Low-Rank Plus Sparse Decompositions of Large-Scale Matrices via Semidefinite Optimization

by

Rui Gong

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Combinatorics and Optimization

Waterloo, Ontario, Canada, 2023

© Rui Gong 2023

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

We study the problem of decomposing a symmetric matrix into the sum of a low-rank symmetric positive semidefinite matrix and a tridiagonal matrix, and a relaxation which looks for symmetric positive semidefinite matrices with small nuclear norms. These problems are generalizations of the problem of decomposing a symmetric matrix into a low-rank symmetric positive semidefinite matrix plus a diagonal matrix and one of its relaxations, the minimum trace factor analysis problem. We also show that for the relaxation of the low-rank plus tridiagonal decomposition problem with regularizations on the tridiagonal matrix, the optimal solution is unique when the nonnegative regularizing coefficient is not 2. Then, given such a coefficient λ 2 R_+ n $f2g$, we consider three problems. The first problem is decomposing a matrix into a low-rank symmetric positive semidefinite matrix and a tridiagonal matrix. The second is to determine the facial structure of $E_n^{\emptyset}$, which is the set of correlation matrices whose absolute values of entries right below and above the diagonal entries are upper bounded by $\lambda/2$. And the third problem is that given strictly positive integers k, n with $n > k$, and points $v_1, \ldots, v_n$ 2 R^k, determine if there exists a centered (degenerate) ellipsoid passing through all these points exactly such that when the points are projected onto the unit ball corresponding to the ellipsoid, for every i, the cosine value of the angle between the projected ith and $(i + 1)$th points is upper bounded by $\lambda/2$ and lower bounded by $\lambda/2$. We then prove that all these three problems are equivalent and when the regularization coefficient λ goes to infinity, we show the equivalence between them and the corresponding properties of the low-rank plus diagonal decomposition problem.

We also provide a sufficient condition on a subspace U for us to find a nonempty face of $E_n^{\emptyset}$ defined by U. By the equivalence above, this is also a sufficient condition for the other two problems.

After that, we prove that the low-rank plus tridiagonal problem can be solved in polynomial time when the rank of the positive semidefinite matrix in the decomposition is bounded above by an absolute constant.

In the end, we consider representing our problem as a conic programming problem and generalizing it to general sparsity patterns.

Acknowledgements

I would like to thank my supervisor, Levent Tunçel, for his continuous guidance, support and patience. He taught me how to be a better, more professional researcher and co-worker. I am grateful for his invaluable comments and advice which inspired and motivated me academically and spiritually.

I thank Walaa Moursi and Stephen Vavasis for taking the time to read the thesis and for their thoughtful comments.

The material in this thesis is based upon research supported in part by Mathematics Faculty Research Chair funds, NSERC Discovery Grants, Sinclair Graduate Scholarship, Math Domestic Graduate Student Award and C&O Graduate Award. The financial support is gratefully acknowledged.

Dedication

This thesis is dedicated to my father Jiandong Gong and my mother Yan Zou for their love and support. I hope you are proud of me.

Table of Contents

List of Figures

Chapter 1

Introduction

We begin with some notations, basic concepts and definitions. We also describe the problems we study with some known results and their applications and motivations. After that, we give an overall map of the rest of the thesis.

For a positive integer n, we denote the set $\{1, \ldots, n\}$ as $[n]$. We refer $\mathbb{R}^n$ as n-dimensional Euclidean space, $\mathbb{R}^n_+$ as the nonnegative orthant of $\mathbb{R}^n$. For $x \in \mathbb{R}^n$, $i \in [n]$, we let x_i denote the ith entry of x. For $X \in \mathbb{R}^{n_1 \times n_2}$, we let X_{ij} denote the ith row, jth column entry of X. When we have $X \in \mathbb{R}^{n_1 \times n_2}$, $v \in \mathbb{R}^{n_2}$, we use $[Xv]_i$ to denote the ith entry of $Xv \in \mathbb{R}^{n_1}$. We let $\mathrm{Null}(X), \mathrm{row}(X) \subseteq \mathbb{R}^{n_2}, \mathrm{col}(X) \subseteq \mathbb{R}^{n_1}$ to denote the nullspace, the row space and the column space of a matrix $X \in \mathbb{R}^{n_1 \times n_2}$ respectively.

A subset C of a vector space is called *convex* if for any $x, y \in C$, the line segment between x, y entirely lies in C. Given a convex set C, an *extreme point* $\hat{x}$ of C is a point in C such that there does not exists $u, v \in C \setminus \{\hat{x}\}$ with $\frac{1}{2}u + \frac{1}{2}v = \hat{x}$. Given a closed convex set C, a *face* F of C is a closed convex subset of C such that if $u, v \in C$, $\alpha \in (0, 1)$ and $\alpha u + (1 - \alpha)v \in F$, then $u, v \in F$. A nonempty face F that is strictly contained in C is called a *proper face*.

A square matrix is called *symmetric* if it is equal to its transpose. A symmetric matrix is *positive semidefinite* if all its eigenvalues are nonnegative. A symmetric matrix is *positive definite* is all its eigenvalues are strictly positive. By $\mathbb{S}^n, \mathbb{S}^n_+$, and $\mathbb{S}^n_{++}$, we denote the sets of symmetric, symmetric positive semidefinite and symmetric positive definite $n \times n$ matrices, respectively. In this thesis, we also use $X \succeq 0, X \succ 0$ to denote $X \in \mathbb{S}^n_+$, $X \in \mathbb{S}^n_{++}$ respectively.

For $X, Y \in \mathbb{R}^{n_1 \times n_2}$, let $\mathrm{tr}(X)$ denote the trace of X, the inner product $\langle X, Y \rangle$ is defined as $\mathrm{tr}(X^\top Y)$. For $x, y \in \mathbb{R}^n$, define the inner product $\langle x, y \rangle$ as $x^\top y$. The default norm we use for $\mathbb{R}^{n_1 \times n_2}$ is the *Frobenius norm*, which is defined as $\|X\|_F := \sqrt{\langle X, X \rangle}$, and the default norm we use for $\mathbb{R}^n$ is the Euclidean norm, $\|x\| := \sqrt{\langle x, x \rangle}$. For $x \in \mathbb{R}^n$, $\|x\|_1 = \sum_{i=1}^n |x_i|$ and $|x| \in \mathbb{R}^n$ is in the form $|x|_i = |x_i|$. We use $\lambda(X)$ to represent the vector of eigenvalues of X, where $\lambda_i(X)$ is the ith largest eigenvalue of X. We also use $\sigma_i(X)$ to present the ith largest singular value of X. Then the *nuclear norm* on $\mathbb{R}^{n_1 \times n_2}$ is defined as

$$\|X\| := \mathrm{tr}(\sqrt{X^\top X}) = \sum_{i=1}^{\min\{n_1, n_2\}} \sigma_i(X).$$

We let $\|X\| := \sigma_1(X)$ denote the spectral norm of X, which is the dual norm of the nuclear norm.

A subset C of a vector space $\mathbb{V}$ is called a *cone* if it is closed under nonnegative scalar multiplication, that is, for every $x \in C$, and $\alpha \geq 0$, $\alpha x \in C$. The *dual cone* of a cone $C \subseteq \mathbb{V}$ defined over a space $\mathbb{V}$ with respect to an inner product $\langle \cdot, \cdot \rangle_{\mathbb{V}} : \mathbb{V} \times \mathbb{V} \to \mathbb{R}$, is denoted as $C^* := \{y \in \mathbb{V} : \langle x, y \rangle_{\mathbb{V}} \geq 0, \forall x \in C\}$, which is the set of elements in $\mathbb{V}$ whose inner product with every element in C is positive. If not specifically mentioned, the space and inner product defining the dual cone will be clear from the context. For example, when we are considering a cone in $\mathbb{R}^n$, its dual cone is defined over $\mathbb{R}^n$ with $\langle x, y \rangle := x^T y$ by default. Similarly, for a cone in $\mathbb{S}^n$, its dual cone is defined over $\mathbb{S}^n$ with $\langle X, Y \rangle := \mathbf{tr}(XY)$ by default. Notice the dual cones of a cone defined over different spaces with different inner products might be different. Consider the nonnegative ray $C := \{x \in \mathbb{R} : x \geq 0\} \subseteq \mathbb{R}$. Its dual cone C^* defined over $\mathbb{R}$ with $\langle x, y \rangle = xy$ is itself. However, if we embed C in $\mathbb{R}^2$, get $C' := \{x \in \mathbb{R}^2 : x_1 \geq 0, x_2 = 0\} \subseteq \mathbb{R}^2$ and consider its dual cone defined over $\mathbb{R}^2$ with $\langle x, y \rangle := x^T y$. Then $C'^* = \{y \in \mathbb{R}^2 : y_1 \geq 0\}$, which is different from the embedding of C^* in $\mathbb{R}^2$. Also, for C', its dual cone defined over $\mathbb{R}^2$ with inner product $\langle x, y \rangle := 2x_1y_1 + 5x_2y_2 - x_1y_2 - x_2y_1$ is $\{y \in \mathbb{R}^2 : 2y_1 - y_2 \geq 0\}$, which is different from C'^*. Note that $\mathbb{S}^n, \mathbb{S}^n_+, \mathbb{S}^n_{++}$ are all cones, and the dual cone of $\mathbb{S}^n_+$ defined over $\mathbb{S}^n$ with respect to $\mathbf{tr}(\cdot)$ is itself.

Let $K \subseteq \mathbb{R}^n$ be a closed convex cone. A convex cone $G \subseteq K$ is a *face* of K if for every $u, v \in K$ such that $(u + v) \in G$, we have $u, v \in G$.

