\, \|Y\|_1 \leq \|B\|_1\right\}$$

**Proof.**

Define $g_1 : \mathbb{R}^{n\times n} \to \mathbb{R}$ and $g_2 : \mathbb{R}^{n\times n} \to \mathbb{R}$ by

$$\begin{aligned} g_1(Y) &:= \|Y\|_1 \\ g_2(Y) &:= \|BY\|_1 + \|I - YB\|_1 + \|B - BYB\|_1 \end{aligned}$$

Then by Proposition (8.1.3), $g_2(Y) \geq g_2(B^T) = \|I\|_1 = n$ for all $Y \in \mathbb{R}^{n\times n}$. Therefore for any $Y \in \mathbb{R}^{n\times n}$ with $\|Y\|_1 > \|B\|_1$ direct calculation reveals

$$\begin{aligned} G_B(Y) &= g_1(Y) + g_2(Y) \\ &= \|Y\|_1 + g_2(Y) \\ &> \|B\|_1 + g_2(Y) \qquad \text{since } \|Y\|_1 > \|B\|_1 \\ &\geq \|B\|_1 + n \qquad \text{since } g_2(Y) \geq n \text{ for all } Y \\ &= G_B(0) \end{aligned}$$

Therefore $Y$ cannot be a minimizer of $G_B(\cdot)$, proving the result.

□

The following is essential to prove the main result of this section. It shows that when the zero matrix and $B^T$ are global minimizers for $G_B$, where $B$ is a real orthogonal matrix, then they are in fact extreme points of the set of global minimizers.

**Proposition 8.1.5.** *Let $B \in \mathbb{R}^{n\times n}$ be a real orthogonal matrix and let*

$$\mathcal{A} = \arg\min_{Y\in\mathbb{R}^{n\times n}} G_B(Y)$$

*If $0 \in \mathcal{A}$ or $B^T \in \mathcal{A}$ then both $0 \in \mathcal{A}$ and $B^T \in \mathcal{A}$ and, in fact, both $0$ and $B^T$ are extreme points of $\mathcal{A}$.*

**Proof.**

First $G_B(0) = n + \|B\|_1 = G_B(B^T)$. Therefore $0 \in \mathcal{A}$ if and only if $B^T \in \mathcal{A}$. Next, by Proposition (8.1.4),

$$\mathcal{A} \subseteq \left\{Y \in \mathbb{R}^{n\times n} \,:\, \|Y\|_1 \leq \|B\|_1\right\}$$

Therefore $\|Y\|_1 \leq \|B\|_1$ for all $Y \in \mathcal{A}$. Now suppose $Y \in \mathcal{A}$ with $\|Y\|_1 = \|B\|_1$. It will be show that, in this case, $Y$ must be $B^T$. To do so, notice by assumption $0 \in \mathcal{A}$ and thus $G_B(0) = n + \|B\|_1$ is the minimum value of $G_B$ over the set of all $n \times n$ matrices. Therefore if $Y \in \mathcal{A}$ and $\|Y\|_1 = \|B\|_1$ then

$$\begin{aligned} n + \|B\|_1 &= G_B(Y) \\ &= \|Y\|_1 + \|I - YB\|_1 + \|BY\|_1 + \|B - BYB\|_1 \\ &= \|B\|_1 + \|I - YB\|_1 + \|BY\|_1 + \|B - BYB\|_1 \end{aligned}$$

Therefore

$$\|I - YB\|_1 + \|BY\|_1 + \|B - BYB\|_1 = n$$

However, it was shown in Proposition (8.1.3) that

$$\|I - ZB\|_1 + \|BZ\|_1 + \|B - BZB\|_1 \geq n$$

for all $Z \in \mathbb{R}^{n \times n}$ with equality if and only if $Z = B^T$. Therefore if $Y \in \mathcal{A}$ and $\|Y\|_1 = \|B\|_1$ then $Y = B^T$. Therefore $\|Y\|_1 \leq \|B\|_1$ for all $Y \in \mathcal{A}$ and $B^T \in \mathcal{A}$ being the unique $W \in \mathcal{A}$ with $\|W\|_1 = \|B\|_1$ establishes that

$$\arg\max_{Y \in \mathcal{A}} \|Y\|_1 = \{B^T\}$$

Therefore by Proposition (8.1.1), $B^T$ is an extreme point of $\mathcal{A}$.

Next let

$$\begin{aligned} \mathcal{B} &= \arg\min_{Y \in \mathbb{R}^{n \times n}} G_B(Y + B^T) \\ &= \arg\min_{Z - B^T \in \mathbb{R}^{n \times n}} G_B(Z) \\ &= \arg\min_{Z \in \mathbb{R}^{n \times n}} G_B(Z) \end{aligned}$$

Now, by assumption, $Z = 0$ is a global minimizer of $G_B$. Therefore, as shown above, $B^T$ is an extreme point of $\mathcal{B}$. However, notice $\mathcal{B} = \mathcal{A} + B^T$ and hence $\mathcal{A} = \mathcal{B} - B^T$. Thus, by Proposition (8.1.2), $B^T - B^T = 0$ is an extreme point of $\mathcal{A}$.

□

The following is the main result of this section. Applying Proposition (6.3.4) to reorder rows if necessary, it states if the rows of an injective matrix $X$ are a concatenation of two orthonormal bases, and there exists a left inverse of $X$ that corresponds to an orthonormal basis that is a minimizer of problem ($P_1'$) then, in fact, all left inverses of $X$ that correspond to an orthonormal basis are extreme points of the set of minimizers of problem ($P_1'$).

**Theorem 8.1.6.** *Let*

$$X = \left[\frac{X_1}{X_2}\right]$$

*where $X_1, X_2 \in \mathbb{R}^{n\times n}$ are real orthogonal matrices. Next let*

$$\mathcal{A} = \underset{YX=I}{\arg\min} \|XY\|_1$$

*If either $[X_1^T \mid 0] \in \mathcal{A}$ or $[0 \mid X_2^T] \in \mathcal{A}$ then both $[X_1^T \mid 0] \in \mathcal{A}$ and $[0 \mid X_2]^T \in \mathcal{A}$ and, in fact, $[X_1^T \mid 0]$ and $[0 \mid X_2]^T$ are both extreme points of $\mathcal{A}$.*

**Proof.**

The result follows directly from the previous proposition along with Propositions (6.3.5) and (6.3.2). That is, suppose $[X_1^T \mid 0] \in \mathcal{A}$. Then by Proposition (6.3.5) the zero matrix is a global minimizer of $G_B$ where $B = X_2X_1^T$. Next, by the previous proposition, if the zero matrix is a global minimizer of $G_B$ then the zero matrix is an extreme point of

$$\mathcal{B} = \underset{Y\in\mathbb{R}^{n\times n}}{\arg\min} G_B(Y)$$

Next by Proposition (6.3.2), each $Y \in \mathbb{R}^{n\times n}$ corresponds to the matrix $Z = [(I - YX_2)X_1^T \mid 0]$ that satisfies $ZX = I$. Now, to reach a contradiction, suppose there exists matrices $Z_1, Z_2 \in \mathcal{A}$ and $0 < \lambda_1, \lambda_2 < 1$ with $\lambda_1 + \lambda_2 = 1$ such that $[X_1^T \mid 0] = \lambda_1 Z_1 + \lambda_2 Z_2$. Then by Proposition (6.3.2), let $Y_1, Y_2 \in \mathbb{R}^{n\times n}$ be such that $Z_i = [(I - Y_iX_2)X_1^T \mid 0]$ for $i = 1, 2$. Then,

$$\begin{aligned}
[X_1^T \mid 0] &= \lambda_1 Z_1 + \lambda_2 Z_2 \\
&= \lambda_1[(I - Y_1X_2)X_1^T \mid 0] + \lambda_2[(I - Y_2X_2)X_1^T \mid 0] \\
&= [(I - (\lambda_1 Y_1 + \lambda_2 Y_2)X_2)X_1^T \mid 0]
\end{aligned}$$

Therefore $X_1^T = (I-(\lambda_1Y_1+\lambda_2Y_2)X_2)X_1^T$ and hence $\lambda_1Y_1+\lambda_2Y_2 = 0$. However, this contradicts the fact that the zero matrix is an extreme point of $\mathcal{B}$. Therefore $[X_1^T \mid 0]$ is an extreme point.

A similar application of the previous proposition and Propositions (6.3.5) and (6.3.2) shows that $[0 \mid X_2^T]$ is also an extreme point of $\mathcal{A}$. Last, a direct calculation shows

$$\left\| \begin{bmatrix} X_1 \\ \hline X_2 \end{bmatrix} [X_1^T \mid 0] \right\|_1 = n + \left\| X_2 X_1^T \right\| = \left\| \begin{bmatrix} X_1 \\ \hline X_2 \end{bmatrix} [0 \mid X_2^T] \right\|_1$$

Therefore $[X_1^T \mid 0] \in \mathcal{A}$ if and only if $[0 \mid X_2^T] \in \mathcal{A}$.

□

## 8.2 Results on Real Orthogonal Matrices That Are Nowhere Zero

The results of this section will show that the zero matrix is a global minimizer of $G_B$ for any real orthogonal matrix $B$ if and only if the zero matrix is a global minimizer of $G_A$ for any real orthogonal matrix $A$ such that $A_{i,j} \neq 0$ for all $i$ and $j$. The assertion that $A$ is nowhere zero will be very useful when using subdifferential analysis to analyze the function $G_A$ in later chapters. First, a definition will be needed.

**Definition 8.2.1.** *Define* $\mathcal{M} : \mathbb{R}^{m \times n} \to \mathbb{R}$ *as*

$$\mathcal{M}(B) := \min_{Y \in \mathbb{R}^{n \times m}} G_B(Y)$$

The following result follows from Proposition (5.1.5) and is very useful as it implies if $B$ is a matrix and $\{B_n\}_{n=1}^{\infty}$ is a sequence of matrices converging to $B$ then

$$\min_{Y \in \mathbb{R}^{n \times m}} G_B(Y) = \mathcal{M}(B) = \mathcal{M}(\lim_{n \to \infty} B_n) = \lim_{n \to \infty} \mathcal{M}(B_n) = \lim_{n \to \infty} \min_{Y \in \mathbb{R}^{n \times m}} G_{B_n}(Y)$$

Thus the global minimum of $G_B$ can be calculated by calculating the global minimum of $G_{B_n}$ for $B_n$ arbitrarily close to $B$.

**Proposition 8.2.1.** *The function* $\mathcal{M}$ *is continuous on* $\mathbb{R}^{m \times n}$.

**Proof.**

Given $X \in \mathbb{R}^{m \times n}$ define the set

$$\mathscr{C}_X := \left\{ C \in \mathbb{R}^{m \times n} \, : \, \|C\|_1 \leq \|X\|_1 \right\}$$

Notice $X \in \mathscr{C}_X$ and, by construction, $\mathscr{C}_X$ is bounded. Furthermore if $\{C_n\}_{n=1}^{\infty} \subseteq \mathscr{C}_X$ such that $\lim C_n = C$ then given $\varepsilon > 0$ there exists $N$ such that $\|C - C_n\|_1 < \varepsilon$ if $n \geq N$. Then for $n \geq N$,

$$\begin{aligned} \|C\|_1 &= \|C - C_n + C_n\|_1 \\ &\leq \|C_n\|_1 + \|C - C_n\|_1 \\ &\leq \|X\|_1 + \varepsilon \end{aligned}$$

Therefore since $\varepsilon > 0$ is arbitrary $\|C\|_1 \leq \|X\|_1$ and thus $C \in \mathscr{C}_X$. Therefore given any $X \in \mathbb{R}^{m\times n}$ the set $\mathscr{C}_X$ is closed and bounded, and hence compact by the Heine-Borel Theorem.

Next given a matrix $X \in \mathbb{R}^{m\times n}$ define the set

$$\mathscr{A}_X := \left\{Y \in \mathbb{R}^{n\times m} \,:\, \|Y\|_1 \leq n + \|X\|_1\right\}$$

Notice that given $X \in \mathbb{R}^{m\times n}$, if $Z \notin \mathscr{A}_X$ then $Z$ cannot be a global minimizer of $G_X$ since $G_X(0) = n + \|X\|_1$ while $Z \notin \mathscr{A}_X$ implies $\|Z\|_1 > n + \|X\|_1$. Hence

$$G_X(Z) = \|Z\|_1 + \|I - ZB\|_1 + \|BZ\|_1 + \|B - BZB\|_1 \geq \|Z\|_1 > n + \|X\|_1$$

Therefore

$$\min_{Y\in\mathbb{R}^{n\times m}} G_X(Y) = \min_{Y\in\mathscr{A}_X} G_X(Y)$$

Furthermore notice then if $C \in \mathscr{C}_X$ for some $X$ then $\mathscr{A}_C \subseteq \mathscr{A}_X$. This follows from the fact if $Y \in \mathscr{A}_C$ then $\|Y\|_1 \leq n + \|C\|_1 \leq n + \|X\|_1$, since $\|C\|_1 \leq \|X\|_1$ as $C \in \mathscr{C}_X$.

Next let $H : \mathbb{R}^{m\times n} \times \mathbb{R}^{n\times m} \to \mathbb{R}$ be defined as $H(X,Y) := G_X(Y)$ and fix $B \in \mathbb{R}^{m\times n}$. Then for any $C \in \mathscr{C}_B$,

$$\begin{aligned} \mathcal{M}(C) &= \min_{Y\in\mathbb{R}^{n\times m}} G_C(Y) \\ &= \min_{Y\in\mathscr{A}_C} G_C(Y) \\ &= \min_{Y\in\mathscr{A}_B} G_C(Y) \qquad \text{since } \mathscr{A}_C \subseteq \mathscr{A}_B \\ &= \min_{Y\in\mathscr{A}_B} H(C,Y) \\ &= \min_{Y\in\mathscr{A}_B} H|_{\mathscr{C}_B\times\mathscr{A}_B}(C,Y) \end{aligned}$$

where the last line follows from the fact that $C \in \mathscr{C}_B$, by assertion, and the minimization is restricted to $Y \in \mathscr{A}_B$.

Now recall for any $X \in \mathbb{R}^{m\times n} \times \mathbb{R}^{n\times m} \to \mathbb{R}$ is defined as

$$H(X,Y) := G_X(Y) = \|Y\|_1 + \|XY\|_1 + \|I - YX\|_1 + \|X - XYX\|_1$$

Furthermore notice $\|\cdot\|_1$ is continuous everywhere since it is a norm by Proposition (2.2.2). Next the functions $(X,Y) \mapsto Y$, $(X,Y) \mapsto XY$, $(X,Y) \mapsto I - YX$, and $(X,Y) \mapsto X - XYX$ are clearly continuous on $\mathbb{R}^{m\times n} \times \mathbb{R}^{n\times m}$. Therefore since $H$ is a sum of a composition of continuous functions on $\mathbb{R}^{m\times n} \times \mathbb{R}^{n\times m}$, $H$ is continuous on $\mathbb{R}^{m\times n} \times \mathbb{R}^{n\times m}$.

Next because $\mathscr{A}_B$ is compact and $\mathscr{C}_B$ is compact, the set $\mathscr{C}_B \times \mathscr{A}_B$ is compact. Therefore since $H$ is continuous on $\mathbb{R}^{m\times n} \times \mathbb{R}^{n\times m}$, the function $H|_{\mathscr{C}_B\times\mathscr{A}_B}$ is continuous on $\mathscr{C}_B \times \mathscr{A}_B$. Further, because $\mathscr{C}_B \times \mathscr{A}_B$ is compact, $H|_{\mathscr{C}_B\times\mathscr{A}_B}$ is uniformly continuous on $\mathscr{C}_B \times \mathscr{A}_B$. Thus, by Proposition (2.2.6), $\mathcal{M}$ is continuous at $\mathscr{C}_B$ and in particular at $B$. Therefore since $B \in \mathbb{R}^{m\times n}$ was arbitrary, $\mathcal{M}$ is continuous on $\mathbb{R}^{m\times n}$.

□

The next results will establish that for any real orthogonal matrix $B$ there exists a real orthogonal matrix $A$ arbitrarily close to $B$ such that $A_{i,j} \neq 0$ for all $i$ and $j$. The above result will then be used to prove the main result of this section.