Every nonempty face of $\mathbb{S}^n_+$ is characterized by a unique subspace U of $\mathbb{R}^n$ such that the face F_U is written as [29]:

$$F_U = \{X \in \mathbb{S}^n_+ : \mathbf{Null}(X) \supseteq U\}. \tag{1.0.1}$$

Let $\mathbb{1}$ represent the vector of all ones, and let $\{e_1, \ldots, e_n\}$ be the standard basis of $\mathbb{R}^n$ where $e_1 = \begin{bmatrix}1 & 0 & \ldots & 0\end{bmatrix}^T, \ldots, e_n = \begin{bmatrix}0 & 0 & \ldots & 1\end{bmatrix}^T$.

We call a matrix $X \in \mathbb{R}^{n_1 \times n_2}$ *diagonal* if $X_{ij} = 0$ for every $i \neq j$. We let $\mathbf{Diag}(x)$ denote a matrix in $\mathbb{S}^n$ such that $\mathbf{Diag}(x)_{ij} = x_i$ if $i = j$, and $\mathbf{Diag}(x)_{ij} = 0$ otherwise. By $\mathbf{diag}(X) \in \mathbb{R}^{\min\{n_1, n_2\}}$, we denote a vector with $\mathbf{diag}(X)_i = X_{ii}$. A matrix X is called *tridiagonal* if $X_{ij} = 0$ for every $|i - j| > 1$. Let $\mathbb{T}^n := \{X \in \mathbb{R}^{n \times n} : X_{ij} = 0, \forall |i - j| > 1\}$ denote the cone of *tridiagonal matrices*, and we call a matrix $X \in \mathbb{R}^{n \times n}$ tridiagonal if $X \in \mathbb{T}^n$. Let $E_n := \{X \in \mathbb{S}^n_+ : \mathbf{diag}(X) = \mathbb{1}\}$ be the set of n-by-n *correlation matrices*, and these convex sets are also called *elliptopes*. A matrix X is *lower-triangular* if $X_{ij} = 0$ for every $j > i$. A matrix X is *upper-triangular* if $X_{ij} = 0$ for every $j < i$.

Given a linear map $\mathcal{A} : \mathbb{V} \to \mathbb{W}$, its *adjoint* $\mathcal{A}^*$ is defined as a linear map $\mathcal{A}^* : \mathbb{W} \to \mathbb{V}$ satisfying

$$\langle \mathcal{A}^*(Y), X \rangle = \langle Y, \mathcal{A}(X) \rangle, \forall X \in \mathbb{V}, Y \in \mathbb{W}.$$

A *centered ellipsoid* in $\mathbb{R}^n$ is a set of the form

$$\{x \in \mathbb{R}^n : x^T A x \leq 1\}$$

where $A \in \mathbb{S}^n_+$. Note that this definition allows for ellipsoidal hypercylinders (also called degenerate ellipsoids), in addition to ellipsoids. That is, we only require A to be symmetric positive

semidefinite instead of being symmetric positive definite. We say a centered ellipsoid *passing through* $v \in \mathbb{R}^n$ if $v^\top A v = 1$.

For two matrices $M, N \in \mathbb{R}^{n_1 \times n_2}$, their *Hadamard product* $M \circ N \in \mathbb{R}^{n_1 \times n_2}$ is given by $(M \circ N)_{ij} = M_{ij} N_{ij}$.

For a subspace U of $\mathbb{R}^n$, $U^\perp$ represents its orthogonal complement which is the set of vectors that are orthogonal to every vector in U. Let $\mathbf{P}_U$ denote the projection matrix onto the subspace U.

Given a simple undirected graph $G = ([n], E)$, $|E| = p$, we define a linear map

$$\mathsf{SparseMat}_G : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{S}^n$$

by letting diagonal entries of $\mathsf{SparseMat}_G(u, v)$ to be u_i, and off-diagonal entries corresponding to E to be v_k. That is, $\mathsf{SparseMat}_G(u, v)_{ij} = 0$ if $i \neq j$ and $ij \notin E$. Hence, we require a bijection between $[p]$ and E. We say a matrix $X \in \mathbb{S}^n$ has a *sparsity pattern* G if there exists u, v such that $\mathsf{SparseMat}_G(u, v) = X$.

Theorem 1.0.1 (Cholesky Decomposition Theorem [13])**.** Let $X \in \mathbb{S}^n$. Then

1. $X \succeq 0$ if and only if there exists lower-triangular $L \in \mathbb{R}^{n \times n}$ such that $X = LL^\top$.
2. $X \succ 0$ if and only if there exists lower-triangular and nonsingular $L \in \mathbb{R}^{n \times n}$ such that $X = LL^\top$.

Corollary 1.0.2 (Square-root-free Cholesky Decomposition Theorem [26])**.** Let $X \in \mathbb{S}^n$. Then

1. $X \succeq 0$ if and only if there exists lower-triangular $L \in \mathbb{R}^{n \times n}$ and $d \geq 0$ such that $X = L\,\mathsf{Diag}(d) L^\top$.
2. If $X \succ 0$ if and only if there exists lower-triangular, nonsingular $L \in \mathbb{R}^{n \times n}$ and $d > 0$ such that $X = L\,\mathsf{Diag}(d) L^\top$.

1.1 Semidefinite Programming

Given $C \in \mathbb{S}^n, b \in \mathbb{R}^m$ and a linear transformation $\mathcal{A} : \mathbb{S}^n \to \mathbb{R}^m$, the *semidefinite programming (SDP)* in standard equality form is defined as:

$$\begin{aligned} &\inf \langle C, X \rangle \\ &s.t.\ \mathcal{A}(X) = b \\ &\quad X \succeq 0, \end{aligned} \tag{SDP}$$

and its dual is defined as:

$$\begin{aligned} &\sup \langle b, y \rangle \\ &s.t.\ \mathcal{A}^*(y) + S = C \\ &\quad S \succeq 0, \end{aligned} \tag{SDD}$$

where $\mathcal{A}^* : \mathbb{R}^m \to \mathbb{S}^n$ is the adjoint of $\mathcal{A}$.

For every linear map $\mathcal{A} : \mathbb{V} \to \mathbb{W}$, there exists $A_1, \ldots, A_m \in \mathbb{S}^n$ such that

$$[\mathcal{A}(X)]_i = \langle A_i, X \rangle = \mathrm{tr}(A_i X), \forall i \in [m]$$

and thus $\mathcal{A}^*(y) = \sum_{i=1}^m y_i A_i$. Then, we can rewrite the SDP problems as:

$$\begin{aligned} \inf \ & \mathrm{tr}(C, X) \\ s.t. \ & \mathrm{tr}(A_i X) = b_i, \forall i \in [m] \\ & X \succeq 0, \end{aligned}$$

and its dual is defined as:

$$\begin{aligned} \sup \ & b^\top y \\ s.t. \ & \sum_{i=1}^m y_i A_i + S = C \\ & S \succeq 0. \end{aligned}$$

A *trace constrained SDP* is a special form of SDP with constraints on the trace of the matrix. Define

$$\Omega_n = \{X \in \mathbb{S}_n : \mathrm{tr}(X) = 1 \text{ and } X \succeq 0\}.$$

Given matrices $C, A_1, \ldots, A_d \in \mathbb{S}^n, b \in \mathbb{R}^d$ and a constant $\alpha \in \mathbb{R}_+$ (or $\mathbb{Z}_+$ sometimes), we have the following trace-constrained SDP:

$$\begin{aligned} \min \ & \mathrm{tr}(CX) \\ s.t. \ & \mathrm{tr}(A_i X) = b_i, \text{ for } i = 1, \ldots, d \\ & X \in \alpha \Omega_n. \end{aligned}$$

Given a pair of SDP in standard equality forms, we can determine when the optimal value is attained. Before that, let us consider a useful definition.

Definition 1.1.1. We say (SDP) has a *Slater point* or *satisfies the Slater condition*, if there exists a $X \in \mathbb{S}^n_{++}$ such that $\mathcal{A}(X) = b$. We say (SDD) has a *Slater point* or *satisfies the Slater condition*, if there exists a $y \in \mathbb{R}^n$ and $S \in \mathbb{S}^n_{++}$ such that $\mathcal{A}^*(y) + S = C$.

Now, we provide a theorem which characterizes the optimal value of an SDP given conditions on the optimal value and Slater point of its dual.

Theorem 1.1.2 (Strong Duality Theorem [29])**.** Suppose (SDD) has a Slater point and the objective value of (SDD) is bounded from above. Then (SDP) attains its optimal value and the optimal values of (SDP) and (SDD) are the same.

Now, we consider the case that both the primal and the dual have Slater points.

Theorem 1.1.3. [29] If both (SDP) and (SDD) have Slater points, then they both attain their optimal values and their optimal values are equal. Also, $X, (y, S)$ are optimal for (SDP) and (SDD) respectively if and only X is feasible for (SDP), (y, S) is feasible for (SDD) and $XS = 0$.

1.2 Convex Programming in Conic Form

A *convex programming in conic form* problem is a convex optimization problem written as minimizing a convex function over the intersection of an affine subspace and a convex cone. In this thesis, we only consider the convex programmings in conic form with a linear objective function and write them as:

$$\begin{aligned} &\inf \langle c, x\rangle \\ &s.t.\ A(x) \ _G\ b \\ &\quad x \ _K\ 0, \end{aligned} \tag{ConicP}$$

where $A : W_1 \to W_2$ is a given linear map from a vector space W_1 to another vector space W_2, and $c \in W_1$, $b \in W_2$ are given. Also, $G \subseteq W_1$, $K \subseteq W_2$ are two closed convex cones. For any closed convex cone G, we use $a \ _G\ b$ to denote $a - b \in G$. Then the corresponding dual of this problem is defined as:

$$\begin{aligned} &\sup \langle b, y\rangle \\ &s.t.\ A^*(y) \ _K\ c \\ &\quad y \ _G\ 0. \end{aligned} \tag{ConicD}$$

For the SDP in standard equality forms, we have $K = S^n_+$ and $G = \{0\} \subseteq R^m$, where $K = S^n_+$ is defined with the inner product $\langle X, Y\rangle := \mathrm{tr}(X^\top Y)$ and $G = R^m$ is defined with the inner product $\langle x, y\rangle := x^\top y$.

Now we define the relative interior points of a set and consider the Slater conditions and Strong Duality Theorem for convex programming in conic form.

Definition 1.2.1. Given a set S, its *relative interior* is defined as

$$\mathrm{relint}(S) := \{x \in S : \exists \epsilon > 0 \text{ such that } B_\epsilon(x) \cap \mathrm{a}(S) \subseteq S\},$$

where $\mathrm{a}(S)$ is the affine hull of S and $B_\epsilon(x)$ is a ball centered at x with radius ϵ.

Definition 1.2.2. We say (ConicP) has a *Slater point* or *satisfied the Slater condition*, if there exists a $x \in \mathrm{relint}(K)$ such that $A(x) - b \in \mathrm{relint}(G)$. We say (ConicD) has a *Slater point* or *satisfied the Slater condition*, if there exists a $y \in \mathrm{relint}(G^*)$ such that $c - A^*(y) \in \mathrm{relint}(K^*)$.

Now, we provide a theorem which characterizes the optimal value of a convex programming in conic form given conditions on the optimal value and Slater point of its dual.

Theorem 1.2.3 (Strong Duality Theorem [16])**.** Suppose (ConicD) has a Slater point and the objective value of (ConicD) is bounded from above. Then (ConicP) attains its optimal value and the optimal values of (ConicP) and (ConicD) are the same.

Now, we consider the case that both the primal and the dual have Slater points.

Theorem 1.2.4. [16] Assume that both (ConicP) and (ConicD) have Slater points. Then they both attain their optimal values and their optimal values are equal. Also, x, y are optimal for (ConicP) and (ConicD) respectively if and only x is feasible for (ConicP), y is feasible for (ConicD) and $\langle \mathcal{A}(x) \quad b, y \rangle = 0$ and $\langle x, c \quad \mathcal{A}^*(y) \rangle = 0$.

Notice that, we have the Strong Duality Theorem holds when the Slater conditions for the primal and the dual are satisfied, then both the primal and dual problems attain their optimal solutions and they are equal. Hence, if we consider the complementary slackness conditions, we have

$$\langle \mathcal{A}(x) \quad b, y \rangle + \langle x, c \quad \mathcal{A}^*(y) \rangle = \langle c, x \rangle \quad \langle b, y \rangle = 0$$

which is equivalent to

$$\langle \mathcal{A}(x) \quad b, y \rangle = 0 \text{ and } \langle x, c \quad \mathcal{A}^*(y) \rangle = 0$$

under cone constraints.

1.3 Affine Rank Minimization and Computational Complexity

In this section, we introduce some problems related to and motivating the thesis. We first introduce the *affine rank minimization problem* which aims to minimize the rank of a matrix over an affine space, and then we show that this problem is NP-hard. After that, we introduce some relaxations of it and their applications.

1.3.1 Computational Complexity

We begin with a special case of *affine rank minimization problem*, then we provide the general problem. We also prove the NP-hardness of the special case which automatically proves the NP-hardness for the general problem.

Let $\rho(x)$ be the number of nonzeros in the vector $x \in \mathbb{R}^n$. Consider the problem: given $A \in \mathbb{R}^{n_1 \quad n_2}, b \in \mathbb{R}^{n_1}$,

$$\begin{aligned} &\min \rho(x) \\ &s.t.\ Ax = b. \end{aligned}$$

This is called the *vector cardinality minimization problem (VCM)*, which aims to find the sparsest vector in an affine space and it is known to be an NP-hard problem [20]. This result is proven by a reduction from *exact cover by 3 sets problem (X3C)*. Given a set S, and a collection C of 3-subsets (subsets with cardinality 3) of S, X3C determines if there is a subset $\hat{C}$ of C such that every element of S appears exactly once in $\hat{C}$. X3C is shown to be equivalent to a special case of the *3 dimensional matching problem (3DM)*, and that case of 3DM is proven to be NP-complete by reduction from the famous *3-satisfiability problem (3SAT)* [11].

Proposition 1.3.1. The VCM is NP-hard.

Proof. We show that X3C is reducible to VCM [20], which shows that VCM is NP-hard. Given an instance of X3C with $S := \{s_1, \ldots, s_{n_1}\}, C := \{c_1, \ldots, c_{n_2}\}$ and $|c_i| = 3, \forall i \in [n_2]$, without loss of generality, assume n_1 is a multiple of 3. Define $A \in \mathbb{R}^{n_1 \times n_2}$ where $A_{ij} = 1$ if $s_j \in C_i$ and $A_{ij} = 0$ otherwise. Let b be $\mathbb{1} \in \mathbb{R}^{n_1}$.

We show that the VCM with given A, b has a solution with at most $n_1/3$ nonzero entries if and only if the given X3C has a solution. If the VCM has a solution x with $n_1/3$ or fewer nonzero entries, then $Ax = b = \mathbb{1}$. By definition, each column of A has three nonzero entries, so x has at least $n_1/3$ nonzero entries. That is, x has exact $n_1/3$ nonzero entries. Let $\hat{C} := \{c_i : x_i \neq 0\}$. Since $\hat{C}$ covers S and its size is $n_1/3$, so it is an exact cover of S. If the given X3C has a solution $\hat{C}$, then let $x \in \mathbb{R}^{n_2}$ be the indicator vector of the solution, that is, if $c_i \in \hat{C}$, $x_i = 1$, otherwise $x_i = 0$. Then $Ax = \mathbb{1} = b$ by the definition of X3C, so VCM has a solution x with $\rho(x) = n_1/3$. □

When we generalize the vector variable in the problem to a matrix variable, we get the general rank minimization problem over an affine space, which is called the *affine rank minimization problem* [21]:

$$\begin{aligned} \min\ & \mathrm{rank}(X) \\ s.t.\ & \mathcal{A}(X) = b, \\ & X \in \mathbb{R}^{n_1 \times n_2}, \end{aligned}$$

where the vector $b \in \mathbb{R}^p$ and the linear map $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^p$ are given . This problem is also NP-hard because the VCM problem is reducible to it. We can see this by restricting X to be a diagonal matrix represented by $\mathrm{Diag}(x)$, so the rank of $\mathrm{Diag}(x)$ is the number of nonzero diagonal entries, $\rho(x)$. Then consider the linear map $\mathcal{A}$ such that $[\mathcal{A}(\mathrm{Diag}(x))]_i = \mathrm{tr}(\mathrm{Diag}(d_i)\,\mathrm{Diag}(x))$ for some given $d_i \in \mathbb{R}^n$, where $\mathrm{tr}(X)$ is the trace of a matrix X. Then $\mathcal{A}(\mathrm{Diag}(x)) = b$ is equivalent to $Ax = b$, where $A = \begin{bmatrix} d_1 & \ldots & d_p \end{bmatrix}^\top$. Then, the problem

$$\begin{aligned} \min\ & \rho(x) \\ s.t.\ & Ax = b, x \in \mathbb{R}^n \end{aligned}$$

is equivalent to

$$\begin{aligned} \min\ & \mathrm{rank}(X) \\ s.t.\ & \mathcal{A}(X) = b, \\ & X \quad \mathrm{Diag}(\mathrm{diag}(X)) = 0, \\ & X \in \mathbb{R}^{n \times n}, \end{aligned}$$

where the map

$$\begin{aligned} \mathcal{A}^0 : \mathbb{R}^{n \times n} &\to \mathbb{R}^p \quad \mathbb{R}^{n \times n}, \\ \mathcal{A}^0(X) &= \begin{bmatrix} \mathcal{A}(X) \\ X \quad \mathrm{Diag}(\mathrm{diag}(X)) \end{bmatrix} \end{aligned}$$

is a linear map. Hence, every VCM can be written in the form of an affine rank minimization problem, which implies that the general affine rank minimization problem is NP-hard.

1.3.2 Matrix Completion Problem

A special case of affine rank minimization problem is *Matrix Completion Problem*. It aims to recover an $n_1 \quad n_2$ matrix M when only some (say m) of its entries are observed. We assume M to be a low-rank matrix, otherwise, say M is an $n \quad n$ matrix with linearly independent columns, then it is impossible to recover it without receiving more information. For example, consider recovering a rank 2 matrix $\begin{bmatrix} 1 & a \\ b & 1 \end{bmatrix}$, there are infinitely many choices for a, b. Also, if the observed entries are not uniformly distributed, we are not able to recover the matrix. For example, consider $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$, there are infinitely many choices for the last column and row. When enough entries are observed and sufficiently uniformly distributed, we consider the main problem

$$\begin{aligned} &\min \ \mathrm{rank}(X) \\ &s.t.\ X_{ij} = M_{ij}, (i,j) \ 2 \quad , \end{aligned} \tag{1.3.1}$$

where is the set of indices of the observed entries.

This problem is a special case of the affine rank minimization problem and it is also NP-hard in general, see Theorem 7 in [25]. We give some known related problems, which might be solved efficiently.