**Lemma 8.2.2.** *Let $A \in \mathbb{R}^{n\times n}$ be a real orthogonal matrix such that $A_{p,q} = 0$ and $A_{q,q} \neq 0$ for some $p$ and $q$. Then, given $\varepsilon > 0$, there exists a real orthogonal matrix $B \in \mathbb{R}^{n\times n}$ such that*

1. $\|A - B\|_\infty < \varepsilon$
2. $B_{p,q} \neq 0$
3. *If $A_{i,j} \neq 0$ then $B_{i,j} \neq 0$ for all $i$ and $j$*

**Proof.**

For $i \neq j$ let $G(i,j,\theta) \in \mathbb{R}^{n\times n}$ denote the matrix where

$$\begin{aligned}
G(i,j,\theta)_{k,k} &= 1 \qquad \text{for } k \neq i \text{ and } k \neq j \\
G(i,j,\theta)_{i,i} &= \cos\theta \\
G(i,j,\theta)_{i,j} &= -\sin\theta \\
G(i,j,\theta)_{j,i} &= \sin\theta \\
G(i,j,\theta)_{j,j} &= \cos\theta \\
G(i,j,\theta)_{s,t} &= 0 \qquad \text{for } s \neq t \text{ and } (s,t) \notin \{(i,j),(j,i)\}
\end{aligned}$$

That is,

$$G(i,j,\theta) = \begin{pmatrix}
1 & 0 & \dots & & \dots & 0 & 0 \\
0 & \ddots & & & & & 0 \\
\vdots & & \cos\theta & \dots & -\sin\theta & & \vdots \\
0 & & \vdots & \ddots & \vdots & & 0 \\
\vdots & & \sin\theta & \dots & \cos\theta & & \vdots \\
0 & & & & & \ddots & 0 \\
0 & 0 & \dots & & \dots & 0 & 1
\end{pmatrix}$$

Now for $0 < \theta < \pi/2$ let $G := G(p,q,\theta)$ and set $B = GA$. Notice $B$ is real orthogonal since $G$ and $A$ are real orthogonal since $B^TB = (GA)^TGA = A^TG^TGA = A^TA = I$. Further notice $B_{i,j} = A_{i,j}$ for $i \neq p$ and $i \neq q$. Next

$$\begin{aligned}
B_{p,j} &= \sum_{k=1}^{n} G_{p,k}A_{k,j} = G_{p,p}A_{p,j} + G_{p,q}A_{q,j} = \cos\theta A_{p,j} - \sin\theta A_{q,j} \\
B_{q,j} &= \sum_{k=1}^{n} G_{q,k}A_{k,j} = G_{q,p}A_{p,j} + G_{q,q}A_{q,j} = \sin\theta A_{p,j} + \cos\theta A_{q,j}
\end{aligned}$$

Let $s := \sin\theta$ and $c := \cos\theta = \sqrt{1-s^2}$. Then

$$\begin{aligned}
B_{p,j} &= cA_{p,j} - sA_{q,j} \\
B_{q,j} &= sA_{p,j} + cA_{q,j}
\end{aligned}$$

and thus

$$B_{p,q} = cA_{p,q} - sA_{q,q} = -sA_{q,q} \neq 0$$

since $A_{q,q} \neq 0$ by assumption and $0 < \theta < \pi/2$ implies $s \neq 0$. Now let $x := A_{p,j}$ and $y := A_{q,j}$ and notice then

$$B_{p,j} = cx - sy$$

$$B_{q,j} = sx + cy$$

Next let $M = \|A\|_\infty$ and notice because $A$ is real orthogonal $M > 0$. Now by direct calculation since $0 < s < 1$,

$$\begin{aligned}
1 + s - \sqrt{1 - s^2} \leq 2s &\iff -\sqrt{1 - s^2} \leq s - 1 \\
&\iff \sqrt{1 - s^2} \geq 1 - s \\
&\iff 1 - s^2 \geq (1 - s)^2 = 1 - 2s + s^2 \\
&\iff -s^2 \geq -2s + s^2 \\
&\iff 0 \geq -2s + 2s^2 = 2(s^2 - s) = 2s(s - 1)
\end{aligned}$$

Thus $2s(s-1) \leq 0$ since $0 < s < 1$ and hence $1 + s - \sqrt{1 - s^2} \leq 2s$. Therefore

$$\begin{aligned}
|B_{p,j} - A_{p,j}| &= |B_{p,j} - x| \\
&= |cx - sy - x| \\
&= |(c - 1)x - sy| \\
&= |(c - 1)x - sy| \\
&\leq |(c - 1)x| + |sy| \\
&= |c - 1| \cdot |x| + |s| \cdot |y| \\
&\leq |c - 1|M + |s|M \\
&= (|c - 1| + |s|)M \\
&= (1 - c + s)M \\
&= (1 + s - \sqrt{1 - s^2})M \\
&\leq 2sM
\end{aligned}$$

Similarly,

$$\begin{aligned}
|B_{q,j} - A_{q,j}| &= |B_{q,j} - y| \\
&= |sx + cy - y| \\
&= |sx + (c-1)y| \\
&\leq |sx| + |(c-1)y| \\
&= |s| \cdot |x| + |c-1| \cdot |y| \\
&\leq |s|M + |c-1|M \\
&= (|s| + |c-1|)M \\
&= (s + 1 - c)M \\
&= (1 + s - \sqrt{1-s^2})M \\
&\leq 2sM
\end{aligned}$$

Now let $\mu = \min\{|A_{i,j}| : A_{i,j} \neq 0\}$. Again since $A$ is real orthogonal $\mu > 0$. Last set $\nu = \min\{\mu, \varepsilon\}$. Now select $\theta \in (0, \pi/2)$ such that $s = \nu/(4M)$. Notice $0 < \nu \leq \mu \leq M < 4M$. Hence $0 < \nu/(4M) < 1$ and thus a $\theta$ exists. Then if $i = p$ or $i = q$,

$$|B_{i,j} - A_{i,j}| \leq 2sM = 2 \cdot \frac{\nu}{4M} \cdot M = \frac{\nu}{2} < \nu \leq \varepsilon$$

Thus by the construction of $\mu$, and since $0 < \nu \leq \mu$ it follows that $|B_{i,j} - A_{i,j}| < \mu$ for $i = p$ or $i = q$. Therefore if $A_{i,j} \neq 0$ then $|A_{i,j} - 0| \geq \mu$ and hence $B_{i,j} \neq 0$. Furthermore since $B_{i,j} = A_{i,j}$ for $i \neq p$ and $i \neq q$, the above calculation shows $\|A - B\|_\infty < \varepsilon$.

□

**Proposition 8.2.3.** *If $A \in \mathbb{R}^{n\times n}$ is real orthogonal and $\varepsilon > 0$ there exists $B \in \mathbb{R}^{n\times n}$ real orthogonal such that $B_{i,j} \neq 0$ for all $i$ and $j$ and $\|A - B\|_\infty < \varepsilon$.*

**Proof.**

This proposition follows from repeated use of the previous proposition since $A$ can have at most a finite number of zeros and because if $A_{p,q} = 0$ then there must exist some $r$ such that $A_{p,r} \neq 0$ since otherwise the $p$th row of $A$ is zero which contradicts the assumption that $A$ is real orthogonal.

In particular set $A_0 = A$ and consider the real orthogonal matrix $A_q$. If $(A_q)_{q,q} \neq 0$ set $\widetilde{A}_q = A_q$ otherwise, since $A_q$ is orthogonal, there must exist $\ell$ such that $(A_q)_{\ell,q} = 0$ since if this wasn't the case the $q$th column of $A_q$ would be a column of zeros. Then set $\widetilde{A}_q$ to be the result of switching the $q$th and $\ell$th rows of $A_q$. In either case notice $\widetilde{A}_q$ is real orthogonal and $(\widetilde{A})_{q,q} \neq 0$.

Now set $\widetilde{A}_q^{(0)} = \widetilde{A}_q$ and given $\widetilde{A}_q^{(p)}$ let $\widetilde{A}_q^{(p+1)} = \widetilde{A}_q^{(p)}$ if $(\widetilde{A}_q^{(p)})_{p,q} \neq 0$ and otherwise let $\widetilde{A}_q^{(p+1)}$ be the matrix obtained from the previous proposition such that $(\widetilde{A}_q^{(p+1)})_{p,q} \neq 0$. Now construct $\widetilde{A}_q^{(p)}$ for $p = 1, \ldots, n$.

Last, set $A_{q+1} = \widetilde{A}_q^{(n)}$ and construct $A_q$ for $q = 1, \ldots, n$. Then, following the procedure outlined above, $B := A_n$ has the property that $\|A - B\|_\infty < \varepsilon$ and $B_{i,j} \neq 0$ for all $i$ and $j$ and $B$ is real orthogonal.

□

Now for a fixed positive integer $n$, consider the sets

$$\mathcal{A} := \left\{ A \in \mathbb{R}^{n \times n} \,:\, A^T A = I \text{ and } A_{i,j} \neq 0 \text{ for all } i, j \right\}$$
$$\mathcal{B} := \left\{ B \in \mathbb{R}^{n \times n} \,:\, B^T B = I \right\}$$

Then the following is the main result of this section and shows that to ensure the zero matrix is a global minimizer of $G_B$ for any $B \in \mathcal{B}$, it is necessary and sufficient to ensure the zero matrix is a global minimizer of $G_A$ for any $A \in \mathcal{A}$. Thus, when minimizing the function $G_B$ for some real orthogonal matrix $B$, one can assume, without loss of generality, that $B_{i,j} \neq 0$ for all $i, j = 1, \ldots, n$. This assumption will simplify calculations in later sections.

**Theorem 8.2.4.** *The zero matrix is a global minimizer of $G_B$ for all $B \in \mathcal{B}$ if and only if the zero matrix is a global minimizer of $G_A$ for all $A \in \mathcal{A}$.*

**Proof.**

First an inequality between $\|\cdot\|_1$ and $\|\cdot\|_\infty$ needs to be established. To do so, notice for any $X \in \mathbb{R}^{n \times n}$

$$\|X\|_1 = \sum_{i,j=1}^{n} |X_{i,j}| \leq \sum_{i,j=1}^{n} \|X\|_\infty = n^2 \|X\|_\infty$$

where the fact that $|X_{i,j}| \leq \|X\|_\infty$ for all $i$ and $j$ is used.

Now fix $B$ a real orthogonal matrix and let $\varepsilon > 0$. Then because, by Proposition (8.2.1), $\mathcal{M}(\cdot)$ is continuous at $B$ there exists $\delta > 0$ such that if $C \in \mathbb{R}^{n\times n}$ is such that $\|C - B\|_\infty < \delta$ then $|\mathcal{M}(C) - \mathcal{M}(B)| < \frac{\varepsilon}{2}$. Here the norm $\|\cdot\|_\infty$ can be used because $\mathbb{R}^{n\times n}$ is finite dimensional and hence all norms on the space are equivalent.

Now let $\mu = \min\{\varepsilon, \delta\}$. Then by Proposition (8.2.3) there exists a real orthogonal matrix $A$ such that $A_{i,j} \neq 0$ for all $i$ and $j$ and $\|A - B\|_\infty < \frac{\mu}{2n^2}$. Further, by assumption, $\mathcal{M}(A) = n + \|A\|_1$. Next notice

$$|\|A\|_1 - \|B\|_1| \leq \|A - B\|_1 \leq n^2 \|A - B\|_\infty < \frac{\mu}{2} \leq \frac{\varepsilon}{2}$$

Further because $\|A - B\|_\infty < \frac{\mu}{2n^2} < \mu \leq \delta$, by the construction of $\mu$, it follows that $|\mathcal{M}(A) - \mathcal{M}(B)| < \frac{\varepsilon}{2}$. Therefore

$$\begin{aligned}
|\mathcal{M}(B) - (n + \|B\|_1)| &= |\mathcal{M}(B) - \mathcal{M}(A) + \mathcal{M}(A) - n - \|B\|_1| \\
&= |\mathcal{M}(B) - \mathcal{M}(A) + n + \|A\|_1 - n - \|B\|_1| \\
&= |\mathcal{M}(B) - \mathcal{M}(A) + \|A\|_1 - \|B\|_1| \\
&\leq |\mathcal{M}(B) - \mathcal{M}(A)| + |\,\|A\|_1 - \|B\|_1| \\
&< \frac{\varepsilon}{2} + \frac{\varepsilon}{2} \\
&= \varepsilon
\end{aligned}$$

Therefore since $\varepsilon > 0$ was arbitrary it has been established that

$$\mathcal{M}(B) = n + \|B\|_1$$

□

## 8.3 Results on Optimality Conditions

This section will develop near necessary conditions a real orthogonal matrix $B \in \mathbb{R}^{n\times n}$ must satisfy to ensure zero is a global minimizer of $G_B$. To describe this further, consider the following definition.

**Definition 8.3.1.** *For a positive integer $n$ and $\mathcal{K} \subseteq \mathbb{Z}_m \times \mathbb{Z}_n$ define*

$$\mathcal{F}(\mathcal{K}) := \left\{ X \in \mathbb{R}^{m\times n} \,:\, \|X\|_\infty \leq 1 \text{ and } X_{i,j} = 0 \text{ for all } (i,j) \in \mathcal{K} \right\}$$

*and for notational convenience let*

$$\mathcal{F}(n) := \mathcal{F}(\mathcal{D}_n)$$

*where*

$$\mathcal{D}_n := \{(1,1), \ldots, (n,n)\} \subseteq \mathbb{Z}_n \times \mathbb{Z}_n$$

Now based on the results of the previous section, one can assume, without loss of generality, that $B_{i,j} \neq 0$ for all $i$ and $j$. Then, with this assumption, recall Theorem (6.3.8) states that the zero matrix is a global minimizer of $G_B$ if and only if there exists matrices $\Lambda_1, \Lambda_2, \Lambda_3 \in \mathbb{R}^{n\times n}$ with $\|\Lambda_i\|_\infty \leq 1$ for $i = 1, \ldots, 4$ and $\Lambda_2 \in \mathcal{F}(n)$ such that

$$B^T \operatorname{Sgn}(B) B^T + B^T = \Lambda_1 + \Lambda_2 B^T + B^T \Lambda_3 \tag{8.2}$$

Now the term "near necessary conditions" is used since the construction developed in this chapter uses $\Lambda_1 = \operatorname{Sgn}(B)^T$, while the choice of this value is not required by Theorem (6.3.8).

That is, if $\Lambda_1 = \operatorname{Sgn}(B)^T$ and $\Lambda_2$ and $\Lambda_3$ can be found that satisfy the above conditions and equation (8.2), then the zero matrix is a global minimizer of $G_B$. However, a proof that shows if zero is a global minimizer of $G_B$ then it is possible to satisfy equation (8.2) with $\Lambda_1 = \operatorname{Sgn}(B)^T$ has not been found.

Instead, setting $\Lambda_1 = \operatorname{Sgn}(B)^T$ adds a great deal of symmetry, which will become apparent later, to equation (8.2). This symmetry makes equation (8.2) much easier to analyze. The overarching results of this chapter show that even with this symmetry, analyzing equation (8.2) is very difficult.

Furthermore, the cylindrical algebraic decomposition functionality build into Mathematica, which is used to find a solution to a system of inequalities, was used to find $\Lambda_1$, $\Lambda_2$, and $\Lambda_3$ satisfying equation (8.2) for thousands of randomly selected real orthogonal matrices of varying size. These numerical tests consistently showed that $\Lambda_1 = \operatorname{Sgn}(B)^T$. Although these results do not prove one can assume $\Lambda_1 = \operatorname{Sgn}(B)^T$ by themselves, the symmetry this assumption adds

to the problem coupled with strong numerical evidence suggests it is worthwhile to consider the consequences of setting $\Lambda_1 = \operatorname{Sgn}(B)^T$.