1. First, in many papers (for example, [3]), the following relaxation was introduced:

$$\begin{aligned} &\min \ kXk \\ &s.t.\ X_{ij} = M_{ij}, (i,j) \ 2 \quad , \end{aligned} \tag{1.3.2}$$

   where $X \ 2 \ \mathsf{R}^{n_1 \ n_2}$. Notice that the singular values are all nonnegative and a matrix with rank r has r non-zero singular values. In this way, the objective function can be viewed as a convexification because the nuclear norm is the convex envelope of the rank function within the unit ball of matrices with respect to spectral norm [9]. Note if we restrict X to be symmetric, then the nuclear norm can be written as $kXk = \sum_{i=1}^{n} j\lambda_i(X)j = k\lambda(X)k_1$. When X is required to be symmetric and positive semidefinite, $kXk = \mathrm{tr}(X)$, then (1.3.2) is in the standard form of SDP.

2. The rank( ) function can also be put in the constraints. That is,

$$\begin{aligned} &\min \ \alpha \\ &s.t.\ X_{ij} = M_{ij}, (i,j) \ 2 \\ &\quad \mathrm{rank}(X) \quad \alpha, \end{aligned} \tag{1.3.3}$$

or we can let the objective function be a constant and find feasible solutions for fixed small α. For this problem, the key is whether it is possible to do relaxations for rank(X), especially when the underlying matrix M is a positive semidefinite matrix (or we define X over the positive semidefinite cone). A possible relaxation is changing rank(X) to kXk , then this problem becomes a trace-constrained SDP problem when X is required to be symmetric positive semidefinite. Further note that such an SDP is a convex relaxation of (1.3.3) where the objective function is kept the same and nonconvex feasible region is replaced by a convex set containing it.

There are many possible applications of the matrix completion problem. Here, we mention just two:

1. The Netflix problem [33]: Netflix provided a data set of ratings of 17779 movies given by 48189 users. Each rating was an integer from 1 to 5. People were given parts of the data and asked to predict the missing entries. This data set could be put into a matrix where users are treated as rows of the data matrix and movies as the columns of the matrix. In this way, this problem becomes a matrix completion problem where the given ratings are the given entries, and the rest are the missing ones. It looks for a low-rank matrix because a low-rank matrix can be stored more efficiently and it is believed that only a few factors contribute to a person's movie preferences.

2. Triangulation from incomplete data [3]: Suppose we have some sensors randomly distributed in a region. Each sensor can only estimate the distance from its nearest fellow sensors by signal strength. Then a partially observed distance matrix can be formed. We can estimate the complete distance matrix whose rank will be two if the sensors are in a plane or three if in a 3-dimensional space [27, 19]. For example, we can let V be the set of sensors and consider $X_i \in \mathbb{R}^{|V| \times 2}$ for every $i \in \{1, \ldots, |V|\}$, and $[X_i]_{j1} = (v_{i1} \quad v_{j1})$, $[X_i]_{j2} = (v_{i2} \quad v_{j2})$ for every $j \in \{1, \ldots, |V|\}$ and then $U = [X_1^\top \ \ldots \ X_{|V|}^\top]^\top \in \mathbb{R}^{|V|^2 \times 2}$ has rank at most 2 and the diagonal entries of $UU^\top$ describe the squares of distances of each pair of sensors. In this case, we only need to a few entries to reconstruct the original distances matrix. When we are reconstructing the original distances matrix $UU^\top$, we aim to find a matrix with rank as small as possible. If we find a matrix with rank 2, then it is very likely we recover the original matrix.

1.4 Minimum Trace Factor Analysis (MTFA) problem

We consider *minimum trace factor analysis (MTFA)* problem in this section. It aims to minimize the trace when we modify the diagonal entries of a given matrix and keep the resulting matrix positive semidefinite. MTFA is a famous and tractable problem which has been studied for almost a century [18]. We show how it is related to *low-rank plus sparse decomposition* problem and how it motivates other problems we consider in this thesis. Consider an MTFA:

$$\begin{aligned} &\min_{x \in \mathbb{R}^n} \ \mathbb{1}^\top x \\ &s.t. \quad {}_0 + \mathrm{Diag}(x) \quad 0, \end{aligned}$$

where $\quad_0$ is a given matrix in $\mathbb{S}^n$. Note MTFA is an SDP. Della Riccia and Shapiro [7, Theorem 4] showed that when MTFA has a solution, it has a unique optimal solution. Some interpretations of this problem are from the *Factor Analysis Model* in Statistics [1], which was studied since 1904 [28]. Factor Analysis Model considers $K, M \in \mathbb{R}^{n \quad m}$ such that K_{ij} is the ith observation of the jth individual and $M_{ij} = \mu_i$ is the observation mean of the ith observation, then write $K \quad M = LF + \varepsilon$ where $L \in \mathbb{R}^{n \quad k}, F \in \mathbb{R}^{k \quad m}, \varepsilon \in \mathbb{R}^{n \quad m}$, F is an unknown factor matrix, ε is an error term and then look for the underlying L. By assuming F, ε are independent, $\mathbb{E}(F) = 0$, and $Cov(F) = I$, we have $Cov(K \quad M) = LL^{\top} + Cov(\varepsilon)$ where $Cov(\varepsilon)$ is a diagonal matrix because columns of ε are assumed to be independent error terms. The dual problem of MTFA is

$$\begin{aligned} \max\ & \langle \quad_0, X \rangle \\ s.t.\ & \mathrm{diag}(X) = \mathbb{1}, \\ & X \quad 0. \end{aligned}$$

Note this primal-dual pair of SDP commonly arises in other contexts as well. For example, we consider the *MaxCut* problem in SDP form [12]. Given a simple graph $G := (V, E)$, and a weight vector $w \in \mathbb{R}^E_+$, find a set $U \quad V$ such that

$$\sum_{\{i,j\} \in \delta(U)} w_{ij} \text{ is maximized,}$$

where $\delta(U) := \{\{i, j\} \in E : i \in U, j \in V \setminus U\}$. Then with $n := |V|$, let us represent each cut $(U, V \setminus U)$ by a vector $u \in \{1, \quad 1\}^n$, where

$$u_i = \begin{cases} 1, & \text{if } i \in U, \\ \quad 1, & \text{if } i \notin U. \end{cases}$$

And set $w_{ij} = 0$ for every $\{i, j\} \notin E$, then the MaxCut problem can be represented as:

$$\begin{aligned} \max\ & \frac{1}{4} \sum_{i \in V} \sum_{j \in V} w_{ij}(1 \quad u_i u_j) \\ s.t.\ & u \in \{1, \quad 1\}^n. \end{aligned}$$

If we define $W \in \mathbb{S}^n$ with $W_{ij} := w_{ij}$, then MaxCut problem can be written as

$$\begin{aligned} \max\ & \frac{1}{4} \langle W, \mathbb{1}\mathbb{1}^{\top} \rangle \quad \frac{1}{4} \langle W, X \rangle \\ s.t.\ & \mathrm{diag}(X) = \mathbb{1} \\ & X \quad 0 \\ & \mathrm{rank}(X) = 1 \end{aligned}$$

Note the only nonconvex constraint here is $\mathrm{rank}(X) = 1$, which is a constraint we may relax and we then get:

$$\begin{aligned} \max \quad & \frac{1}{4}\mathrm{tr}(WX) \quad \left(+\frac{1}{4}\mathbb{1}^\top W\mathbb{1}\right) \\ s.t. \ & \mathrm{diag}(X) = \mathbb{1}, \\ & X \succeq 0 \end{aligned}$$

which is a special case of the dual problem of MTFA, and its dual is defined as:

$$\begin{aligned} \min \ & \mathbb{1}^\top y \quad \left(+\frac{1}{4}\mathbb{1}^\top W\mathbb{1}\right) \\ s.t. \ & \mathrm{Diag}(y) + \frac{1}{4}W \succeq 0 \end{aligned}$$

which is a special case of MTFA.

Note MTFA is minimizing $\mathbb{1}^\top x$, in another way, it is minimizing $\mathrm{tr}(\ _0 + \mathrm{Diag}(x))$. Since $_0 + \mathrm{Diag}(x) \succeq 0$, the objective function is equivalent to $\|\ _0 + \mathrm{Diag}(x)\|$, which is the *convex envelope* of $\mathrm{rank}(\ _0 + \mathrm{Diag}(x))$ within the unit ball of matrices with respect to spectral norm [9]. The *convex envelop* $\hat{f}$ of $f : C \to \mathbb{R}$, where C is convex, is defined as

$$\hat{f} := \sup\{g(x) : g \text{ is convex and } g(y) \le f(y), \forall y \in C\}.$$

Now we consider an equivalent form of MTFA, which is

$$\begin{aligned} \min_{x\in\mathbb{R}^n, L\in\mathbb{S}^n} \ & \mathrm{tr}(L) \\ s.t. \ & \ _0 = L + \mathrm{Diag}(x), \\ & L \succeq 0. \end{aligned}$$

Note that this form minimizes $\mathrm{tr}(L) = \mathrm{tr}(\ _0) \quad \mathbb{1}^\top x$ which is equivalent to minimizing $\mathbb{1}^\top x$ in the original form after changing the sign of x, and the constraints are equivalent to the original MTFA. With this form, we can see that it is a relaxation of

$$\begin{aligned} \min_{x\in\mathbb{R}^n, L\in\mathbb{S}^n} \ & \mathrm{rank}(L) \\ s.t. \ & \ _0 = L + \mathrm{Diag}(x), \\ & L \succeq 0. \end{aligned}$$

Since a diagonal matrix is clearly sparse, this is a *low-rank plus sparse decomposition problem*, which aims to write a matrix as a sum of a low-rank symmetric positive semidefinite matrix and a sparse matrix. Results and analysis in [4] can apply to the problem above. We can also interpret MTFA as a positive semidefinite matrix completion problem where the off-diagonal entries are given by $_0$ and diagonal entries are to be determined.

As we will see below, many problems can be relaxed as an MTFA problem. Another equivalent form of (1.3.1) is discussed in "Statistical inference of semidefinite programming" by Alexander

Shapiro [24]:

$$\begin{aligned} \min_{X \in \mathcal{V}} \ & \operatorname{rank}(M + X) \\ s.t. \ & M + X \succeq 0, \end{aligned} \tag{1.4.1}$$

where $M \in \mathbb{S}^p$, $p = n_1 + n_2$, and it is in the form $M = \begin{bmatrix} 0 & M^0 \\ M^{0\top} & 0 \end{bmatrix}$, where $M^0 \in \mathbb{R}^{n_1 \times n_2}$, $M^0_{ij} = M_{ij}, \forall (i,j) \in \Omega$ and 0 otherwise. Here,

$$\mathcal{V} := \{X \in \mathbb{S}^p : X_{ij} = 0, (i,j) \in \Omega^0\}$$

where $\Omega^0 := \bigcup_{(i,j)\in\Omega} \{(i, n_1 + j), (n_1 + j, i)\}$ is the set of indices corresponding to the known entries of M.

Proof of equivalence between (1.3.1) *and* (1.4.1) *[24].* First, consider an arbitrary solution Y of (1.3.1) with rank r. By singular value decomposition, it can be written as $Y = VW^\top$, where V, W are matrices with orders $n_1 \times r, n_2 \times r$ and a common rank r. Then consider

$$X := UU^\top \succeq M, \text{ where } U := \begin{bmatrix} V \\ W \end{bmatrix}, \text{ i.e. } X = \begin{bmatrix} VV^\top & Y \\ (Y & M^0)^\top \end{bmatrix} \quad \begin{matrix} M^0 \\ WW^\top \end{matrix}$$

This matrix X is feasible for (1.4.1) and $\operatorname{rank}(X + M) = \operatorname{rank}(UU^\top) = r$. Hence, the optimal value of (1.4.1) is less than or equal to the optimal value of (1.3.1).

Now let X be a feasible point of problem (1.4.1) and $r = \operatorname{rank}(X + M)$, then $X + M = UU^\top$ for some $p \times r$ matrix U of rank r. Partition U into V, W as above, then $Y = VW^\top$ is feasible for (1.3.1) with $\operatorname{rank}(Y) \le r$ because V has r columns. That is, the optimal value of (1.3.1) is less than or equal to the optimal value of (1.4.1). So the optimal values of these two problems are equal. □

Also the above transformation can be approximated by

$$\begin{aligned} \min_{X \in \mathcal{V}} \ & \operatorname{tr}(X) \\ s.t. \ & M + X \succeq 0 \end{aligned} \tag{1.4.2}$$

by relaxing $\operatorname{rank}(M + X)$ to $\operatorname{tr}(M + X) = \operatorname{tr}(X)$. And when X is restricted to being a diagonal matrix, this problem becomes an MTFA problem.

We have seen the low-rank plus sparse decomposition problem. It provides a more efficient way to store the information and also helps us to interpret the given data matrix. In Chapter 2, we extend the low-rank plus diagonal decomposition problem to the low-rank plus tridiagonal decomposition problem, which allows us to control the entries right above and under the diagonal entries. For some applications, low-rank plus diagonal decompositions use the low-rank matrix to represent the main factors and the diagonal matrix to present some noises or variables which are assumed to be independently and identically distributed. For this setting, tridiagonal generalization allows the ith variable to be possibly dependent on the $(i-1)$th and/or $(i+1)$th variables but none

of the others. This structure arises in applications like time-dependent models or truss design (as a rough approximation). Also for some applications, it might be reasonable to give different weights to the diagonal entries of the tridiagonal matrix and the other entries of it.

There are some algorithms solving some problems we mentioned above exactly. For example, in the paper "Exact Matrix Completion via Convex Optimization" by Emmanuel J. Candès and Benjamin Recht [3], it was proven that the Matrix Completion problem can be solved exactly under some conditions with high probability. Also, for general large-scale and weakly constrained SDPs, Yurtsever et. al provided a provably correct randomized algorithm which solves them efficiently [34].

The structure of the remainder of this thesis is as follows. In Chapter 2, we introduce the low-rank plus diagonal decomposition, one of its relaxations and the realizability, the recoverability and the ellipsoid fitting property of a subspace. Then we study the low-rank plus tridiagonal decomposition and provide an optimality condition for one of its relaxations. We extend the three properties of the low-rank plus diagonal decomposition to the low-rank plus tridiagonal decomposition. In Chapter 3, we study the coherence of a subspace and provide a condition on it which is sufficient for the realizability of a subspace with respect to the low-rank plus tridiagonal decomposition problem. We then provide some computational examples to verify the condition and study the realizability of one-dimensional subspaces. In Chapter 4, we provide an algorithm to solve the low-rank plus tridiagonal decomposition problem in polynomial time when the optimal value is bounded by an absolute constant. Finally, in Chapter 5, we generalize a relaxation of the low-rank plus tridiagonal decomposition problem to a convex programming in conic form and consider the general low-rank plus sparse matrices decomposition problem. Then we summarize the thesis and discuss some future research directions.

Chapter 2

Subspace Realizability, Recoverability and Ellipsoid Fitting

In the first section of this chapter, we introduce the *diagonal perturbation problem*, which aims to minimize the rank of a positive semidefinite matrix by perturbing its diagonal entries. We then introduce the definitions of *subspace realizability, subspace recoverability, ellipsoid fitting property*, which describe the uniqueness of an optimization problem, characterize faces of the elliptopes and show if the columns of certain matrices can be passed through exactly by an ellipsoid respectively. After that, we show these three definitions are equivalent. This section is an exposition of the results in [22].

In the second section, we move to the *tridiagonal perturbation problem*, which allows us to also perturb the tridiagonal entries while minimizing the rank of the resulting positive semidefinite matrix. Then, we consider a relaxation of the problem with and without a regularization term on the absolute values of the perturbation and develop the conditions for the uniqueness of optimal solutions. After that, we extend the definitions of *subspace realizability, subspace recoverability, ellipsoid fitting property* to the tridiagonal case and prove they are equivalent.

2.1 Diagonal Perturbation

Recall the low-rank plus sparse matrices decomposition problem and some of its applications. For example, after the decomposition of a given matrix A as the sum of a sparse matrix S and a low-rank matrix L, the cost to store the matrix A is reduced by storing the sparse matrix S and the low-rank matrix L instead. Then, the solutions of associate linear systems $Ax = b$ can be computed efficiently by only considering the sparse matrix and a basis for the column space of the low-rank matrix. Also, if the matrix is built from real data or measurements, such as the covariance matrix of a sample, it might give some useful interpretations for the data, like the direction of arrival estimation in [22]. For this section, we focus on the low-rank plus diagonal decomposition, which is discussed when we presented MTFA. The factor analysis model studied by Spearman brings this

decomposition problem [28]. We can write the problem as: given $A \in \mathbb{S}^n$,

$$\begin{aligned} \min_{x \in \mathbb{R}^n, L \in \mathbb{S}^n} \ & \operatorname{rank}(L) \\ s.t. \ & A = L + \operatorname{Diag}(x) \\ & L \succeq 0, \end{aligned}$$

which can be written in an equivalent form, the *diagonal perturbation problem*,

$$\begin{aligned} \min \ & \operatorname{rank}(A + \operatorname{Diag}(y)) \\ s.t. \ & A + \operatorname{Diag}(y) \succeq 0 \\ & y \in \mathbb{R}^n. \end{aligned}$$

and we may and do assume $\operatorname{diag}(A) = 0$ for this form. This form shows that the problem is minimizing the rank of a positive semidefinite matrix when a given matrix A is fixed and a perturbation on the diagonal entries is allowed. Like the matrix completion problem, we may relax this problem by replacing the rank function with the nuclear norm. Then for every feasible solution y, since $A + \operatorname{Diag}(y) \succeq 0$, we have $y \geq 0$ and $\|A + \operatorname{Diag}(y)\|_* = \operatorname{tr}(A + \operatorname{Diag}(y)) = \langle \mathbb{1}, y \rangle$, so we can write the relaxed problem as

$$\begin{aligned} \min \ & \langle \mathbb{1}, y \rangle \\ s.t. \ & A + \operatorname{Diag}(y) \succeq 0 \\ & y \in \mathbb{R}^n. \end{aligned}$$

Note we can write $L = A + \operatorname{Diag}(y) = A - \operatorname{Diag}(-y)$. Then, by replacing $A + \operatorname{Diag}(y)$ in the problem above by $A - \operatorname{Diag}(-y)$ and replacing $\min \langle \mathbb{1}, y \rangle$ by $-\max \langle \mathbb{1}, -y \rangle$, change y to $-y$ and dropping the " $-$ " in the objective function, we obtain a problem with the negative optimal value and the same feasible region:

$$\begin{aligned} \max \ & \langle \mathbb{1}, y \rangle \\ s.t. \ & L + \operatorname{Diag}(y) = A \\ & L \succeq 0 \\ & y \in \mathbb{R}^n \end{aligned} \tag{MTFA}$$

whose dual is defined as

$$\begin{aligned} \min \ & \langle A, X \rangle \\ s.t. \ & \operatorname{diag}(X) = \mathbb{1} \\ & X \succeq 0. \end{aligned} \tag{MTFAD}$$

Notice now, the dual problem (MTFAD) is in the standard equality form of SDP.