Thus suppose $\Lambda_1 = \operatorname{Sgn}(B)^T$. Then equation (8.2) is satisfied if

$$\Lambda_3 = B(-\Lambda_2)B^T + \operatorname{Sgn}(B)B^T - B\operatorname{Sgn}(B)^T + I \tag{8.3}$$

Notice $\operatorname{Sgn}(B)B^T - B\operatorname{Sgn}(B)^T = \operatorname{Sgn}(B)B^T - (\operatorname{Sgn}(B)B^T)^T$ which is one reason why the choice of setting $\Lambda_1 = \operatorname{Sgn}(B)^T$ adds symmetry to equation (8.2). Other nice properties of the expression $\operatorname{Sgn}(B)B^T - (\operatorname{Sgn}(B)B^T)^T$ will be studied later, and because this expression will be referenced many times, the following definition is provided.

**Definition 8.3.2.** *Define* $\mathcal{Z} : \mathbb{R}^{n\times m} \to \mathbb{R}^{n\times n}$ *as*

$$\mathcal{Z}(B) := \operatorname{Sgn}(B)B^T - (\operatorname{Sgn}(B)B^T)^T$$

Therefore equation (8.3) can be written as

$$-\Lambda_3 = B\Lambda_2 B^T - (\mathcal{Z}(B) + I) \tag{8.4}$$

The above work establishes the following theorem.

**Theorem 8.3.1.** *Let* $B \in \mathbb{R}^{n\times n}$ *be a real orthogonal matrix. If there exists* $X \in \mathcal{F}(n)$ *such that* $\left\|BXB^T - (\mathcal{Z}(B) + I)\right\|_\infty \leq 1$ *then the zero matrix is a global minimizer of* $G_B$.

**Proof.**

By Theorem (6.3.8), the zero matrix is a global minimizer of $G_B$ if and only if

$$B^T\operatorname{Sgn}(B)B^T + B^T = \Lambda_1 + \Lambda_2 B^T + B^T\Lambda_3 + B^T\Lambda_4 B^T \tag{8.5}$$

for some $\Lambda_1 \in \mathbb{R}^{n\times n}$, $\Lambda_2 \in \mathbb{R}^{n\times n}$, $\Lambda_3 \in \mathbb{R}^{n\times n}$, $\Lambda_4 \in \mathbb{R}^{n\times n}$ such that

1. $\|\Lambda_1\|_\infty, \|\Lambda_2\|_\infty, \|\Lambda_3\|_\infty, \|\Lambda_4\|_\infty \leq 1$
2. $(\Lambda_2)_{i,i} = 0$ for all $i = 1, \ldots, n$
3. $(\Lambda_4)_{i,j} = 0$ for all $i, j = 1, \ldots, n$ such that $B_{i,j} \neq 0$

Next, since $B$ is real orthogonal, equation (8.5) holds if and only if the following equation holds. This equation is obtained by multiplying (8.5) on the left by $B$.

$$\operatorname{Sgn}(B)B^T + I = B\Lambda_1 + B\Lambda_2 B^T + \Lambda_3 + \Lambda_4 B^T \tag{8.6}$$

Now suppose there exists $X \in \mathcal{F}(n)$ such that $\left\|BXB^T - (\mathcal{Z}(B) + I)\right\|_\infty \leq 1$. Then set $\Lambda_1 = \operatorname{Sgn}(B)^T$, $\Lambda_2 = X$, $\Lambda_3 = -(BXB^T - (\mathcal{Z}(B) + I))$, and $\Lambda_4 = 0$. Then

$$\begin{aligned} B\Lambda_1 + B\Lambda_2 B^T + \Lambda_3 + \Lambda_4 B^T &= B\operatorname{Sgn}(B)^T + BXB^T - (BXB^T - (\mathcal{Z}(B) + I)) \\ &= B\operatorname{Sgn}(B)^T + (\operatorname{Sgn}(B)B^T - B\operatorname{Sgn}(B)^T + I) \\ &= \operatorname{Sgn}(B)B^T + I \end{aligned}$$

Therefore the above $\Lambda_1$, $\Lambda_2$, $\Lambda_3$, and $\Lambda_4$ solve equation (8.6) and therefore equation (8.5). Next since $\Lambda_1$, $\Lambda_2$, $\Lambda_3$, and $\Lambda_4$ meet the conditions imposed by Theorem (6.3.8), the zero matrix is a global minimizer of $G_B$.

□

In fact, the previous theorem can be relaxed somewhat. That is, if one finds $X \in \mathbb{R}^{n\times n}$ with $\|X\|_\infty \leq 1$ (but not necessarily $X_{i,i} = 0$ for all $i = 1, \ldots, n$ as in the previous theorem), and $X$ satisfies $\left\|BXB^T - \mathcal{Z}(B)\right\|_\infty \leq 1$ (as apposed to $\left\|BXB^T - (\mathcal{Z}(B) + I)\right\|_\infty \leq 1$ as in the previous theorem) then zero is a global minimizer of $G_B$.

To establish this fact, the next two lemmas will be needed. Recall a matrix $C \in \mathbb{R}^{n\times n}$ is skew-symmetric if $C^T = -C$. Furthermore notice if $C$ is skew-symmetric then $C_{i,i} = (C^T)_{i,i} = -C_{i,i}$ for all $i = 1, \ldots, n$ and hence $C_{i,i} = 0$ for all $i = 1, \ldots, n$.

**Lemma 8.3.2.** *Given a matrix $B \in \mathbb{R}^{n\times n}$ and a skew-symmetric matrix $C \in \mathbb{R}^{n\times n}$ define the function $f : \mathbb{R}^{n\times n} \to \mathbb{R}$ as*

$$f(X) := \left\|BXB^T + C\right\|_\infty$$

*Then there exists $X \in \mathbb{R}^{n\times n}$ with $\|X\|_\infty \leq 1$ such that $f(X) \leq 1$ if and only if there exists a skew-symmetric matrix $Y \in \mathcal{F}(n)$ such that $f(Y) \leq 1$.*

**Proof.**

The reverse direction clearly holds. To prove the forward direction, suppose there exists $X \in \mathbb{R}^{n\times n}$ with $\|X\|_\infty \leq 1$ such that $f(X) \leq 1$. Now let $Y = \frac{1}{2}(X - X^T)$. Notice then

$$\|Y\|_\infty = \left\|\frac{1}{2}(X - X^T)\right\|_\infty = \frac{1}{2}\left\|X - X^T\right\|_\infty \leq \frac{1}{2}(\|X\|_\infty + \left\|-X^T\right\|_\infty) \leq \frac{1}{2}(1+1) = 1$$

Further, $Y^T = \frac{1}{2}(X^T - X) = -Y$ and therefore $Y$ is skew-symmetric. Thus $Y_{i,i} = 0$ for all $i = 1, \ldots, n$ and therefore $Y \in \mathcal{F}(n)$. Next, notice

$$\begin{aligned}
f(-X^T) &= \left\|B(-X^T)B^T + C\right\|_\infty \\
&= \left\|-BX^TB^T + C\right\|_\infty \\
&= \left\|-BX^TB^T - C^T\right\|_\infty \\
&= \left\|BX^TB^T + C^T\right\|_\infty \\
&= \left\|(BXB^T)^T + C^T\right\|_\infty \\
&= \left\|(BXB^T + C)^T\right\|_\infty \\
&= \left\|BXB^T + C\right\|_\infty \\
&= f(X)
\end{aligned}$$

Furthermore $f$ is convex on $\mathbb{R}^{n\times n}$ since for any $X_1, X_2 \in \mathbb{R}^{n\times n}$ and $\lambda_1, \lambda_2 \in [0,1]$ such that $\lambda_1 + \lambda_2 = 1$ one has that

$$\begin{aligned}
f(\lambda_1 X_1 + \lambda_2 X_2) &= \left\|B(\lambda_1 X_1 + \lambda_2 X_2)B^T + C\right\|_\infty \\
&= \left\|\lambda_1 BX_1B^T + \lambda_2 BX_2B^T + C\right\|_\infty \\
&= \left\|\lambda_1 BX_1B^T + \lambda_2 BX_2B^T + \lambda_1 C + \lambda_2 C\right\|_\infty \\
&= \left\|\lambda_1(BX_1B^T + C) + \lambda_2(BX_2B^T + C)\right\|_\infty \\
&\leq \left\|\lambda_1(BX_1B^T + C)\right\|_\infty + \left\|\lambda_2(BX_2B^T + C)\right\|_\infty \\
&= \lambda_1\left\|BX_1B^T + C\right\|_\infty + \lambda_2\left\|BX_2B^T + C\right\|_\infty \\
&= \lambda_1 f(X_1) + \lambda_2 f(X_2)
\end{aligned}$$

Therefore,

$$f(Y) = f(\frac{1}{2}X + \frac{1}{2}(-X^T)) \leq \frac{1}{2}f(X) + \frac{1}{2}f(-X^T) = \frac{1}{2}f(X) + \frac{1}{2}f(X) = f(X) \leq 1$$

□

**Lemma 8.3.3.** *Given a matrix $B \in \mathbb{R}^{n\times n}$ and a skew-symmetric matrix $C \in \mathbb{R}^{n\times n}$ define the functions $f, g : \mathbb{R}^{n\times n} \to \mathbb{R}$ as*

$$f(X) := \left\|BXB^T + C\right\|_\infty$$
$$g(X) := \left\|BXB^T + C + I\right\|_\infty$$

*Then there exists $X \in \mathcal{F}(n)$ such that $g(X) \leq 1$ if and only if there exists $Y \in \mathcal{F}(n)$ that $f(Y) \leq 1$.*

**Proof.**

For the forward direction, suppose there exists $X \in \mathcal{F}(n)$ such that $g(X) \leq 1$. Then setting $Z = BXB^T + C + I$, one has that $\|Z\|_\infty \leq 1$. Further,

$$Z^T = BX^TB^T + C^T + I = BX^TB^T - C + I$$

Therefore $-Z^T = -BX^TB^T + C - I$. Hence

$$\begin{aligned} Z - Z^T &= (BXB^T + C + I) + (-BX^TB^T + C - I) \\ &= B(X - X^T)B^T + 2C \end{aligned}$$

Thus $BYB^T + C = W$ where $Y = \frac{1}{2}(X - X^T)$ and $W = \frac{1}{2}(Z - Z^T)$. Notice $Y$ is skew symmetric since

$$Y^T = \frac{1}{2}(X^T - X) = -Y$$

Therefore, in particular, $Y_{i,i} = 0$ for all $i = 1, \ldots, n$. Furthermore,

$$\|Y\|_\infty = \frac{1}{2}\left\|X - X^T\right\|_\infty \leq \frac{1}{2}(\|X\|_\infty + \left\|X^T\right\|_\infty) \leq 1$$

Thus $Y \in \mathcal{F}(n)$. Additionally,

$$\|W\|_\infty = \frac{1}{2}\left\|Z - Z^T\right\|_\infty \leq \frac{1}{2}(\|Z\|_\infty + \left\|Z^T\right\|_\infty) \leq 1$$

Therefore $f(Y) = \|W\|_\infty \leq 1$.

For the reverse direction, suppose there exists $Y \in \mathbb{R}^{n\times n}$ with $\|Y\|_\infty \leq 1$ such that $f(Y) = \left\|BYB^T + C\right\|_\infty \leq 1$. Then, by the previous result, there exists a skew-symmetric $X \in \mathbb{R}^{n\times n}$

with $\|X\|_\infty \leq 1$ such that $f(X) \leq 1$. Notice then $(BXB^T)^T = BX^TB^T = B(-X)B^T = -BXB^T$. Therefore $BXB^T$ is skew-symmetric and hence $(BXB^T)_{i,i} = 0$ for all $i$. Similarly since $C$ is skew-symmetric $C_{i,i} = 0$ for all $i$. Thus because $f(X) = \left\|BXB^T + C\right\|_\infty \leq 1$ and $(BXB^T + C)_{i,i} = 0$ for all $i$ it follows that $\left|(BXB^T + C + I)_{i,j}\right| = \left|(BXB^T + C)_{i,j}\right| \leq 1$ for $i \neq j$ and $\left|(BXB^T + C + I)_{i,i}\right| = |I_{i,i}| = 1$. Therefore $g(X) = \left\|BXB^T + C + I\right\|_\infty \leq 1$. Last $X \in \mathcal{F}(n)$ since $\|X\|_\infty \leq 1$ and $X_{i,i} = 0$ for all $i$ since $X$ is skew-symmetric.

□

Now the previously mentioned modification of Theorem (8.3.1) can be formally stated and proved.

**Theorem 8.3.4.** *Let $B \in \mathbb{R}^{n\times n}$ be a real orthogonal matrix. If there exists $X \in \mathbb{R}^{n\times n}$ such that $\|X\|_\infty \leq 1$ and $\left\|BXB^T - \mathcal{Z}(B)\right\|_\infty \leq 1$ then the zero matrix is a global minimizer of $G_B$.*

**Proof.**

Suppose there exists $X \in \mathbb{R}^{n\times n}$ with $\|X\|_\infty \leq 1$ such that $\left\|BXB^T - \mathcal{Z}(B)\right\|_\infty \leq 1$. Notice then $\left\|B(-X)B^T + \mathcal{Z}(B)\right\|_\infty \leq 1$. Therefore, by the previous result, there exists $Y \in \mathcal{F}(n)$ such that

$$1 \geq \left\|BYB^T + \mathcal{Z}(B) + I\right\|_\infty = \left\|-(BYB^T + \mathcal{Z}(B) + I\right\|_\infty = \left\|B(-Y)B^T - (\mathcal{Z}(B) + I)\right\|_\infty$$

Thus there exists $W = -Y \in \mathcal{F}$ such that $\left\|BWB^T - (\mathcal{Z}(B) + I)\right\|_\infty \leq 1$. Therefore, by Theorem (8.3.1), the zero matrix is a global minimizer of $G_B$.

□

Now let $\mathcal{B}_\infty$ denote the closed unit ball in $\mathbb{R}^{n\times n}$ in terms of the $\|\cdot\|_\infty$ norm. That is,

$$\mathcal{B}_\infty := \left\{X \in \mathbb{R}^{n\times n} \,:\, \|X\|_\infty \leq 1\right\}$$

Notice then if $B \in \mathbb{R}^{n\times n}$ is a real orthogonal matrix then

$$B\mathcal{B}_\infty B^T := \left\{BXB^T \,:\, X \in \mathcal{B}_\infty\right\}$$

is the set obtained by rotating the unit ball $\mathcal{B}_\infty$ as described by the transformation $T : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ given by $T(X) = BXB^T$. That is, the set $B\mathcal{B}_\infty B^T$ is a rotated hypercube.

Next notice there exists $X \in \mathbb{R}^{n\times n}$ with $\|X\|_\infty \leq 1$ such that $\left\|BXB^T - \mathcal{Z}(B)\right\|_\infty \leq 1$ if and only if $(B\mathcal{B}_\infty B^T - \mathcal{Z}(B)) \cap \mathcal{B}_\infty \neq \emptyset$ if and only if $(B\mathcal{B}_\infty B^T) \cap (\mathcal{B}_\infty + \mathcal{Z}(B)) \neq \emptyset$.

That is, there exists $X \in \mathbb{R}^{n\times n}$ with $\|X\|_\infty \leq 1$ such that $\left\|BXB^T - \mathcal{Z}(B)\right\|_\infty \leq 1$ if and only if the rotated and shifted hypercube $B\mathcal{B}_\infty B^T - \mathcal{Z}(B)$ intersects the hypercube $\mathcal{B}_\infty$ if and only if the rotated hypercube $B\mathcal{B}_\infty B^T$ intersects the shifted hypercube $\mathcal{B}_\infty + \mathcal{Z}(B)$.

Now to characterize, when two hypercubes intersect, the concept of the distance from a point to a set will be needed. Based on the definition given in [3], given a closed nonempty set $\mathcal{C} \subseteq \mathbb{R}^{m\times n}$, a point $X_0 \in \mathbb{R}^{m\times n}$, and a norm $\|\cdot\|$ on $\mathbb{R}^{m\times n}$, the *distance from* $X_0$ *to* $\mathcal{C}$ is defined as

$$\operatorname{dist}(X_0, \mathcal{C}, \|\cdot\|) := \inf_{X\in\mathcal{C}} \|X - X_0\|$$

Notice since $\mathcal{C}$ is nonempty there exists $Y_0 \in \mathcal{C}$. Now let

$$\mathcal{B} = \left\{X \in \mathbb{R}^{m\times n} \,:\, \|X - Y_0\| \leq \|X_0 - Y_0\|\right\}$$

and notice $\mathcal{B}$ is closed and bounded. Therefore $\mathcal{C} \cap \mathcal{B}$ is closed and bounded and hence compact by the Heine-Borel theorem. Furthermore notice

$$\inf_{X\in\mathcal{C}} \|X - X_0\| = \inf_{X\in\mathcal{C}\cap\mathcal{B}} \|X - X_0\|$$

since if $X \in \mathcal{C} \cap \mathcal{B}^C$ then $\|X - X_0\| > \|Y_0 - X_0\|$. Last, since the function $\|\cdot - X_0\|$ is continuous on $\mathbb{R}^{m\times n}$ and $\mathcal{C} \cap \mathcal{B}$ is compact, there exists $X_* \in \mathcal{C}$ where the above infimum is attained. That is, $\operatorname{dist}(X_0, \mathcal{C}, \|\cdot\|) = \|X_* - X_0\|$.

Now the previous theorem can be characterized in terms of the distance from the point $-\mathcal{Z}(B^T)$ and the set $\mathcal{B}_\infty$ with respect to the appropriate norm. The next result makes this precise.

**Corollary 8.3.5.** *Let $B \in \mathbb{R}^{n\times n}$ be a real orthogonal matrix. Next let $|||\cdot||| : \mathbb{R}^{n\times n} \to \mathbb{R}$ be defined as $|||X|||_B := \left\|BXB^T\right\|_\infty$. Then there exists $X \in \mathbb{R}^{n\times n}$ with $\|X\|_\infty \leq 1$ such that $\left\|BXB^T - \mathcal{Z}(B)\right\|_\infty \leq 1$ if and only if*

$$\operatorname{dist}(-\mathcal{Z}(B^T), \mathcal{B}_\infty, |||\cdot|||_B) \leq 1.$$

*Therefore, if*

$$\operatorname{dist}(-\mathcal{Z}(B^T), \mathcal{B}_\infty, |||\cdot|||_B) \leq 1$$

*then the zero matrix is a global minimizer of* $G_B$.

**Proof.**

Notice

$$\begin{aligned} B^T\mathcal{Z}(B)B &= B^T(\operatorname{Sgn}(B)B^T - B\operatorname{Sgn}(B)^T)B \\ &= B^T\operatorname{Sgn}(B) - \operatorname{Sgn}(B)^T B \\ &= -(\operatorname{Sgn}(B)^T B - B^T\operatorname{Sgn}(B)) \\ &= -(\operatorname{Sgn}(B^T)(B^T)^T - B^T\operatorname{Sgn}(B^T)^T) \\ &= -\mathcal{Z}(B^T) \end{aligned}$$

Therefore given $X \in \mathbb{R}^{n\times n}$,

$$\begin{aligned} \left\|BXB^T - \mathcal{Z}(B)\right\|_\infty &= \left\|BXB^T - BB^T\mathcal{Z}(B)BB^T\right\|_\infty \\ &= \left\|B(X - B^T\mathcal{Z}(B)B)B^T\right\|_\infty \\ &= \left\|B(X + \mathcal{Z}(B^T))B^T\right\|_\infty \\ &= \left|\left|\left|X + \mathcal{Z}(B^T)\right|\right|\right|_B \end{aligned}$$

Therefore if there exists $X \in \mathbb{R}^{n\times n}$ with $\|X\|_\infty \leq 1$ such that $\left\|BXB^T - \mathcal{Z}(B)\right\|_\infty \leq 1$ if and only if there exists $X \in \mathcal{B}_\infty$ such that $\left|\left|\left|X - (-\mathcal{Z}(B^T))\right|\right|\right|_B \leq 1$ if and only if $\operatorname{dist}(-\mathcal{Z}(B^T), \mathcal{B}_\infty, |||\cdot|||_B) \leq 1$.

□

For an arbitrary real orthogonal matrix $B$, deriving a complete characterization of the norm $|||\cdot|||_B$ can be difficult. The next result shows that, for any real orthogonal matrix $B$, this norm can be approximated by the $\|\cdot\|_2$ which describes standard Euclidean distance.

**Corollary 8.3.6.** *Let* $B \in \mathbb{R}^{n\times n}$ *be a real orthogonal matrix. If* $\operatorname{dist}(-\mathcal{Z}(B^T), \mathcal{B}_\infty, \|\cdot\|_2) \leq 1$ *then the zero matrix is a global minimizer of* $G_B$.

**Proof.**

Recall the norm $|||\cdot||| : \mathbb{R}^{n\times n} \to \mathbb{R}$ defined in Corollary (8.3.5) defined as $|||X|||_B := \left\|BXB^T\right\|_\infty$.

Now given $X \in \mathbb{R}^{n\times n}$ let $p$ and $q$ be such that $|X_{p,q}| = \|X\|_\infty$. Notice then

$$\|X\|_\infty^2 = |X_{p,q}|^2 \leq |X_{p,q}|^2 + \sum_{(i,j)\neq(p,q)} |X_{i,j}|^2 = \|X\|_2^2$$

Hence $\|X\|_\infty \leq \|X\|_2$. Next, notice

$$\mathrm{Tr}(X^TX) = \sum_{i=1}^n (X^TX)_{i,i} = \sum_{i,j=1}^n X_{i,j}^T X_{j,i} = \sum_{i,j=1}^n (X_{j,i})^2 = \|X\|_2^2$$

Thus

$$\begin{aligned}
\left\|BXB^T\right\|_2^2 &= \mathrm{Tr}((BXB^T)^T(BXB^T)) \\
&= \mathrm{Tr}(BX^TB^TBXB^T) \\
&= \mathrm{Tr}(BX^TXB^T) \\
&= \mathrm{Tr}(X^TXB^TB) \\
&= \mathrm{Tr}(X^TX) \\
&= \|X\|_2^2
\end{aligned}$$

Hence, following the details from the previous result,

$$\left|\left|\left|X + \mathcal{Z}(B^T)\right|\right|\right|_B = \left\|B(X + \mathcal{Z}(B^T))B^T\right\|_\infty \leq \left\|B(X + \mathcal{Z}(B^T))B^T\right\|_2 = \left\|X + \mathcal{Z}(B^T)\right\|_2$$

Therefore, if $\mathrm{dist}(-\mathcal{Z}(B^T), \mathcal{B}_\infty, \|\cdot\|_2) \leq 1$ then there exists $X \in \mathcal{B}_\infty$ such that $\left\|X + \mathcal{Z}(B^T)\right\|_2 \leq 1$. Hence, by the above computation, there exists $X \in \mathcal{B}_\infty$ such that $\left|\left|\left|X + \mathcal{Z}(B^T)\right|\right|\right|_B \leq 1$. Thus, by the previous corollary, the zero matrix is a global minimizer of $G_B$.

□

## 8.4 Analyzing the Optimality Conditions

Recall by Theorem (8.3.1) if there exists $X \in \mathcal{F}(n)$ such that $\left\|BXB^T - (\mathcal{Z}(B) + I)\right\|_\infty \leq 1$, then the zero matrix is a global minimizer of $G_B$. That is, if there exists $X \in \mathbb{R}^{n\times n}$ that satisfies the system

$$\begin{aligned}
-1 \leq (BXB^T)_{i,j} - (\mathcal{Z}(B) + I)_{i,j} \leq 1 \qquad & \text{for all } i,j = 1,\ldots,n \\
X_{i,i} = 0 \qquad & \text{for all } i = 1,\ldots,n \\
-1 \leq X_{i,j} \leq 1 \qquad & \text{for all } i \neq j
\end{aligned} \tag{8.7}$$

then the zero matrix is a global minimizer of $G_B$. This chapter will use Fourier-Motzkin elimination to analyze when systems of the above form have a solution.

Notice by Corollary (8.3.5) it is enough to show there exists $X \in \mathcal{F}(n)$ such that

$$\left\| BXB^T - \mathcal{Z}(B) \right\|_\infty \leq 1$$

to ensure the zero matrix is a global minimizer of $G_B$. However, the analysis in the following chapter does actually not rely on Corollary (8.3.5). Because of this, the following chapter will address the analysis of system (8.7).

Fourier-Motzkin elimination was first discovered by Fourier and later rediscovered by Motzkin. It is an extension of Gaussian-Elimination to systems of linear inequalities, and is a method to systematically remove variables from a system to find a solution to the system. In particular, suppose one has a system of linear inequalities of the form

$$a_{i,1}x_1 + \cdots + a_{i,n}x_n \leq b_i$$

for $i = 1, \ldots, m$. To remove the variable $x_1$ from the system, group the inequalities into three groups depending on whether $a_{i,1} > 0$, $a_{i,1} < 0$, or $a_{i,1} = 0$. Suppose there are $P$ inequalities of the first kind, $N$ inequalities of the second, and $Z$ inequalities of the third. The process of Fourier-Motzkin Elimination takes the $P$ inequalities where $a_{i,1} > 0$ for some $i$ and the $N$ inequalities where $a_{i,1} < 0$ for some $i$ and replaces them in a new system with $PN$ inequalities not involving the variable $x_1$. Next, the $Z$ inequalities in the original system where $a_{i,1} = 0$ for some $i$ are simply placed in the new system.

It was shown by Fourier and Motzkin that this new system, not involving the variable $x_1$, has a solution if and only if the original system had a solution. Note, however, that the original system had $P + N + Z$ inequalities while the new system has $PN + Z$ inequalities.

To remove the variable $x_2$ one would just repeat the process for the new system generated. In theory, one could step-by-step remove all variables from the system, and upon removing the last variable from the system, the newly constructed system would simply be a system of linear inequalities only involving constants. One then would check if each inequality in that system was valid to determine if the original series had a solution. In practice, however, this

method has the disadvantage that at each iteration of the method, the number of inequalities in the system grows exponentially. This is a problem when using Fourier-Motzkin Elimination in practice, but for theoretic results, this problem can sometimes be sidestepped.

The remainder of this section will formally state Fourier-Motzkin elimination and prove that it can be used as described above to determine if a system of inequalities has a solution. Then Fourier-Motzkin elimination will be used to analyze system (8.7). The ultimate goal is then to find conditions a real orthogonal matrix $B \in \mathbb{R}^{n\times n}$ must satisfy provided it is known that there exists $X \in \mathcal{F}(n)$ such that $\left\|BXB^T - (\mathcal{Z}(B) + I)\right\|_\infty \leq 1$. The results of the next section will then show all real orthogonal matrices satisfy these conditions.

First a system of inequalities is said to be *feasible* if there exists values for all the variables of the system that satisfy all the inequalities in the system. The next result shows how to remove any constraint that requires a variable in a system be zero to yield a new system that is feasible if and only if the original system was feasible.

**Proposition 8.4.1.** *Let $\mathcal{I}$ be a finite index set that indexes a collection of inequalities with variables $\{x_k\}_{k\in\mathcal{K}_1\cup\mathcal{K}_0}$ for some finite disjoint index sets $\mathcal{K}_1$ and $\mathcal{K}_0$. Further suppose $a_{i,k} \neq 0$ for all $i \in \mathcal{I}$ and $k \in \mathcal{K}_1 \cup \mathcal{K}_0$. The system*

$$\begin{aligned}
\sum_{k\in\mathcal{K}_1\cup\mathcal{K}_0} a_{i,k}x_k &\leq b_i \qquad \textit{for all} \quad i \in \mathcal{I} \\
x_k &\leq 1 \qquad \textit{for all} \quad k \in \mathcal{K}_1 \\
-x_k &\leq 1 \qquad \textit{for all} \quad k \in \mathcal{K}_1 \\
x_k &= 0 \qquad \textit{for all} \quad k \in \mathcal{K}_0
\end{aligned} \tag{$*'$}$$

*is feasible if and only if the system*

$$\begin{aligned}
\sum_{k\in\mathcal{K}_1} a_{i,k}x_k &\leq b_i \qquad i \in \mathcal{I} \\
x_k &\leq 1 \qquad k \in \mathcal{K}_1 \\
-x_k &\leq 1 \qquad k \in \mathcal{K}_1
\end{aligned} \tag{$*''$}$$

*is feasible.*

**Proof.**

Notice since $x_k = 0$ for all $k \in \mathcal{K}_0$ one has that

$$\sum_{k \in \mathcal{K}_1 \cup \mathcal{K}_0} a_{i,k} x_k = \sum_{k \in \mathcal{K}_1} a_{i,k} x_k$$

Therefore the first condition and the fourth condition hold simultaneously in system ($*'$) if and only if

$$\sum_{k \in \mathcal{K}_1} a_{i,k} x_k \leq b_i$$

for all $i \in \mathcal{I}$. Next the second and fourth conditions in system ($*'$) hold simultaneously if and only if the second condition holds since $x_k = 0$ for $k \in \mathcal{K}_0$ does not influence the second condition. The same holds for the third condition. Thus system ($*'$) is equivalent to the system ($*''$).

□

The next result is extremely useful for taking a system of linear inequalities and deriving another system of inequalities, which can be more easily handled with Fourier-Motzkin Elimination, such that the feasibility of the first system implies the feasibility of the second system. The derivation of this second system is a way to search for a counterexample to the statement that the first system must be feasible.

**Proposition 8.4.2.** *Let $\mathcal{I}$ be a finite index set that indexes a collection of inequalities with variables $\{x_k\}_{k \in \mathcal{K}}$ for some finite index set $\mathcal{K}$. If the system of inequalities*

$$\begin{aligned} \sum_{k \in \mathcal{K}} a_{i,k} x_k &\leq b_i && \text{for all } i \in \mathcal{I} \\ x_k &\leq 1 && \text{for all } k \in \mathcal{K} \\ -x_k &\leq 1 && \text{for all } k \in \mathcal{K} \end{aligned} \tag{$*$}$$

*is feasible then so is the system*

$$\begin{aligned} \sum_{k \in \mathcal{K}} y_{i,k} &\leq b_i && \text{for all } i \in \mathcal{I} \\ y_{i,k} &\leq |a_{i,k}| && \text{for all } i \in \mathcal{I} \text{ and } k \in \mathcal{K} \\ -y_{i,k} &\leq |a_{i,k}| && \text{for all } i \in \mathcal{I} \text{ and } k \in \mathcal{K} \end{aligned} \tag{$**$}$$

**Proof.**

Suppose $x = \{x_k\}_{k\in\mathcal{K}}$ is a solution to system (∗) and define $y_{i,k} := a_{i,k}x_k$ for $i \in \mathcal{I}$ and $k \in \mathcal{K}$. Then for any $i \in \mathcal{I}$,

$$\sum_{k\in\mathcal{K}} y_{i,k} = \sum_{k\in\mathcal{K}} a_{i,k}x_k \leq b_i$$

Next, since $x$ solves system (∗) it follows that

$$x_k \leq 1 \qquad k \in \mathcal{K} \tag{A}$$

$$-x_k \leq 1 \qquad k \in \mathcal{K} \tag{B}$$

Now for any $i \in \mathcal{I}$ either $a_{i,k} \geq 0$ or $a_{i,k} < 0$. If $a_{i,k} \geq 0$ then (A) and (B) respectively imply $y_{i,k} = a_{i,k}x_k \leq a_{i,k} = |a_{i,k}|$ and $-y_{i,k} = -a_{i,k}x_k \leq a_{i,k} = |a_{i,k}|$. Otherwise if $a_{i,k} < 0$ then (A) and (B) respectively imply $y_{i,k} = a_{i,k}x_k \geq a_{i,k} = -|a_{i,k}|$ and $-y_{i,k} = -a_{i,k}x_k \geq a_{i,k} = -|a_{i,k}|$ which imply $-y_{i,k} \leq |a_{i,k}|$ and $y_{i,k} \leq |a_{i,k}|$. Therefore, in either case, $y_{i,k} \leq |a_{i,k}|$ and $-y_{i,k} \leq |a_{i,k}|$ for $i \in \mathcal{I}$ and $k \in \mathcal{K}$ and hence $y$ satisfies system (∗∗).

□

The following is a formal statement and proof of Fourier-Motzkin Elimination as described in [28] and [20].