With (MTFA), we consider three problems [22]:

1. Suppose $X \in \mathbb{S}^n$ can be written in the form of $X = L + \operatorname{Diag}(y)$, where L is symmetric positive semidefinite. What properties or conditions of (L, y) will ensure that (L, y) is

the unique optimal solution of (MTFA) with the input $A = X$?

2. Recall that F_U in (1.0.1) is defined as the face of S^n_+ defined by a subspace U of R^n. We have that every face of E_n, the set of correlation matrices, is in the form

$$E_n \setminus F_U = \{X \succeq 0 : \text{Null}(X) \supseteq U, \text{diag}(X) = \mathbb{1}\},$$

where U is a subspace of R^n. However, for some subspace U of R^n, the set $E_n \setminus F_U$ is empty. For example, consider $U = \text{span}\{e_1\}$, there is no $X \in E_n$ such that $\text{Null}(X) \supseteq U$, so $E_n \setminus F_U = \emptyset$. Thus, another problem is, which subspaces define a nonempty face of E_n?

3. Consider the lemma:

Lemma 2.1.1. [22] Suppose V is a $k \times n$ matrix with row space V. If there exists a centered ellipsoid in R^k passing through each column (which is a point in R^k) of V, then for every matrix W with row space V, there exists a centered ellipsoid passing through all columns of W.

The above lemma shows that, given $v_1, \ldots, v_n \in R^k$, the row space of the matrix $[v_1, \ldots, v_n]$ decides if it is possible to fit an ellipsoid to $v_1, \ldots, v_n$. Then we consider the problem: for which subspaces V of R^n, do there exist a positive integer k and a $k \times n$ matrix V with row space V and a centered ellipsoid passing through all its columns?

Being motivated by the three problems above, we consider the following three definitions [22]:

Definition 2.1.2.

1. A subspace U of R^n is *diagonally recoverable* by MTFA if for every $y \in R^n$ and every $L \in S^n$ with column space U, (L, y) is the unique optimal solution of (MTFA) with input $A = \text{Diag}(y) + L$.

2. A subspace U of R^n is *diagonally realizable* if there exists a correlation matrix $Q \in E_n$ such that $\text{Null}(Q) \supseteq U$.

3. A subspace V of R^n has the *ellipsoid fitting property* if there exists $V \in R^{k \times n}$ with row space V such that there is a centered ellipsoid in R^k passing through each column of V.

The following proposition shows that these three definitions are equivalent:

Proposition 2.1.3. [22] Let U be a subspace of R^n, then the following are equivalent:

1. U is diagonally recoverable.

2. U is diagonally realizable.

3. $U^\perp$ has the ellipsoid fitting property.

Definition 2.1.4. Given a subspace U of $\mathbb{R}^n$, we say (A_1, y_1, L_1),$(A_1, y_1, L_1) \in \mathbb{S}^n \times \mathbb{R}^n \times \mathbb{S}^n_+$ are *equivalent with respect to* U if $\mathrm{col}(L_1) = \mathrm{col}(L_2) = U$ and $(L_1, y_1), (L_2, y_2)$ are the unique optimal solutions of (MTFA) with input $A_1 = \mathrm{Diag}(y_1) + L_1$, $A_2 = \mathrm{Diag}(y_2) + L_2$ respectively.

In the definition above, each equivalence class is defined by a subspace U of $\mathbb{R}^n$. Given an instance $(A, y, L) \in \mathbb{S}^n \times \mathbb{R}^n \times \mathbb{S}^n_+$ of one such equivalence class, we first obtain the corresponding subspace $U = \mathrm{col}(L)$. Then for every $L \in \mathbb{S}^n_+$ with $\mathrm{col}(L) = U$ and every $y \in \mathbb{R}^n$, we have that $(\mathrm{Diag}(y) + L, y, L)$ is in the equivalence class. For example, if an instance (A, y, L) is given, then $\alpha(A, y, L)$ is also in the equivalence class for every $\alpha > 0$.

Proposition 2.1.5. A subspace U of $\mathbb{R}^n$ is diagonally recoverable if and only if there exists a nonempty equivalence class in $\mathbb{S}^n \times \mathbb{R}^n \times \mathbb{S}^n$ with respect to U.

Proof. Assume U is diagonally recoverable. By the definition of U being diagonally recoverable and picking arbitrary $y \in \mathbb{R}^n$ and positive semidefinite $L \in \mathbb{S}^n$ with column space U, we know (L, y) is the unique optimal solution of (MTFA) with input $A = \mathrm{Diag}(y) + L$. Notice such L exists because we can consider a projection matrix onto $\mathbf{P}_U$. Then we know there exists an equivalence class with respect to U containing $(\mathrm{Diag}(y) + L, y, L)$.

Now assume there exists a nonempty equivalence class with respect to U, then there exists y and L with column space U, such that (L, y) is the unique optimal solution of (MTFA) with input $A = \mathrm{Diag}(y) + L$. Since (MTFA) and (MTFAD) are strictly feasible, by Theorem 1.1.3, we know (y, L) is an optimal solution of (MTFA) if and only if there exists a feasible solution Q of (MTFAD) such that $QL = 0$. Since $\mathrm{col}(L) = U$, we know $U \subseteq \mathrm{Null}(Q)$, so U is diagonally realizable by definition, hence by Prop. 2.1.3, U is diagonally recoverable. □

2.2 Tridiagonal Symmetric Positive Semidefinite Matrices

Being motivated by the diagonal perturbation problem, we now consider the *tridiagonal perturbation problem*, which allows us to perturb more entries. In this section, we introduce some properties of positive semidefinite *tridiagonal matrices*. We then define tridiagonal perturbation problems with and without a regularization term and analyze their optimality conditions. After that, we discuss the recoverability, realizability and ellipsoid fitting property for the tridiagonal case.

Recall that $\mathbb{T}^n$ is the space of n-by-n tridiagonal matrices. Note that $\mathbb{T}^n$ is isomorphic to $\mathbb{R}^{3n-2}$ by mapping all its $3n-2$ possibly nonzero entries to $\mathbb{R}^{3n-2}$. Similarly, $\mathbb{T}^n \setminus \mathbb{S}^n$ is isomorphic to $\mathbb{R}^{2n-1}$. We also define

$$K_n := \left\{ X \in \mathbb{T}^n : \begin{pmatrix} X_{ii} & X_{i(i+1)} \\ X_{(i+1)i} & X_{(i+1)(i+1)} \end{pmatrix} \succeq 0, \forall i \in [n-1] \right\},$$

$$B_i := \{X \in \mathbb{S}^n_+ : X_{jk} = 0, \forall (j,k) \notin \{(i,i), (i,i+1), (i+1,i), (i+1,i+1)\}\}, \forall i \in [n-1].$$

Lemma 2.2.1. Let $X \in (\mathsf{T}^n \setminus \mathsf{S}_+^n)$. Then, there exist $\ell \in \mathsf{R}^{n-1}$ and $d \in \mathsf{R}_+^n$ such that

$$X = \begin{bmatrix} 1 & & & & \\ \ell_1 & 1 & & 0 & \\ & \ell_2 & 1 & & \\ & & \ddots & \ddots & \\ & 0 & & \ell_{n-1} & 1 \end{bmatrix} \begin{bmatrix} d_1 & & & \\ & d_2 & & 0 \\ & & \ddots & \\ & 0 & & d_n \end{bmatrix} \begin{bmatrix} 1 & \ell_1 & & & \\ & 1 & \ell_2 & 0 & \\ & & 1 & \ddots & \\ & & & \ddots & \ell_{n-1} \\ & 0 & & & 1 \end{bmatrix}.$$

Moreover, X can be expressed as $\sum_{i=1}^{n-1} \lambda_i X^{(i)}$ where $\lambda \in \mathsf{R}_+^{n-1}$, $\sum_{i=1}^{n-1} \lambda = 1$, $X^{(i)} \in B_i, i \in \{1, \ldots, n-1\}$.

Proof. By the square-root-free Cholesky Decomposition and Corollary 1.0.2, we can express X as

$$X = LDL^\top,$$

where $D = \mathsf{Diag}(d)$ is a diagonal matrix with $d \geq 0$ and L is a lower-triangular matrix with diagonal entries all being 1. Let L_i represent the ith column of L, then

$$X = \sum_{i=1}^{n} d_i (L_i L_i^\top).$$

Without loss of generality, we assume $d_i \neq 0$ for every $i \in [n]$. We claim L is a tridiagonal matrix (i.e. $L \in \mathsf{T}^n$). We prove it by doing an induction on i, consider $d_1 L_1 L_1^\top$ and without loss of generality, assume $[L_1]_3 \neq 0$, then $X_{31} = d_1 [L_1]_3 \neq 0$, $X \notin \mathsf{T}^n$, then we reach a contradiction, so $L_1 L_1^\top \in B_1$. Consider $L_1 L_1^\top \in B_1$ as the base case and assume that for an $i \in [n]$, $L_j L_j^\top \in B_j$ for all $j < i$. For contradiction, suppose $L_i L_i^\top \notin B_i$, without loss of generality, say $[L_i]_{i+2} \neq 0$. Then, since $L_j L_j^\top \in B_j$ for every $j < i$ and $[L_k]_i = 0$ for every $k > i$, we have $X_{(i+2)i} = d_i([L_i]_{i+2}) \cdot 1 \neq 0$, which contradicts $X \in \mathsf{T}^n$. Hence, when $L_j L_j^\top \in B_j$ for every $j < i$, we have $L_i L_i^\top \in B_i$, so L is a tridiagonal matrix. Hence, we can write X as

$$X = d_1 \begin{bmatrix} 1 \\ \ell_1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \begin{bmatrix} 1 & \ell_1 & 0 & \cdots & 0 \end{bmatrix} + d_2 \begin{bmatrix} 0 \\ 1 \\ \ell_2 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \begin{bmatrix} 0 & 1 & \ell_2 & 0 & \cdots & 0 \end{bmatrix} + \cdots + d_n \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \begin{bmatrix} 0 & \cdots & 0 & 1 \end{bmatrix}.$$