**Proposition 8.4.3. (Fourier-Motzkin Elimination)** *Consider for three mutually disjoint index sets $\mathcal{I}_+$, $\mathcal{I}_-$, and $\mathcal{I}_0$ the system of inequalities*

$$x_1 + \sum_{j=2}^{n} a_{i,j}x_j \leq b_i \qquad \text{for all } i \in \mathcal{I}_+ \tag{8.8}$$

$$-x_1 + \sum_{j=2}^{n} a_{i,j}x_j \leq b_i \qquad \text{for all } i \in \mathcal{I}_- \tag{8.9}$$

$$\sum_{j=2}^{n} a_{i,j}x_j \leq b_i \qquad \text{for all } i \in \mathcal{I}_0 \tag{8.10}$$

*Then the above system is feasible if and only if the system*

$$\sum_{j=2}^{n} a_{i,j}x_j \leq b_i \qquad \text{for all } i \in \mathcal{I}_0 \tag{8.11}$$

$$\sum_{j=2}^{n} (a_{i,j} + a_{k,j})x_j \leq b_i + b_k \qquad \text{for all } i \in \mathcal{I}_- \text{ and } k \in \mathcal{I}_+ \tag{8.12}$$

*is feasible.*

**Proof.**

Suppose $x \in \mathbb{R}^n$ is a solution to the first system. Then adding inequalities (8.8) and (8.9) one has that for any $i \in \mathcal{I}_-$ and $k \in \mathcal{I}_+$,

$$\sum_{j=2}^{n} a_{i,j}x_j + \sum_{j=2}^{n} a_{k,j}x_j \leq b_i + b_k$$

Hence

$$\sum_{j=2}^{n} a_{i,j}x_j - b_i \leq b_k - \sum_{j=2}^{n} a_{k,j}x_j$$

for $i \in \mathcal{I}_-$ and $k \in \mathcal{I}_+$. Therefore $x$ is a solution to the second system.

Now, for the converse, suppose $(y_2, \ldots, y_n)$ is a solution to the second system and let

$$U = \min\left\{ b_k - \sum_{j=2}^{n} a_{k,j}y_j \ : \ k \in \mathcal{I}_+ \right\}$$
$$L = \max\left\{ \sum_{j=2}^{n} a_{i,j}y_j - b_i \ : \ i \in \mathcal{I}_- \right\}$$

Notice if $L > U$ then because $U$ and $L$ must be achieved for some $i$ and $k$ one has

$$\sum_{j=2}^{n} a_{i,j}y_j - b_i = L > U = b_k - \sum_{j=2}^{n} a_{k,j}y_j$$

which contradicts the assumption that $(y_2, \ldots, y_n)$ solves the second system. Thus $L \leq U$.

Now set $y_1 = (L + U)/2$. Then $L \leq y_1 \leq U$. Hence for any $k \in \mathcal{I}_+$ it follows that

$$y_1 \leq U \leq b_k - \sum_{j=2}^{n} a_{k,j}y_j$$

and therefore

$$y_1 + \sum_{j=2}^{n} a_{k,j}y_j \leq b_k$$

Similarly, for any $i \in \mathcal{I}_-$ it follows that

$$\sum_{j=2}^{n} a_{i,j}y_j - b_i \leq L \leq y_1$$

Therefore

$$-y_1 + \sum_{j=2}^{n} a_{i,j}y_j \leq b_i$$

and hence $(y_1, y_2, \ldots, y_n)$ solves the first system.

□

By applying Fourier-Motzkin Elimination to the second system in Proposition (8.4.2) necessary conditions that the coefficients of the system must satisfy to ensure feasibility of the system can be derived. This is the topic of the next result.

**Proposition 8.4.4.** *Let $\mathcal{I}$ be a finite index set that indexes a collection of inequalities with variables $\{y_{i,k}\}_{(i,k)\in\mathcal{I}\times\mathcal{K}}$ for some finite index set $\mathcal{K}$. The system*

$$\begin{aligned} \sum_{k\in\mathcal{K}} y_{i,k} &\leq b_i && i\in\mathcal{I} \\ y_{i,k} &\leq |a_{i,k}| && i\in\mathcal{I} \text{ and } k\in\mathcal{K} \\ -y_{i,k} &\leq |a_{i,k}| && i\in\mathcal{I} \text{ and } k\in\mathcal{K} \end{aligned}$$

*is feasible if and only if*

$$b_i + \sum_{k\in\mathcal{K}} |a_{i,k}| \geq 0$$

*for all $i\in\mathcal{I}$.*

**Proof.**

For fixed $p\in\mathcal{I}$ and $q\in\mathcal{K}$ write the initial system in the form

$$\left.\begin{aligned} y_{p,q} + \sum_{k\in\mathcal{K}\setminus\{q\}} y_{p,k} &\leq b_p \\ y_{p,q} &\leq |a_{p,q}| \\ -y_{p,q} &\leq |a_{p,q}| \end{aligned}\right\} \text{(A)}$$

$$\left.\begin{aligned} y_{p,k} &\leq |a_{p,k}| && k\in\mathcal{K}\setminus\{q\} \\ -y_{p,k} &\leq |a_{p,k}| && k\in\mathcal{K}\setminus\{q\} \end{aligned}\right\} \text{(B)}$$

$$\left.\begin{aligned} \sum_{k\in\mathcal{K}} y_{i,k} &\leq b_i && i\in\mathcal{I}\setminus\{p\} \\ y_{i,k} &\leq |a_{i,k}| && i\in\mathcal{I}\setminus\{p\} \text{ and } k\in\mathcal{K} \\ -y_{i,k} &\leq |a_{i,k}| && i\in\mathcal{I}\setminus\{p\} \text{ and } k\in\mathcal{K} \end{aligned}\right\} \text{(C)}$$

Here the system has been broken into three subsystems. The first three lines defines a subsystem, which will be denoted as subsystem (A), that involves the variable $y_{p,q}$. The next two lines defines a subsystem, which will be denoted as subsystem (B), that describes the inequalities involving the variables $y_{i,j}$ where $i = p$ but $j \neq q$. The last three lines defines a subsystem, which will be denoted as subsystem (C), that describes the variables $y_{i,j}$ where $i \neq p$.

Therefore the only subsystem that involves the variable $y_{p,q}$ is subsystem (A). Thus, using Fourier-Motzkin elimination, the above system is feasible if and only if a second system is feasible. This second system contains exactly the subsystems (B) and (C) since they do not involve the variable $y_{p,q}$. However, the second system involves a modified version of subsystem (A) where the variable $y_{p,q}$ has been removed.

In particular, every inequality in subsystem (A) that involves a $-y_{p,q}$ term is compared to every other inequality that involves a $y_{p,q}$ term. There is one inequality, on the third line of subsystem (A), of the former type and there are two inequalities, on the first two lines of subsystem (A), of the latter type. Thus if subsystem (A') denotes the new system constructed from subsystem (A) after performing Fourier-Motzkin elimination to remove the variable $y_{p,q}$ from the system, the number of inequalities in subsystem (A') is the product of the number of inequalities with a $-y_{p,q}$ term with the number of inequalities with a $y_{p,q}$ term. In this case, subsystem (A') will consist of two inequalities.

Specifically the inequalities on the first and third lines of subsystem (A) are replaced by the inequality

$$-|a_{p,q}| \leq b_p - \sum_{k \in \mathcal{K} \setminus \{q\}} y_{p,k}$$

in subsystem (A'). Further the second and the third lines of subsystem (A) are replaced by the inequality

$$-|a_{p,q}| \leq |a_{p,q}|$$

in subsystem (A'). However, this inequality is always true for any $a_{p,q} \in \mathbb{R}$. Therefore subsystem (A) is replaced by subsystem (A') consisting of the single inequality

$$\sum_{k \in \mathcal{K} \setminus \{q\}} y_{p,k} \leq b_p + |a_{p,q}|$$

Hence, in summary, the original system is feasible if and only if the following system (a system with the variable $y_{p,q}$ removed) is feasible,

$$\left.\begin{aligned} \sum_{k\in\mathcal{K}\setminus\{q\}} y_{p,k} \le b_p + |a_{p,q}| \end{aligned}\right\} \text{(A')}$$

$$\left.\begin{aligned} y_{p,k} &\le |a_{p,k}| \qquad k \in \mathcal{K}\setminus\{q\} \\ -y_{p,k} &\le |a_{p,k}| \qquad k \in \mathcal{K}\setminus\{q\} \end{aligned}\right\} \text{(B)}$$

$$\left.\begin{aligned} \sum_{k\in\mathcal{K}} y_{i,k} &\le b_i \qquad\qquad i \in \mathcal{I}\setminus\{p\} \\ y_{i,k} &\le |a_{i,k}| \qquad i \in \mathcal{I}\setminus\{p\} \text{ and } k \in \mathcal{K} \\ -y_{i,k} &\le |a_{i,k}| \qquad i \in \mathcal{I}\setminus\{p\} \text{ and } k \in \mathcal{K} \end{aligned}\right\} \text{(C)}$$

Now because $\mathcal{K}$ is a finite set the above procedure can be repeated to eliminate $y_{p,k}$ from the original system for any $k \in \mathcal{K}$. Then, after a finite number of iterations, the original system is feasible if and only if the following system is feasible,

$$\sum_{k\in\mathcal{K}\setminus\{\mathcal{K}\}} y_{p,k} \le b_p + \sum_{k\in\mathcal{K}} |a_{p,k}|$$

$$\begin{aligned} y_{p,k} &\le |a_{p,k}| \qquad\qquad k \in \mathcal{K}\setminus\{\mathcal{K}\} \\ -y_{p,k} &\le |a_{p,k}| \qquad\qquad k \in \mathcal{K}\setminus\{\mathcal{K}\} \end{aligned}$$

$$\begin{aligned} \sum_{k\in\mathcal{K}} y_{i,k} &\le b_i \qquad\qquad i \in \mathcal{I}\setminus\{p\} \\ y_{i,k} &\le |a_{i,k}| \qquad i \in \mathcal{I}\setminus\{p\} \text{ and } k \in \mathcal{K} \\ -y_{i,k} &\le |a_{i,k}| \qquad i \in \mathcal{I}\setminus\{p\} \text{ and } k \in \mathcal{K} \end{aligned}$$

This system is

$$0 \leq b_p + \sum_{k \in \mathcal{K}} |a_{p,k}|$$

$$\begin{aligned} \sum_{k \in \mathcal{K}} y_{i,k} &\leq b_i && i \in \mathcal{I} \setminus \{p\} \\ y_{i,k} &\leq |a_{i,k}| && i \in \mathcal{I} \setminus \{p\} \text{ and } k \in \mathcal{K} \\ -y_{i,k} &\leq |a_{i,k}| && i \in \mathcal{I} \setminus \{p\} \text{ and } k \in \mathcal{K} \end{aligned}$$

Further as the above system does not involve any variable $y_{i,k}$ if $i = p$ since all the variables $\{y_{p,k}\}_{k \in \mathcal{K}}$ have been removed from the original system to form a new system. Thus, again, this process can be repeated for every $p \in \mathcal{I}$ since $\mathcal{I}$ is finite. After doing so, the system created is,

$$0 \leq b_i + \sum_{k \in \mathcal{K}} |a_{i,k}| \qquad \text{for all } i \in \mathcal{I}$$

Thus, because at each step the new system constructed is feasible if and only if the original system is feasible, it follows that the original system is feasible if and only if the above system is feasible, completing the proof.

□

Again the goal of this section is to find conditions a real orthogonal matrix $B \in \mathbb{R}^{n \times n}$ must satisfy provided it is known that there exists $X \in \mathcal{F}(n)$ such that $\left\| BXB^T - (\mathcal{Z}(B) + I) \right\|_\infty \leq 1$. The next result will allow such conditions to be derived.

**Theorem 8.4.5.** *Fix $A \in \mathbb{R}^{m \times n}$, $B \in \mathbb{R}^{M \times N}$, and $C \in \mathbb{R}^{m \times N}$ such that $A_{i,j} \neq 0$ and $B_{i,j} \neq 0$ for all $i$ and $j$ and let $\mathcal{K} \subseteq \mathbb{Z}_n \times \mathbb{Z}_M$. If there exists $X \in \mathcal{F}(\mathcal{K})$ such that $\|AXB + C\|_\infty \leq 1$ then*

$$\sum_{(k,\ell) \notin \mathcal{K}} |A_{i,k} B_{\ell,j}| \geq |C_{i,j}| - 1$$

*for all $i = 1, \ldots, m$ and $j = 1, \ldots, N$.*

**Proof.**

Notice there exists $X \in \mathcal{F}(\mathcal{K})$ such that $\|AXB + C\|_\infty \leq 1$ if and only if the following system

is feasible.

$$\begin{aligned}
(AXB+C)_{i,j} &\leq 1 \qquad \text{for all } i=1,\dots,m,\ j=1,\dots,N \\
-(AXB+C)_{i,j} &\leq 1 \qquad \text{for all } i=1,\dots,m,\ j=1,\dots,N \\
X_{i,j} &\leq 1 \qquad \text{for all } (i,j)\notin\mathcal{K} \\
-X_{i,j} &\leq 1 \qquad \text{for all } (i,j)\notin\mathcal{K} \\
X_{i,j} &= 0 \qquad \text{for all } (i,j)\in\mathcal{K}
\end{aligned}$$

Expanding this system yields the following,

$$\begin{aligned}
\sum_{\ell=1}^{M}\sum_{k=1}^{n} A_{i,k}X_{k,\ell}B_{\ell,j}+C_{i,j} &\leq 1 \qquad \text{for all } i=1,\dots,m,\ j=1,\dots,N \\
\sum_{\ell=1}^{M}\sum_{k=1}^{n} -A_{i,k}X_{k,\ell}B_{\ell,j}-C_{i,j} &\leq 1 \qquad \text{for all } i=1,\dots,m,\ j=1,\dots,N \\
X_{i,j} &\leq 1 \qquad \text{for all } (i,j)\notin\mathcal{K} \\
-X_{i,j} &\leq 1 \qquad \text{for all } (i,j)\notin\mathcal{K} \\
X_{i,j} &= 0 \qquad \text{for all } (i,j)\in\mathcal{K}
\end{aligned}$$

That is defining $D_{i,j,k,\ell} = A_{i,k}B_{\ell,j}$ for $i=1,\dots,m$, $j=1,\dots,N$, $k=1,\dots,n$, and $\ell=1,\dots,M$,

$$\begin{aligned}
\sum_{\ell=1}^{M}\sum_{k=1}^{n} D_{i,j,k,\ell}X_{k,\ell} &\leq 1-C_{i,j} \qquad \text{for all } i=1,\dots,m,\ j=1,\dots,N \\
\sum_{\ell=1}^{M}\sum_{k=1}^{n} -D_{i,j,k,\ell}X_{k,\ell} &\leq 1+C_{i,j} \qquad \text{for all } i=1,\dots,m,\ j=1,\dots,N \\
X_{i,j} &\leq 1 \qquad \text{for all } (i,j)\notin\mathcal{K} \\
-X_{i,j} &\leq 1 \qquad \text{for all } (i,j)\notin\mathcal{K} \\
X_{i,j} &= 0 \qquad \text{for all } (i,j)\in\mathcal{K}
\end{aligned}$$

By Proposition (8.4.1) the above system is feasible if and only if the following system is feasible.

$$\begin{aligned}
\sum_{(k,\ell)\notin\mathcal{K}} D_{i,j,k,\ell}X_{k,\ell} &\leq 1-C_{i,j} \qquad \text{for all } i=1,\dots,m,\ j=1,\dots,N \\
\sum_{(k,\ell)\notin\mathcal{K}} -D_{i,j,k,\ell}X_{k,\ell} &\leq 1+C_{i,j} \qquad \text{for all } i=1,\dots,m,\ j=1,\dots,N
\end{aligned}$$

$$X_{i,j} \leq 1 \qquad \text{for all } (i,j) \notin \mathcal{K}$$

$$-X_{i,j} \leq 1 \qquad \text{for all } (i,j) \notin \mathcal{K}$$

Now defining $E_{i,j,k,\ell}$ for $i = 1, \ldots, 2m$ and $j = 1, \ldots, N$ by

$$E_{i,j,k,\ell} := \begin{cases} D_{i,j,k,\ell} & \text{if } i = 1, \ldots, m \\ -D_{i,j,k,\ell} & \text{if } i = m+1, \ldots, 2m \end{cases}$$

and defining $F_{i,j}$ for $i = 1, \ldots, 2m$ and $j = 1, \ldots, N$ by

$$F_{i,j} := \begin{cases} 1 - C_{i,j} & \text{if } i = 1, \ldots, m \\ 1 + C_{i,j} & \text{if } i = m+1, \ldots, 2m \end{cases}$$

the above system can be written in the form

$$\sum_{(k,\ell) \notin \mathcal{K}} E_{i,j,k,\ell} X_{k,\ell} \leq F_{i,j} \qquad \text{for all } i = 1, \ldots, 2m,\ j = 1, \ldots, N$$

$$X_{k,\ell} \leq 1 \qquad \text{for all } (k,\ell) \notin \mathcal{K}$$

$$-X_{k,\ell} \leq 1 \qquad \text{for all } (k,\ell) \notin \mathcal{K}$$

Then by, Proposition (8.4.2), the above system is feasible if and only if the following system is feasible.

$$\sum_{(k,\ell) \notin \mathcal{K}} Y_{i,j,k,\ell} \leq F_{i,j} \qquad \text{for all } i = 1, \ldots, 2m,\ j = 1, \ldots, N$$

$$Y_{i,j,k,\ell} \leq |E_{i,j,k,\ell}| \qquad \text{for all } i = 1, \ldots, 2m,\ j = 1, \ldots, N,\ (k,\ell) \notin \mathcal{K}$$

$$-Y_{i,j,k,\ell} \leq |E_{i,j,k,\ell}| \qquad \text{for all } i = 1, \ldots, 2m,\ j = 1, \ldots, N,\ (k,\ell) \notin \mathcal{K}$$

Next by, Proposition (8.4.4), the above system is feasible if and only if

$$F_{i,j} + \sum_{(k,\ell) \notin \mathcal{K}} |E_{i,j,k,\ell}| \geq 0$$

for all $i = 1, \ldots, 2m$ and $j = 1, \ldots, N$ if and only if

$$F_{i,j} + \sum_{(k,\ell) \notin \mathcal{K}} |E_{i,j,k,\ell}| \geq 0 \qquad \text{for all } i = 1, \ldots, m,\ j = 1, \ldots, N$$

$$F_{i,j} + \sum_{(k,\ell) \notin \mathcal{K}} |E_{i,j,k,\ell}| \geq 0 \qquad \text{for all } i = m+1, \ldots, 2m,\ j = 1, \ldots, N$$

if and only if

$$1 - C_{i,j} + \sum_{(k,\ell)\notin\mathcal{K}} |D_{i,j,k,\ell}| \geq 0 \qquad \text{for all } i = 1,\dots,m,\ j = 1,\dots,N$$

$$1 + C_{i,j} + \sum_{(k,\ell)\notin\mathcal{K}} |-D_{i,j,k,\ell}| \geq 0 \qquad \text{for all } i = 1,\dots,m,\ j = 1,\dots,N$$

if and only if

$$\sum_{(k,\ell)\notin\mathcal{K}} |D_{i,j,k,\ell}| \geq C_{i,j} - 1 \qquad \text{for all } i = 1,\dots,m,\ j = 1,\dots,N$$

$$\sum_{(k,\ell)\notin\mathcal{K}} |D_{i,j,k,\ell}| \geq -C_{i,j} - 1 \qquad \text{for all } i = 1,\dots,m,\ j = 1,\dots,N$$

if and only if

$$\sum_{(k,\ell)\notin\mathcal{K}} |D_{i,j,k,\ell}| \geq \max\{C_{i,j} - 1, -C_{i,j} - 1\} \qquad \text{for all } i = 1,\dots,m,\ j = 1,\dots,N \tag{8.13}$$

Now define $f : \mathbb{R} \to \mathbb{R}$ by $f(x) := \max\{x - 1, -x, -1\}$. Notice then if $x \geq 0$ then $-x \leq x$, hence $-x - 1 \leq x - 1$, and therefore $f(x) = x - 1 = |x| - 1$. Similarly if $x \leq 0$ then $x \leq -x$, hence $x - 1 \leq -x - 1$, and therefore $f(x) = -x - 1 = |x| - 1$. Therefore for any $x \in \mathbb{R}$ it has been shown $f(x) = |x| - 1$. Thus inequality (8.13) holds if and only if

$$\sum_{(k,\ell)\notin\mathcal{K}} |D_{i,j,k,\ell}| \geq |C_{i,j}| - 1 \qquad \text{for all } i = 1,\dots,m,\ j = 1,\dots,N \tag{8.14}$$

if and only if

$$\sum_{(k,\ell)\notin\mathcal{K}} |A_{i,k}B_{\ell,j}| \geq |C_{i,j}| - 1 \qquad \text{for all } i = 1,\dots,m,\ j = 1,\dots,N \tag{8.15}$$

□

The following is the main result of the section. It states that if the zero matrix is a global minimizer of $G_B$, and $\Lambda_1 = \mathrm{Sgn}(B)^T$ in Proposition (6.3.5), then roughly the entries of $\mathcal{Z}(B)$ have to be bounded in a special way in terms of the $\ell_1$ norms of the rows of $B$.

**Theorem 8.4.6.** *Let $B \in \mathbb{R}^{n\times n}$ be a real orthogonal matrix such that $B_{i,j} \neq 0$ for all $i, j = 1,\dots,n$. If there exists $X \in \mathcal{F}(n)$ such that $\left\|BXB^T - (\mathcal{Z}(B) + I)\right\|_\infty \leq 1$ then*

$$\langle |b_i|, |b_j| \rangle + \mathcal{Z}(B)_{i,j} \leq 1 + \|b_i\|_1 \cdot \|b_j\|_1$$

*for all $i \neq j$, where $b_i$ denotes the $i$th row of $B$.*

**Proof.**

Notice from Proposition (8.4.5) if there exists $X \in \mathcal{F}(n)$ such that $\left\|BXB^T - (\mathcal{Z}(B) + I)\right\|_\infty \leq 1$ then

$$\sum_{k \neq \ell} \left|B_{i,k} B^T_{\ell,j}\right| \geq \left|\mathcal{Z}(B)_{i,j} + I_{i,j}\right| - 1 \tag{8.16}$$

for all $i, j = 1, \ldots, n$. Next, for any $i$ and $j$,

$$\begin{aligned}
\sum_{k \neq \ell} \left|B_{i,k} B^T_{\ell,j}\right| &= \sum_{k=1}^{n} \sum_{\ell=1}^{n} \left|B_{i,k} B^T_{\ell,j}\right| - \sum_{k=1}^{n} \left|B_{i,k} B^T_{k,j}\right| \\
&= \sum_{k=1}^{n} \sum_{\ell=1}^{n} |b_{i,k}| \cdot |b_{j,\ell}| - \sum_{k=1}^{n} |b_{i,k}| \cdot |b_{j,k}| \\
&= \left(\sum_{k=1}^{n} |b_{i,k}|\right) \left(\sum_{\ell=1}^{n} |b_{j,\ell}|\right) - \sum_{k=1}^{n} |b_{i,k}| \cdot |b_{j,k}| \\
&= \|b_i\|_1 \cdot \|b_j\|_1 - \langle |b_i|, |b_j| \rangle
\end{aligned}$$

where $b_i$ denotes the $i$th row of $B$. Therefore inequality (8.16) holds for all $i$ and $j$ if and only if the following inequality holds for all $i$ and $j$.

$$\langle |b_i|, |b_j| \rangle + \left|\mathcal{Z}(B)_{i,j} + I_{i,j}\right| \leq 1 + \|b_i\|_1 \cdot \|b_j\|_1 \tag{8.17}$$

Next notice for any $i$ and $j$,

$$\begin{aligned}
\mathcal{Z}(B)_{i,j} &= (\mathrm{Sgn}(B)B^T)_{i,j} - (\mathrm{Sgn}(B)B^T)^T_{i,j} \\
&= (\mathrm{Sgn}(B)B^T)_{i,j} - (\mathrm{Sgn}(B)B^T)_{j,i} \\
&= \langle \mathrm{Sgn}(b_i), b_j \rangle - \langle \mathrm{Sgn}(b_j), b_i \rangle \\
&= \langle \mathrm{Sgn}(b_i), b_j \rangle - \langle b_i, \mathrm{Sgn}(b_j) \rangle
\end{aligned}$$

Hence for any $i$,

$$\mathcal{Z}(B)_{i,i} = \langle \mathrm{Sgn}(b_i), b_i \rangle - \langle b_i, \mathrm{Sgn}(b_i) \rangle = 0$$

Thus for $i = j$ inequality (8.17) reduces to

$$\langle |b_i|, |b_i| \rangle + 1 \leq 1 + \|b_i\|_1 \cdot \|b_i\|_1$$

which is equivalent to

$$\langle |b_i|, |b_i| \rangle \leq \|b_i\|_1 \cdot \|b_i\|_1$$

In fact for any $x, y \in \mathbb{R}^n$ one has by the Cauchy-Schwarz inequality that

$$\langle |x|, |y| \rangle \leq \| \, |x| \, \|_2 \cdot \| \, |y| \, \|_2 = \|x\|_2 \cdot \|y\|_2 \leq \|x\|_1 \cdot \|y\|_1$$

since $\|z\|_2 \leq \|z\|_1$ for any $z \in \mathbb{R}^n$. Thus inequality (8.17) necessarily holds for $i = j$. Thus inequality (8.17) holds for all $i$ and $j$ if and only if it holds for all $i \neq j$. Furthermore $I_{i,j} = 0$ for $i \neq j$ and thus inequality (8.17) simplifies to

$$\langle |b_i|, |b_j| \rangle + |\mathcal{Z}(B)_{i,j}| \leq 1 + \|b_i\|_1 \cdot \|b_j\|_1$$

Last since $\mathcal{Z}(B)_{i,j} = -\mathcal{Z}(B)_{j,i}$ it follows that the above inequality holds for all $i \neq j$ if and only if

$$\langle |b_i|, |b_j| \rangle + \mathcal{Z}(B)_{i,j} \leq 1 + \|b_i\|_1 \cdot \|b_j\|_1 \tag{8.18}$$

holds for all $i \neq j$.

□

## 8.5 Verifying the Optimality Conditions

Based on the results of the previous section, the goal of this section is to show that for any real orthogonal matrix $B \in \mathbb{R}^{n \times n}$,

$$\langle |b_i|, |b_j| \rangle + \mathcal{Z}(B)_{i,j} \leq 1 + \|b_i\|_1 \cdot \|b_j\|_1$$

for all $i, j = 1, \ldots, n$ where $b_i$ denotes the $i$th row of $B$. Next recall $\mathcal{Z}(B)_{i,j} = \langle \text{Sgn}(b_i), b_j \rangle - \langle b_i, \text{Sgn}(b_j) \rangle$. Thus the goal of this section is to show

$$\langle |b_i|, |b_j| \rangle + \langle \text{Sgn}(b_i), b_j \rangle - \langle b_i, \text{Sgn}(b_j) \rangle \leq 1 + \|b_i\|_1 \cdot \|b_j\|_1$$

for all $i, j = 1, \ldots, n$. That is,

$$\langle |b_i|, |b_j| \rangle + \langle \text{Sgn}(b_i), b_j \rangle - \langle b_i, \text{Sgn}(b_j) \rangle - \|b_i\|_1 \cdot \|b_j\|_1 - 1 \leq 0$$

for all $i, j = 1, \ldots, n$. The above inequality prompts the following definition.

**Definition 8.5.1.** *Define* $J : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ *as*

$$J(x, y) := \langle |x|, |y| \rangle + \langle \text{Sgn}(x), y \rangle - \langle x, \text{Sgn}(y) \rangle - \|x\|_1 \cdot \|y\|_1 - 1$$

Thus the goal of this section is to show if $B \in \mathbb{R}^{n\times n}$ is real orthogonal then $J(b_i, b_j) \leq 0$ for all $i, j = 1, \ldots, n$. To do so notice for any $a \in \mathbb{R}$ one has $a = \operatorname{sgn}(a)|a|$. Thus for any $x, y \in \mathbb{R}^n$,

$$\begin{aligned}
\langle \operatorname{Sgn}(x), y\rangle - \langle x, \operatorname{Sgn}(y)\rangle &\leq |\langle \operatorname{Sgn}(x), y\rangle - \langle x, \operatorname{Sgn}(y)\rangle| \\
&= \left|\sum_{i=1}^{n} \operatorname{sgn}(x_i)y_i - \sum_{i=1}^{n} x_i \operatorname{sgn}(y_i)\right| \\
&= \left|\sum_{i=1}^{n} (\operatorname{sgn}(x_i)y_i - \operatorname{sgn}(y_i)x_i)\right| \\
&= \left|\sum_{i=1}^{n} (\operatorname{sgn}(x_i)\operatorname{sgn}(y_i)|y_i| - \operatorname{sgn}(y_i)\operatorname{sgn}(x_i)|x_i|)\right| \\
&= \left|\sum_{i=1}^{n} \operatorname{sgn}(x_iy_i)(|y_i| - |x_i|)\right| \\
&\leq \sum_{i=1}^{n} |\operatorname{sgn}(x_iy_i)(|y_i| - |x_i|)| \\
&= \sum_{i=1}^{n} |\operatorname{sgn}(x_iy_i)| \cdot \big||y_i| - |x_i|\big| \\
&\leq \sum_{i=1}^{n} \big||y_i| - |x_i|\big| \\
&= \| \, |y| - |x| \, \|_1
\end{aligned}$$

The above work prompts the following definition.

**Definition 8.5.2.** *Define* $K : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ *as*

$$K(x, y) := \langle |x|, |y|\rangle + \| \, |x| - |y| \, \|_1 - \|x\|_1 \cdot \|y\|_1 - 1$$

Then using the above observation, notice for any $x, y \in \mathbb{R}^n$,

$$\begin{aligned}
J(x, y) &= \langle |x|, |y|\rangle + \langle \operatorname{Sgn}(x), y\rangle - \langle x, \operatorname{Sgn}(y)\rangle - \|x\|_1 \cdot \|y\|_1 - 1 \\
&\leq \langle |x|, |y|\rangle + \| \, |y| - |x| \, \|_1 - \|x\|_1 \cdot \|y\|_1 - 1 \\
&= K(x, y)
\end{aligned}$$

The goal, then is to show $K(b_i, b_j) \leq 0$ for all $i, j = 1, \ldots, n$ where $b_i$ is the $i$th row of a real orthogonal matrix $B$. Notice then if $B$ is real orthogonal then $\|b_i\|_2 = 1$ for all $i = 1, \ldots, n$. Furthermore notice for any $x, y \in \mathbb{R}$ then $K(x, y) = K(|x|, |y|)$ by the definition of $K$. Thus

one only needs to establish $K(x,y) \leq 0$ for all $x, y \in \mathbb{R}^n$ such that $\|x\|_2 = \|y\|_2 = 1$ and $x_i \geq 0$ and $y_i \geq 0$ for all $i = 1, \ldots, n$. Based on this observation consider the following which formally defines the nonnegative orthant.

**Definition 8.5.3.** *For a positive integer $n$ define*

$$\mathbb{R}^n_+ := \{x \in \mathbb{R}^n : x_i \geq 0 \text{ for all } i = 1, \ldots, n\}$$

Furthermore the following defines the boundary of the $\ell_2$ ball restricted to the nonnegative orthant. This set will be useful later.

**Definition 8.5.4.** *For a positive integer $n$ define*

$$\mathbb{S}^{n-1} := \{x \in \mathbb{R}^n \, : \, \|x\|_2 = 1\}$$

*and*

$$\mathbb{S}^{n-1}_+ := \mathbb{S}^{n-1} \cap \mathbb{R}^n_+$$

Thus because $K(x,y) = K(|x|, |y|)$ for all $x, y \in \mathbb{S}^{n-1}$ to show $K(x,y) \leq 0$ for all $x, y \in \mathbb{S}^{n-1}$ it is enough to show $K(x,y) \leq 0$ for all $x, y \in \mathbb{S}^{n-1}_+$. In fact, it will be shown $K(x,y) \leq 0$ for all $x, y \in \mathcal{Q}(n)$ where the set $\mathcal{Q}(n)$ is defined below.