The ith term is in B_i for $i < n-1$. And the sum of the last two terms is

$$d_{n-1} \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ \ell_{n-1} \end{bmatrix} \begin{bmatrix} 0 & \cdots & 0 & 1 & \ell_{n-1} \end{bmatrix} + d_n \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \begin{bmatrix} 0 & \cdots & 0 & 1 \end{bmatrix},$$

which is in $B_{n\ 1}$. Then with $\bar{d} := \sum_{i=1}^{n} d_i$, we have

$$X = \sum_{i=1}^{n\ 2} \frac{d_i}{\bar{d}} (\underbrace{\bar{d} L_i L_i^{>}}_{2B_i}) + \frac{d_{n\ 1} + d_n}{\bar{d}} (\underbrace{\bar{d} (L_{n\ 1} L_{n\ 1}^{>} + L_n L_n^{>})}_{2B_n}).$$

□

Lemma 2.2.1 is a special case of the following more general, well-known fact: all symmetric positive semidefinite matrices with a chordal sparsity pattern admit zero fill-in Cholesky factorization [10] (The above proof of Lemma 2.2.1 can be adopted with simple changes to provide a proof for the general statement).

Next, we show that the cone of symmetric positive semidefinite tridiagonal matrices can be decomposed as a Minkowski sum of 2-by-2 symmetric positive semidefinite matrices which are suitably paddled with zeros. This result can also be generalized to the chordal sparsity pattern.

Proposition 2.2.2.

$$(\mathsf{T}^n \setminus \mathsf{S}_+^n) = B_1 + \quad + B_{n\ 1}$$

where the " + " here represents Minkowski addition.

Proof.

- ( ) By Lemma 2.2.1, we can write any $X \in (\mathsf{T}^n \setminus \mathsf{S}_+^n)$ as $X = \sum_{i=1}^{n\ 1} \lambda_i X^{(i)}$, where $X^{(i)} \in B_i$ and $\sum_{i=1}^{n\ 1} \lambda_i = 1, \lambda \quad 0$. Since B_i are cones, $X \in B_1 + \quad + B_{n\ 1}$.
- ( ) Let $X = X_1 + \quad + X_{n\ 1}$ where $X_i \in B_i$. By the definition of B_i, it is clear that $X \in \mathsf{T}^n \setminus \mathsf{S}^n$. And for every $y \in \mathsf{R}^n$, we have

$$y^{>} X y = y^{>} (X_1 + \quad + X_{n\ 1}) y = \underbrace{y^{>} X_1 y}_{0} + \quad + \underbrace{y^{>} X_{n\ 1} y}_{0} \quad 0,$$

  so $X \in \mathsf{S}_+^n$. That is, $X \in (\mathsf{T}^n \setminus \mathsf{S}_+^n)$.

□

As the next proposition shows, for some choices of the inner product and underlying space, the dual cone of $\mathsf{T}^n \setminus \mathsf{S}_+^n$ contains the primal cone itself.

Proposition 2.2.3.

$$\mathsf{T}^n \setminus \mathsf{S}_+^n \quad K_n = (\mathsf{T}^n \setminus \mathsf{S}_+^n)$$

where the dual cone is defined over the space T^n, and with respect to the trace inner product.

Proof. Let $\overline{X} \in \mathsf{T}^n \setminus \mathsf{S}_+^n$. Then, for every $i \in [n \quad 1]$, the 2-by-2 symmetric submatrix

$$\begin{bmatrix} X_{ii} & X_{i(i+1)} \\ X_{(i+1)i} & X_{(i+1)(i+1)} \end{bmatrix}$$

of $\overline{X}$ is PSD. Therefore, $\overline{X} \in K_n$. Hence $\mathbb{T}^n \cap \mathbb{S}_+^n \subseteq K_n$.

Suppose $\overline{S} \in (\mathbb{T}^n \cap \mathbb{S}_+^n)^*$. Note that, $B_1, \ldots, B_{n-1} \in \mathbb{T}^n \cap \mathbb{S}_+^n$. Since $\overline{S} \in \mathbb{T}^n$ (the dual cone is defined over the space $\mathbb{T}^n$), $\langle \overline{S}, X \rangle \geq 0$ for every $X \in \sum_{i=1}^{n-1} B_i$. Hence, the projection of $\overline{S}$ onto B_i is positive semidefinite for every $i \in [n-1]$, so $\overline{S} \in K_n$ which implies $(\mathbb{T}^n \cap \mathbb{S}_+^n)^* \subseteq K_n$.

Then, we consider $\overline{S} \in K_n$. Let $X \in \mathbb{T}^n \cap \mathbb{S}_+^n$. By the Lemma 2.2.1, we can write $X = \sum_{i=1}^{n-1} \lambda_i X^{(i)}$ where $X^{(i)} \in B_i, \lambda \in \mathbb{R}_+^{n-1}$. Then we have

$$\langle \overline{S}, X \rangle = \sum_{i=1}^{n-1} \underbrace{\lambda_i}_{\geq 0} \underbrace{\langle \overline{S}, X^{(i)} \rangle}_{\geq 0} \geq 0$$

Therefore, $\overline{S} \in (\mathbb{T}^n \cap \mathbb{S}_+^n)^*$ which implies $K_n \subseteq (\mathbb{T}^n \cap \mathbb{S}_+^n)^*$. That is, $K_n = (\mathbb{T}^n \cap \mathbb{S}_+^n)^*$. □

The inclusion above is strict. We can see that from the following example.

Example 2.2.4. Consider a matrix

$$A := \begin{bmatrix} 4 & 4 & 0 \\ 4 & 5 & 5 \\ 0 & 5 & 5 \end{bmatrix} \in K_n$$

but

$$\begin{bmatrix} 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 4 & 4 & 0 \\ 4 & 5 & 5 \\ 0 & 5 & 5 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = -4,$$

which implies $A \notin \mathbb{S}_+^n$.

The dual cone also has a connection to the matrix completion problem.

Lemma 2.2.5. [14] $(\mathbb{T}^n \cap \mathbb{S}_+^n)^*$ is the set of all symmetric tridiagonal matrices which are positive semidefinite completable. That is, given $\overline{S} \in \mathbb{T}^n$,

$$\overline{S} \in (\mathbb{T}^n \cap \mathbb{S}_+^n)^* \text{ if and only if } \exists \hat{S} \in \mathbb{S}_+^n \text{ such that } \overline{S} = \mathbf{P}_{\mathbb{T}^n}(\hat{S})$$

where $\mathbf{P}_{\mathbb{T}^n}$ represents the projection onto $\mathbb{T}^n$ which only keeps the tridiagonal part of a symmetric matrix.

Proof. Let $\overline{S} \in (\mathbb{T}^n \cap \mathbb{S}_+^n)^* = K_n$. By Theorem 7 of [14], we know $\overline{S}$ is positive semidefinite completable.

Let $\overline{S} = \mathbf{P}_{\mathbb{T}^n}(\hat{S})$, $\hat{S} \succeq 0$, then for every $X \in \mathbb{T}^n \cap \mathbb{S}_+^n$, we have $\langle \overline{S}, X \rangle = \langle \hat{S}, X \rangle \geq 0$, so $\overline{S}$ is in $(\mathbb{T}^n \cap \mathbb{S}_+^n)^*$. □

The last two results, Proposition 2.2.3 and Lemma 2.2.5 can also be generalized to the chordal sparsity setting. For example, given a simple undirected graph G, the cone of positive semidefinite matrices with sparsity pattern G is the dual cone of the cone of positive semidefinite completable matrices with sparsity pattern G. Hence, all results presented in this section are generalizable to the chordal sparsity setting.

2.3 Tridiagonal Perturbation Problem without Regularizations

We first introduce the low-rank plus tridiagonal decomposition problem, then consider the *tridiagonal perturbation problem* and one of its relaxations which only considers the diagonal entries in the objective function. Then we prove the uniqueness of optimal solutions for this semidefinite programming relaxation.

Given $A \in \mathbb{S}^n$, a low-rank plus tridiagonal decomposition problem is defined as:

$$\begin{aligned} \min\ & \operatorname{rank}(L) \\ s.t.\ & L + Y = A \\ & L \succeq 0 \\ & Y \in \mathbb{T}^n \cap \mathbb{S}^n, \end{aligned}$$

where we assume the integer $n \geq 2$. By changing the sign of Y, we have

$$\begin{aligned} \min\ & \operatorname{rank}(A + Y) \\ s.t.\ & A + Y \succeq 0 \\ & Y \in \mathbb{T}^n \cap \mathbb{S}^n. \end{aligned} \tag{2.3.1}$$

We introduced the general diagonal perturbation problem in the previous sections, now we have a similar problem (2.3.1) where the tridiagonal entries (the diagonal entries and the entries right above and below them, which we call *bidiagonal entries* for the rest of this thesis) are allowed to be perturbed, we call this the *tridiagonal perturbation problem*. In this case, we have more control over the entries and the positive semidefinite structure of a symmetric n-by-n matrix while the cost of storing the decomposition is increased by $n - 1$. Also, we can treat this problem as a low-rank, sparse decomposition problem, because for a large or even moderate n, a tridiagonal matrix is very sparse. Since the tridiagonal perturbation problem is a generalization to the diagonal perturbation problem, some techniques and results can also be generalized.