**Definition 8.5.5.** *For a positive integer $n$ define*

$$\mathcal{Q}(n) = \left\{x \in \mathbb{R}^n_+ \, : \, \|x\|_1 \geq 1\right\} \cap \left\{x \in \mathbb{R}^n_+ \, : \, \|x\|_\infty \leq 1\right\}$$

The set $\mathcal{Q}(n)$ will be used because, first, it will be shown that $\mathcal{Q}(n)$ is nonempty, compact, convex, and has a finite number of extreme points, second, $\mathbb{S}^{n-1}_+ \subseteq \mathcal{Q}(n)$, and third, $K(\cdot,\cdot)$ is convex on $\mathcal{Q}(n)$. Thus the Maximum Principle can be applied to establish that $K(\cdot,\cdot) \leq 0$ on $\mathcal{Q}(n)$ and hence $\mathbb{S}^{n-1}_+$. First all of the aforementioned statements about $\mathcal{Q}(n)$ and $K(\cdot,\cdot)$ will be proven.

**Definition 8.5.6.** *For fixed $y \in \mathbb{R}^n_+$ define $L_y : \mathbb{R}^n_+ \to \mathbb{R}$ as*

$$L_y(x) := \sum_{i=1}^n y_i x_i + \sum_{i=1}^n |x_i - y_i| - \left(\sum_{i=1}^n y_i\right)\left(\sum_{i=1}^n x_i\right) - 1$$

Notice then for any $x, y \in \mathbb{R}^n_+$ one has, by the definition above, $L_y(x) = K(x, y)$. Now it will be shown that for any $y \in \mathbb{R}^n_+$ the function $L_y(\cdot)$ is convex on $\mathbb{R}^n_+$.

**Proposition 8.5.1.** *For fixed $y \in \mathbb{R}^n_+$ the function $L_y$ is convex on $\mathbb{R}^n_+$.*

**Proof.**

Let $x_1, x_2 \in \mathbb{R}^n_+$ and $\lambda_1, \lambda_2 \in [0, 1]$ such that $\lambda_1 + \lambda_2 = 1$. Then

$$\begin{aligned}
L_y(\lambda_1 x_1 + \lambda_2 x_2) &= \sum_{i=1}^n y_i(\lambda_1 x_1 + \lambda_2 x_2)_i + \sum_{i=1}^n |(\lambda_1 x_1 + \lambda_2 x_2)_i - y_i| \\
&\qquad - \left(\sum_{i=1}^n y_i\right)\left(\sum_{i=1}^n (\lambda_1 x_1 + \lambda_2 x_2)_i\right) - 1 \\
&= \lambda_1 \sum_{i=1}^n y_i (x_1)_i + \lambda_2 \sum_{i=1}^n y_i (x_2)_i + \sum_{i=1}^n |\lambda_1 (x_1)_i + \lambda_2 (x_2)_i - \lambda_1 y_i - \lambda_2 y_i| \\
&\qquad - \left(\sum_{i=1}^n y_i\right)\left(\sum_{i=1}^n (\lambda_1 (x_1)_i + \lambda_2 (x_2)_i)\right) - 1 \\
&= \lambda_1 \sum_{i=1}^n y_i (x_1)_i + \lambda_2 \sum_{i=1}^n y_i (x_2)_i + \sum_{i=1}^n |\lambda_1 ((x_1)_i - y_i) + \lambda_2 ((x_2)_i - y_i)| \\
&\qquad - \left(\sum_{i=1}^n y_i\right)\left(\lambda_1 \sum_{i=1}^n (x_1)_i + \lambda_2 \sum_{i=1}^n (x_2)_i\right) - \lambda_1 - \lambda_2 \\
&\leq \lambda_1 \sum_{i=1}^n y_i (x_1)_i + \lambda_2 \sum_{i=1}^n y_i (x_2)_i + \sum_{i=1}^n (|\lambda_1 ((x_1)_i - y_i)| + |\lambda_2 ((x_2)_i - y_i)|) \\
&\qquad - \left(\sum_{i=1}^n y_i\right)\left(\lambda_1 \sum_{i=1}^n (x_1)_i + \lambda_2 \sum_{i=1}^n (x_2)_i\right) - \lambda_1 - \lambda_2 \\
&= \lambda_1 \sum_{i=1}^n y_i (x_1)_i + \lambda_2 \sum_{i=1}^n y_i (x_2)_i + \lambda_1 \sum_{i=1}^n |(x_1)_i - y_i| + \lambda_2 \sum_{i=1}^n |(x_2)_i - y_i| \\
&\qquad - \lambda_1 \left(\sum_{i=1}^n y_i\right)\left(\sum_{i=1}^n (x_1)_i\right) - \lambda_2 \left(\sum_{i=1}^n y_i\right)\left(\sum_{i=1}^n (x_2)_i\right) - \lambda_1 - \lambda_2 \\
&= \lambda_1 \left(\sum_{i=1}^n y_i (x_1)_i + \sum_{i=1}^n |(x_1)_i - y_i| - \left(\sum_{i=1}^n y_i\right)\left(\sum_{i=1}^n (x_1)_i\right) - 1\right) + \\
&\qquad \lambda_2 \left(\sum_{i=1}^n y_i (x_2)_i + \sum_{i=1}^n |(x_2)_i - y_i| - \left(\sum_{i=1}^n y_i\right)\left(\sum_{i=1}^n (x_2)_i\right) - 1\right) \\
&= \lambda_1 L_y(x_1) + \lambda_2 L_y(x_2)
\end{aligned}$$

Therefore $L_y$ is convex on $\mathbb{R}^n_+$.

□

The next result shows that $\mathcal{Q}(n)$ is in fact, nonempty, compact, and convex. Thus by the Krein-Milman Theorem, $\mathcal{Q}(n)$ is the closed convex hull of its extreme points. The following definition will be used to characterize these extreme points.

**Definition 8.5.7.** *For $\mathcal{I} \subseteq \{1, \ldots, n\}$ define $e_{\mathcal{I}} \in \mathbb{R}^n$ as*

$$(e_{\mathcal{I}})_i = \begin{cases} 1 & \text{if } i \in \mathcal{I} \\ 0 & \text{else} \end{cases}$$

**Proposition 8.5.2.** *For a positive integer $n$, the set $\mathcal{Q}(n)$ is non-empty, compact, and convex and if $z$ is an extreme point of $\mathcal{Q}(n)$ then $z = e_{\mathcal{I}}$ for some $\mathcal{I} \subseteq \{1, \ldots, n\}$. Furthermore $\mathbb{S}_+^{n-1} \subseteq \mathcal{Q}(n)$.*

**Proof.**

Clearly $\mathcal{Q}(n)$ is bounded since $\|x\|_\infty \leq 1$ for any $x \in \mathcal{Q}(n)$. Next let

$$\mathcal{Q}_1 := \left\{x \in \mathbb{R}_+^n \,:\, \|x\|_1 \geq 1\right\}$$
$$\mathcal{Q}_2 := \left\{x \in \mathbb{R}_+^n \,:\, \|x\|_\infty \leq 1\right\}$$

Now let $x, y \in \mathcal{Q}_1$ with $0 \leq \lambda \leq 1$. Then

$$\sum_{i=1}^n x_i \geq 1$$
$$\sum_{i=1}^n y_i \geq 1$$

Hence $\lambda, 1-\lambda \geq 0$ means

$$\sum_{i=1}^n \lambda x_i \geq \lambda$$
$$\sum_{i=1}^n (1-\lambda) y_i \geq 1-\lambda$$

Thus

$$\sum_{i=1}^n (\lambda x + (1-\lambda)y)_i = \sum_{i=1}^n \lambda x_i + \sum_{i=1}^n (1-\lambda) y_i \geq \lambda + 1 - \lambda = 1$$

Thus $\lambda x + (1-\lambda)y \in \mathcal{Q}_1$ and hence $\mathcal{Q}_1$ is convex. Similarly let $u, v \in \mathcal{Q}_2$ and $0 \leq \mu \leq 1$. Then

$$\|\mu u + (1-\mu)v\|_\infty \leq \|\mu u\|_\infty + \|(1-\mu)v\|_\infty = \mu \|u\|_\infty + (1-\mu)\|v\|_\infty \leq \mu + (1-\mu) = 1$$

Thus $\mu u + (1-\mu)v \in \mathcal{Q}_2$ and therefore $\mathcal{Q}_2$ is convex. Hence $\mathcal{Q} = \mathcal{Q}_1 \cap \mathcal{Q}_2$ is convex. Next, if $w \in \mathbb{S}_+^{n-1}$ then $w_i \geq 0$ for all $i = 1, \ldots, n$. Next, $\|w\|_1 \geq \|w\|_2 = 1$. Hence $w \in \mathcal{Q}_1$. Furthermore $\|w\|_2 = 1$ implies $w_i \leq 1$ for all $i$ and hence $w \in \mathcal{Q}_2$. Therefore $w \in \mathcal{Q}$ and hence $\mathbb{S}_+^{n-1} \subseteq \mathcal{Q}(n)$. Therefore $\mathcal{Q}(n)$ is nonempty.

Next, to show $\mathcal{Q}(n)$ is closed, it will be show that $\mathcal{Q}(n)^C$ is open. To do so notice if $z \in \mathcal{Q}(n)^C$ then either $\|z\|_\infty > 1$ or $\|z\|_1 < 1$. Suppose $\|z\|_\infty > 1$. Then there exists $i$ such that $z_i > 1$. Without loss of generality, assume $z_1 > 1$. Now let $\varepsilon = z_1 - 1$. Notice then for for any

$$x \in A := \{x \in \mathbb{R}^n \,:\, \|z - x\|_\infty < \varepsilon\}$$

one has $|x_1 - z_1| \leq \|z - x\|_\infty < \varepsilon$. Hence $-\varepsilon < x_1 - z_1 < \varepsilon$ which implies $x_1 > z_1 - \varepsilon = 1$. Therefore $x \in \mathcal{Q}(n)^C$ and hence $A \subseteq \mathcal{Q}(n)^C$.

Next suppose $z \in \mathcal{Q}(n)^C$ such that $\|z\|_1 < 1$ and let $\varepsilon = 1 - \|z\|_1$. Notice then for any

$$x \in B := \{x \in \mathbb{R}^n \,:\, \|z - x\|_\infty < \varepsilon\}$$

one has

$$\|x\|_1 = \|x - z + z\|_1 \leq \|x - z\|_1 + \|z\|_1 < \varepsilon + \|z\|_1 = 1 - \|z\|_1 + \|z\|_1 = 1$$

Hence $x \in \mathcal{Q}(n)^C$ and therefore $B \subseteq \mathcal{Q}(n)^C$. Thus it has been shown that for any $z \in \mathcal{Q}(n)^C$ there exists an open neighborhood of $z$ contained in $\mathcal{Q}(n)^C$. Thus $\mathcal{Q}(n)^C$ is open and hence $\mathcal{Q}(n)$ is closed. Thus by the Heine-Borel theorem $\mathcal{Q}(n)$ is compact.

To prove if $z$ is an extreme point of $\mathcal{Q}(n)$ then $z = e_{\mathcal{I}}$ for some $\mathcal{I} \subseteq \{1, \ldots, n\}$, the contrapositive will be established. First, if $z \in \mathcal{Q}(n)$ such that $\|z\|_1 > 1$ and $\|z\|_\infty < 1$ then I claim $z$ is not an extreme point of $\mathcal{Q}(n)$. To see why suppose, without loss of generality, that $z_1 \geq z_i$ for all $i$. Now let $\varepsilon_1 = \frac{1}{2}(\|z\|_1 - 1)$ and notice since $\|z\|_1 > 1$ one has that $\varepsilon_1 > 0$. Next let $\varepsilon_2 = (1 - z_1)/2$ and notice $\|z\|_\infty < 1$ implies $\varepsilon_2 > 0$. Last let $\varepsilon = \min\{\varepsilon_1, \varepsilon_2, z_1/2\}$. Now define $x$ as $x_1 = z_1 - \varepsilon$ and $x_i = z_i$ for $i \neq 1$. Similarly, define $y$ as $y_1 = z_1 + \varepsilon$ and $y_i = z_i$ for

$i \neq 1$. Notice then by the construction of $\varepsilon$

$$\|x\|_1 = \sum_{i=1}^{n} x_i = z_1 - \varepsilon + \sum_{i=2}^{n} x_i = z_1 - \varepsilon + \sum_{i=2}^{n} z_i = \|z\|_1 - \varepsilon \geq \|z\|_1 - \frac{1}{2}(\|z\|_1 - 1) \geq 1$$
$$\|y\|_1 = \sum_{i=1}^{n} y_i = z_1 + \varepsilon + \sum_{i=2}^{n} y_i = z_1 + \varepsilon + \sum_{i=2}^{n} z_i = \|z\|_1 + \varepsilon \geq \|z\|_1 > 1$$

Also, necessarily $0 \leq x_i = z_i < 1$ and $0 \leq y_i = z_i < 1$ for $i \neq 1$. Further, $z_1 - \varepsilon \geq z_1 - z_1/2 = z_1/2 \geq 0$. Also, $z_1 - \varepsilon \leq z_1 < 1$. Hence $0 \leq x_1 < 1$. Similarly $z_1 + \varepsilon \geq z_1 \geq 0$ and

$$z_1 + \varepsilon \leq z_1 + \frac{1}{2}(1 - z_1) = \frac{1}{2} + \frac{1}{2}z_1 < \frac{1}{2} + \frac{1}{2} = 1$$

Thus $0 \leq y_1 < 1$. Hence $x, y \in \mathcal{Q}(n)$. Furthermore

$$\frac{1}{2}x_1 + \frac{1}{2}y_1 = \frac{1}{2}(z_1 - \varepsilon) + \frac{1}{2}(z_1 + \varepsilon) = z_1$$

and $\frac{1}{2}x_i + \frac{1}{2}y_i = \frac{1}{2}z_i + \frac{1}{2}z_i = z_i$ for $i \neq 1$. Thus

$$z = \frac{1}{2}x + \frac{1}{2}y$$

That is $z$ is a convex sum of the elements $x, y \in \mathcal{Q}(n)$ with $x, y \neq z$ and thus $z$ is not an extreme point of $\mathcal{Q}(n)$.

Now suppose $z \in \mathcal{Q}(n)$ is such that $\|z\|_\infty = 1$. Then there exists $i$ such that $z_i = 1$. Without loss of generality assume $z_1 = 1$. It will then be shown that if there does not exist $\mathcal{I} \subseteq \{1, \ldots, n\}$ such that $z = e_{\mathcal{I}}$ then $z$ cannot be an extreme point of $\mathcal{Q}(n)$. Then, under this condition, there exists $i$ such that $z_i \in (0, 1)$. Without loss of generality suppose $z_2 \in (0, 1)$. Now let $\varepsilon = \frac{1}{2}\min\{1 - z_1, z_1\}$ and define $x$ as $x_2 = z_2 - \varepsilon$ and $x_i = z_i$ for $i \neq 2$ and define $y$ as $y_2 = z_2 + \varepsilon$ and $y_i = z_i$ for $i \neq 2$. Then, by construction, $\frac{1}{2}x + \frac{1}{2}y = z$. Furthermore, $0 < z_2 - \varepsilon < 1$ and $0 < z_2 + \varepsilon < 1$ by the construction of $\varepsilon$. Therefore since $0 \leq z_i \leq 1$ for all $i$ it follows that $\|x\|_\infty, \|y\|_\infty \leq 1$. Furthermore, since $z_2 - \varepsilon > 0$ it follows that

$$\|x\|_1 = \sum_{i=1}^{n} x_i \geq x_1 + x_2 = 1 + z_2 - \varepsilon \geq 1$$

and since $z_2 + \varepsilon > 0$ it follows that

$$\|y\|_1 = \sum_{i=1}^{n} y_i \geq y_1 + y_2 = 1 + z_2 + \varepsilon \geq 1$$

Hence $\|x\|_1, \|y\|_1 \geq 1$. Thus, $z$ is a convex sum of two elements $x, y \in \mathcal{Q}(n)$ with $x, y \neq z$ and hence $z$ cannot be an extreme point of $\mathcal{Q}(n)$.

Last suppose $z \in \mathcal{Q}(n)$ is such that $\|z\|_1 = 1$. It will then be shown if there does not exist $\mathcal{I} \subseteq \{1, \ldots, n\}$ such that $z = e_{\mathcal{I}}$ then $z$ cannot be an extreme point of $\mathcal{Q}(n)$. In particular notice, under this condition, necessarily $z_i < 1$ for all $i$. To see this notice if there exists $\ell$ such that $z_\ell = 1$ then $z_i \geq 0$ for all $i$ and $\|z\|_1 = 1$ imply $z = e_{\{\ell\}}$, a contradiction. Thus $z_i < 1$ for all $i$. Furthermore, there must exist at least two distinct indices $j$ and $k$ such that $z_j > 0$ and $z_k > 0$. To see why notice if there are no such indices then $\|z\|_1 = 0$ contradicting the fact that $\|z\|_1 = 1$. Next, if there is one such index $j$ then $\|z\|_1 = 1$ implies $z_j = 1$ which has already been shown to form a contradiction.

Thus without loss of generality, suppose $0 < z_1, z_2 < 1$. Now define

$$\varepsilon = \frac{1}{2} \min \{z_1, 1 - z_1, z_2, 1 - z_2\}$$

and define $x$ as $x_1 = z_1 - \varepsilon$ and $x_2 = z_2 + \varepsilon$ with $x_i = z_i$ for $i = 3, \ldots, n$, and define $y$ as $y_1 = z_1 + \varepsilon$ and $y_2 = z_2 - \varepsilon$ with $y_i = z_i$ for $i = 3, \ldots, n$. Then, by the construction of $\varepsilon$, it follows that $0 < z_1 \pm \varepsilon < 1$ and $0 < z_2 \pm \varepsilon < 1$. Hence $0 \leq x_i \leq 1$ for all $i$ and $0 \leq y_i \leq 1$ for all $i$. Also,

$$\|x\|_1 = \sum_{i=1}^{n} x_i = x_1 + x_2 + \sum_{i=3}^{n} x_i = z_1 - \varepsilon + z_2 + \varepsilon + \sum_{i=3}^{n} z_i = \sum_{i=1}^{n} z_i = 1$$
$$\|y\|_1 = \sum_{i=1}^{n} y_i = y_1 + y_2 + \sum_{i=3}^{n} y_i = z_1 + \varepsilon + z_2 - \varepsilon + \sum_{i=3}^{n} z_i = \sum_{i=1}^{n} z_i = 1$$

Thus $x, y \in \mathcal{Q}(n)$. Last,

$$\frac{1}{2}x_1 + \frac{1}{2}y_1 = \frac{1}{2}(z_1 - \varepsilon) + \frac{1}{2}(z_1 + \varepsilon) = z_1$$
$$\frac{1}{2}x_2 + \frac{1}{2}y_2 = \frac{1}{2}(z_2 + \varepsilon) + \frac{1}{2}(z_2 - \varepsilon) = z_2$$

and $\frac{1}{2}x_i + \frac{1}{2}y_i = \frac{1}{2}z_i + \frac{1}{2}z_i = z_i$ for $i = 3, \ldots, n$. Thus $z = \frac{1}{2}x + \frac{1}{2}y$, a convex combination of $x, y \in \mathcal{Q}(n)$ with $x, y \neq z$ and hence $z$ is not an extreme point of $\mathcal{Q}(n)$.

Therefore all possibilities have been covered and thus it has been shown that if $z \in \mathcal{Q}(n)$ is such that there does not exist $\mathcal{I} \subseteq \{1, \ldots, n\}$ such that $z = e_{\mathcal{I}}$ then $z$ cannot be an extreme

point of $\mathcal{Q}(n)$. That is, forming the contrapositive, if $z$ is an extreme point of $\mathcal{Q}(n)$ then $z = e_{\mathcal{I}}$ for some $\mathcal{I} \subseteq \{1, \ldots, n\}$.

□

Thus if $z$ is an extreme point of $\mathcal{Q}(n)$ then it must be of the form $e_{\mathcal{I}}$ for some $\mathcal{I} \subseteq \{1, \ldots, n\}$. The next result shows that for any $y \in \mathbb{R}^n$ the function $L_y$ is nonnegative at the extreme points of $\mathcal{Q}(n)$.

**Proposition 8.5.3.** *For a positive integer $n$ fix $y \in \mathcal{Q}(n)$ and let $\mathcal{I} \subseteq \{1, \ldots, n\}$. Then $L_y(e_{\mathcal{I}}) \leq 0$.*

**Proof.**

The fact that $y \in \mathcal{Q}(n)$ implies $\|y\|_\infty \leq 1$ and $\|y\|_1 \geq 1$. Thus, by direct calculation,

$$\begin{aligned}
L_y(e_{\mathcal{I}}) &= \sum_{i \in \mathcal{I}} y_i + \sum_{i \in \mathcal{I}} |1 - y_i| + \sum_{i \notin \mathcal{I}} |0 - y_i| - |\mathcal{I}| \sum_{i=1}^{n} y_i - 1 \\
&= \sum_{i \in \mathcal{I}} y_i + \sum_{i \in \mathcal{I}} |1 - y_i| + \sum_{i \notin \mathcal{I}} y_i - |\mathcal{I}| \cdot \|y\|_1 - 1 \\
&= \sum_{i \in \mathcal{I}} y_i + \sum_{i \in \mathcal{I}} (1 - y_i) + \sum_{i \notin \mathcal{I}} y_i - |\mathcal{I}| \cdot \|y\|_1 - 1 \\
&\qquad\qquad \text{since } \|y\|_\infty \leq 1 \text{ implies } 1 - y_i \geq 0 \\
&= \sum_{i \in \mathcal{I}} y_i + |\mathcal{I}| - \sum_{i \in \mathcal{I}} y_i + \sum_{i \notin \mathcal{I}} y_i - |\mathcal{I}| \cdot \|y\|_1 - 1 \\
&= \sum_{i \in \mathcal{I}} y_i + \sum_{i \notin \mathcal{I}} y_i + |\mathcal{I}| - \sum_{i \in \mathcal{I}} y_i - |\mathcal{I}| \cdot \|y\|_1 - 1 \\
&= \sum_{i=1}^{n} y_i + |\mathcal{I}| - \sum_{i \in \mathcal{I}} y_i - |\mathcal{I}| \cdot \|y\|_1 - 1 \\
&= \|y\|_1 + |\mathcal{I}| - \sum_{i \in \mathcal{I}} y_i - |\mathcal{I}| \cdot \|y\|_1 - 1 \\
&= \|y\|_1 + |\mathcal{I}| - \sum_{i \in \mathcal{I}} y_i - (|\mathcal{I}| - 1 + 1) \|y\|_1 - 1 \\
&= \|y\|_1 + |\mathcal{I}| - \sum_{i \in \mathcal{I}} y_i - (|\mathcal{I}| - 1) \|y\|_1 - \|y\|_1 - 1 \\
&= \|y\|_1 - \|y\|_1 + |\mathcal{I}| - \sum_{i \in \mathcal{I}} y_i - (|\mathcal{I}| - 1) \|y\|_1 - 1
\end{aligned}$$

$$
\begin{aligned}
&= |\mathcal{I}| - \sum_{i\in\mathcal{I}} y_i - (|\mathcal{I}| - 1)\,\|y\|_1 - 1 \\
&= |\mathcal{I}| - 1 - \sum_{i\in\mathcal{I}} y_i - (|\mathcal{I}| - 1)\,\|y\|_1 \\
&= (|\mathcal{I}| - 1) - (|\mathcal{I}| - 1)\,\|y\|_1 - \sum_{i\in\mathcal{I}} y_i \\
&= (|\mathcal{I}| - 1)(1 - \|y\|_1) - \sum_{i\in\mathcal{I}} y_i \\
&\leq 0
\end{aligned}
$$

since $|\mathcal{I}| \geq 1$ implies $|\mathcal{I}|-1 \geq 0$ but $\|y\|_1 \geq 1$ implies $1-\|y\|_1 \leq 0$. Hence $(|\mathcal{I}|-1)(1-\|y\|_1) \leq 0$ and clearly $-\sum_{i\in\mathcal{I}} y_i \leq 0$ since $y_i \geq 0$ for all $i$.

□

Thus it has been shown that $L_y(\cdot)$ is convex on $\mathcal{Q}(n)$ for any for any $y \in \mathbb{R}^n_+$. Furthermore, $\mathcal{Q}(n)$ is compact, convex, and has a finite number of extreme points and $L_y$ is nonnegative at these extreme points. This provides enough information to prove, in the next theorem, that $K(\cdot,\cdot) \leq 0$ on $\mathcal{Q}(n)$. The following is the main result of this section.

**Theorem 8.5.4.** *If $x, y \in \mathbb{R}^n$ such that $\|x\|_1, \|y\|_1 \geq 1$ and $\|x\|_\infty, \|y\|_\infty \leq 1$ then*

$$\langle |x|, |y| \rangle + \langle \operatorname{Sgn}(x), y \rangle - \langle x, \operatorname{Sgn}(y) \rangle \leq 1 + \|x\|_1 \cdot \|y\|_1$$

*In particular, the above inequality holds for any $x, y \in \mathbb{S}^{n-1}$.*

**Proof.**

Fix $y \in \mathbb{R}^n$ such that $\|y\|_1 \geq 1$ and $\|y\|_\infty \leq 1$ and consider $L_{|y|}(x)$ on $\mathbb{R}^n_+$. By Proposition (8.5.1) $L_{|y|}(x)$ is convex on $\mathcal{Q}(n)$. Next, $\mathcal{Q}(n)$ is non-empty, compact, and convex and has a finite number of extreme points by Proposition (8.5.2). Therefore by Proposition (4.2.3), $L_{|y|}(x)$ must attain its maximum on $\mathcal{Q}(n)$ at an extreme point of $\mathcal{Q}(n)$. Next, by Proposition (8.5.2), if $z$ is an extreme point of $\mathcal{Q}(n)$ then $z = e_{\mathcal{I}}$ for some $\mathcal{I} \subseteq \{1, \ldots, n\}$. Furthermore, by Proposition (8.5.3), $L_{|y|}(e_{\mathcal{I}}) \leq 0$ for all $\mathcal{I} \subseteq \{1, \ldots, n\}$. Hence

$$\max_{x\in\mathcal{Q}(n)} L_{|y|}(x) = \max_{\mathcal{I}\subseteq\{1,\ldots,n\}} L_{|y|}(e_{\mathcal{I}}) \leq 0$$

Therefore $L_{|y|}(x) \leq 0$ for all $x \in \mathcal{Q}(n)$.

Now given $x \in \mathbb{R}^n$ such that $\|x\|_1 \geq 1$ and $\|x\|_\infty \leq 1$ notice $|x| \in \mathcal{Q}(n)$. Thus $L_{|y|}(|x|) \leq 0$. Next, by the construction of $L_{|y|}(\cdot)$ and $K(\cdot,\cdot)$, one has $L_{|y|}(|x|) = K(|x|,|y|)$. Further by the construction of $K(\cdot,\cdot)$ one has $K(|x|,|y|) = K(x,y)$. Thus $L_{|y|}(|x|) = K(x,y)$. Next, $J(u,v) \leq K(u,v)$ for any $u, v \in \mathbb{R}^n$. In particular, $J(x,y) \leq K(x,y) = L_{|y|}(|x|) \leq 0$. Last notice

$$\langle |x|, |y| \rangle + \langle \mathrm{Sgn}(x), y \rangle - \langle x, \mathrm{Sgn}(y) \rangle \leq 1 + \|x\|_1 \cdot \|y\|_1$$

if and only if

$$\langle |x|, |y| \rangle + \langle \mathrm{Sgn}(x), y \rangle - \langle x, \mathrm{Sgn}(y) \rangle - \|x\|_1 \cdot \|y\|_1 - 1 \leq 0$$

if and only if $J(x,y) \leq 0$ since $J(x,y)$ was defined to be the left hand side of the above inequality.

□

The above result shows that every real orthogonal matrix $B$ satisfies the near necessary conditions that zero is a global minimizer of $G_B$.

# BIBLIOGRAPHY

[1] Baraniuk, R. (2007). Compressive sensing [lecture notes]. *Signal Processing Magazine, IEEE*, 24(4):118–121.

[2] Bertsekas, D., Nedi, A., Ozdaglar, A., et al. (2003). *Convex analysis and optimization*. Athena Scientific.

[3] Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge Univ Pr.

[4] Candès, E. (2006). Compressive sampling. In *Proceedings oh the International Congress of Mathematicians: Madrid, August 22-30, 2006: invited lectures*, pages 1433–1452.

[5] Candès, E. and Wakin, M. (2008). An introduction to compressive sampling. *Signal Processing Magazine, IEEE*, 25(2):21–30.

[6] Casazza, P. G. (1999). The Art of Frame Theory.

[7] Casazza, P. G., Christensen, O., Lindner, A. M., and Vershynin, R. (2005). Frames and the Feichtinger conjecture. *Proc. Amer. Math. Soc.*, 133(4):1025–1033 (electronic).

[8] Casazza, P. G., Fickus, M., Tremain, J. C., and Weber, E. (2006). The Kadison-Singer problem in mathematics and engineering: a detailed account. In *Operator theory, operator algebras, and applications*, volume 414 of *Contemp. Math.*, pages 299–355. Amer. Math. Soc., Providence, RI.

[9] Casazza, P. G., Kutyniok, G., Speegle, D., and Tremain, J. C. (2008). A decomposition theorem for frames and the Feichtinger conjecture. *Proc. Amer. Math. Soc.*, 136(6):2043–2053.

[10] Chen, S., Donoho, D., and Saunders, M. (2001). Atomic decomposition by basis pursuit. *SIAM review*, pages 129–159.

[11] Christensen, O. (2008). *Frames and bases: An introductory course*. Birkhauser.

[12] Dantzig, G. B. and Thapa, M. N. (1997). *Linear Programming 1: Introduction*. Springer.

[13] De Souza, P. and Silva, J. (2004). *Berkeley problems in mathematics*. Springer Verlag.

[14] Do, M. N. (2003). A Friendly Guide to the Frame Theory and Its Application to Signal Processing.

[15] Donoho, D. and Elad, M. (2003). Optimally sparse representation in general (nonorthogonal) dictionaries via L1 minimization. *Proceedings of the National Academy of Sciences*, 100(5):2197.

[16] Elad, M. and Bruckstein, A. (2002). A generalized uncertainty principle and sparse representation in pairs of bases. *Information Theory, IEEE Transactions on*, 48(9):2558–2567.

[17] Han, D., Kornelson, K., Larson, D., and Weber, E. (2007). *Frames for undergraduates*, volume 40 of *Student Mathematical Library*. American Mathematical Society, Providence, RI.

[18] Horn, R. and Johnson, C. (1990). *Matrix analysis*. Cambridge Univ Pr.

[19] Kadison, R. and Singer, I. (1959). Extensions of pure states. *American Journal of Mathematics*, 81(2):383–400.

[20] Korte, B. and Vygen, J. (2006). *Combinatorial optimization: theory and algorithms*, volume 21. Springer Verlag.

[21] Lauritzen, N. (2009). Lectures on convex sets. *Notas de aula, Aarhus University: http://home. imf. au. dk/niels/lecconset. pdf*.

[22] Luenberger, D. and Ye, Y. (2008). *Linear and nonlinear programming*, volume 116. Springer Verlag.

[23] Mital, K. (1976). *Optimization Methods in operations research and systems analysis*. John Wiley & Sons.

[24] Natarajan, B. (1995). Sparse approximate solutions to linear systems. *SIAM J. Comput.*, 24(2):227–234.

[25] Rockafellar, R. (1997). *Convex analysis*, volume 28. Princeton Univ Pr.

[26] Royden, H. (1988). *Real analysis*. Prentice Hall, 3rd edition.

[27] Rudin, W. (1991). *Functional Analysis*. McGraw-Hill, Inc., 2nd edition.

[28] Subramani, K. (2001). Fourier-Motzkin Elimination.