Now, for problem (2.3.1), instead of using $Y \in \mathbb{T}^n \cap \mathbb{S}^n$, we use $(u, v) \in \mathbb{R}^n \times \mathbb{R}^{n-1}$ to represent Y.

Definition 2.3.1. We define a linear map $\operatorname{BiDiag} : \mathbb{R}^{n-1} \to \mathbb{S}^n$, where

$$\operatorname{BiDiag}(v) := \begin{bmatrix} 0 & v_1 & & 0 \\ v_1 & 0 & v_2 & \\ & v_2 & 0 & \ddots \\ 0 & & \ddots & \ddots \end{bmatrix}$$

and its adjoint $\operatorname{bidiag} : \mathbb{S}^n \to \mathbb{R}^{n-1}$.

We can represent $\operatorname{bidiag}(X)$ in terms of bidiagonal entries of X.

Remark 2.3.2.

$$\mathsf{bidiag}(X) := \mathsf{BiDiag}^*(X) = 2\begin{bmatrix} X_{12} \\ X_{23} \\ \vdots \\ X_{(n-1)n} \end{bmatrix}.$$

Proof. For every $v \in \mathbb{R}^{n-1}, X \in \mathbb{S}^n$, we have

$$\begin{aligned} &\langle \mathsf{BiDiag}(v), X \rangle \\ &= \sum_{i=1}^{n-1} v_i (X_{i(i+1)} + X_{(i+1)i}) \\ &= \sum_{i=1}^{n-1} 2 v_i X_{i(i+1)} \\ &= \langle v, \left[2X_{12}, \ldots, 2X_{(n-1)n}\right]^\top \rangle, \end{aligned}$$

so we have $\mathsf{BiDiag}^*(X) = 2\left[X_{12}, \ldots, X_{(n-1)n}\right]^\top$. □

Definition 2.3.3. Given the linear maps above, we define $\mathsf{TriDiag} : \mathbb{R}^n \times \mathbb{R}^{n-1} \to \mathbb{S}^n$, where

$$\mathsf{TriDiag}(u, v) := \mathsf{Diag}(u) + \mathsf{BiDiag}(v)$$

and its adjoint $\mathsf{tridiag}(X) : \mathbb{S}^n \to \mathbb{R}^n \times \mathbb{R}^{n-1}$.

We can represent $\mathsf{tridiag}(X)$ in terms of $\mathsf{diag}(X)$ and $\mathsf{bidiag}(X)$.

Remark 2.3.4.

$$\mathsf{tridiag}(X) := \mathsf{TriDiag}^*(X) = \begin{bmatrix} \mathsf{diag}(X) \\ \mathsf{bidiag}(X) \end{bmatrix}.$$

Proof. For every $u \in \mathbb{R}^n, v \in \mathbb{R}^{n-1}, X \in \mathbb{S}^n$, we have

$$\begin{aligned} &\langle \mathsf{TriDiag}(u, v), X \rangle \\ &= \sum_{i=1}^{n} u_i X_{ii} + \sum_{i=1}^{n-1} v_i (X_{i(i+1)} + X_{(i+1)i}) \\ &= \langle u, \mathsf{diag}(X) \rangle + \sum_{i=1}^{n-1} 2 v_i X_{i(i+1)} \\ &= \langle u, \mathsf{diag}(X) \rangle + \langle v, \mathsf{bidiag}(X) \rangle \end{aligned}$$

so we have $\mathsf{TriDiag}^*(X) = \begin{bmatrix} \mathsf{diag}(X) \\ \mathsf{bidiag}(X) \end{bmatrix}$. □

Then for the problem (2.3.1), we again relax it by replacing the rank function by the nuclear norm. Since our problem is defined over positive semidefinite matrices, $\|A + Y\|_* = \mathrm{tr}(A + Y)$. Also, since A is fixed, the problem becomes

$$\begin{aligned} \min\ & \langle \mathbb{1}, u \rangle \\ s.t.\ & S = \mathrm{TriDiag}(u, v) + A \\ & S \succeq 0 \\ & u \in \mathbb{R}^n \\ & v \in \mathbb{R}^{n-1} \end{aligned} \tag{2.3.2}$$

where $\mathbb{1} \in \mathbb{R}^n$, and its dual is defined as

$$\begin{aligned} \max\ & \langle A, X \rangle \\ s.t.\ & \mathrm{tridiag}(X) = \begin{bmatrix} \mathbb{1} \\ 0 \end{bmatrix} \\ & X \succeq 0. \end{aligned} \tag{2.3.3}$$

Remark 2.3.5. For the dual problem, the constraints are on the tridiagonal entries of the matrix X while the objective value only depends on the off-diagonal entries. Hence this problem is equivalent to positive semidefinitely completing the tridiagonal matrix with ones on the diagonal and zero on the bidiagonal entries while minimizing $\langle A, X \rangle$.

Proposition 2.3.6. Both the primal (2.3.2) and the dual (2.3.3) attain their optimal values and their optimal values are equal.

Proof. The primal (2.3.2) is strictly feasible with any solution $u \in \mathbb{R}^n, v \in \mathbb{R}^{n-1}$ that makes the matrix $A + \mathrm{TriDiag}(u, v)$ strictly diagonally dominant. The dual (2.3.3) is also strictly feasible with the identity matrix $I_{n \times n}$. Thus, by Theorem 1.1.3, we know both problems attain the same optimal value, and X , $(S$, (u, v) $)$ are optimal respectively if and only if they are feasible respectively and $S \cdot X = 0$. □

In many convex optimization approaches to solve hard nonconvex problems, we solve a convex relaxation of the original nonconvex problem. In such situations, if the objective function is linear and there is no gap between the attained optimal values of the convex relaxation and the original nonconvex problem, uniqueness of the optimal solution to the convex relaxation implies it is also optimal for the original nonconvex problem. Here, our objective function is not linear (we are minimizing the rank of a matrix); however, for recovery of a feasible solution to the original nonconvex problem, we are in a similar situation. If the solution to the SDP relaxation is not unique, then in addition to the lowest rank optimal solutions to the SDP relaxation, we will have infinitely many strictly higher rank optimal solutions (by taking the convex combinations of the lowest rank optimal solutions). Then, since we hope that there exists an optimal solution of our relaxation being optimal for the original problem, we need to find the lowest rank optimal solution for the SDP relaxation. That is, we need to solve for the lowest rank solution over the set of optimal solutions of the SDP relaxation, which might not be simpler than solving the original

problem. Thus, the uniqueness of the optimal solution of the SDP relaxation is essential for the recovery of a solution for the original nonconvex problem.

We now prove that the primal SDP problem (2.3.2) has a unique optimal solution. Before that, we prove a useful lemma.

Lemma 2.3.7. For every $u \in \mathbb{R}^n, v \in \mathbb{R}^{n-1}, (u,v) \neq 0$, there exist $u^0 \in \mathbb{R}^n, v^0 \in \mathbb{R}^{n-1}$ with $\langle \mathbb{1}, u^0 \rangle \neq 0$ such that $\mathrm{Null}(\mathrm{TriDiag}(u,v)) \quad \mathrm{Null}(\mathrm{TriDiag}(u^0,v^0))$.

Proof. If $\langle \mathbb{1}, u \rangle \neq 0$, pick $(u^0,v^0) = (u,v)$ and we are done, so assume $\langle \mathbb{1}, u \rangle = 0$. We prove this by strong induction over the dimension n. Clearly, this statement holds when $n = 1,2$. Now, assume that given an integer $n > 2$, the statement holds for all $i \in [n-1]$. Write $u = [u_1\ u_2\ \dots u_n]^\top$ and $v = [v_1 \dots\ v_{n-1}]^\top$. If $v_i = 0$ for some $i \in [n-1]$, we can write $\mathrm{TriDiag}(u,v)$ as

$$\mathrm{TriDiag}(u,v) = \begin{bmatrix} \mathrm{TriDiag}(u^{(1)},v^{(1)}) & 0 \\ 0 & \mathrm{TriDiag}(u^{(2)},v^{(2)}) \end{bmatrix},$$

where $u^{(1)} = [u_1 \dots u_i]^\top$, $v^{(1)} = [v_1 \dots v_{i-1}]^\top$, $u^{(2)} = [u_{i+1} \dots u_n]^\top$, $v^{(2)} = [v_{i+1} \dots v_{n-1}]^\top$. By the induction hypothesis, there exists $(u^{0(1)}, v^{0(1)})$, $(u^{0(2)}, v^{0(2)})$ such that $\mathrm{Null}(\mathrm{TriDiag}(u^{(1)},v^{(1)}))$ $\mathrm{Null}(\mathrm{TriDiag}(u^{0(1)},v^{0(1)}))$ and $\mathrm{Null}(\mathrm{TriDiag}(u^{(2)},v^{(2)}))$ $\mathrm{Null}(\mathrm{TriDiag}(u^{0(2)},v^{0(2)}))$, $\langle \mathbb{1}, u^{0(1)} \rangle \neq 0$ and $\langle \mathbb{1}, u^{0(2)} \rangle \neq 0$. Then by considering $u^0 = \begin{bmatrix} u^{0(1)} \\ u^{0(2)} \end{bmatrix}$ and $v^0 = \begin{bmatrix} v^{0(1)} \\ 0 \\ v^{0(2)} \end{bmatrix}$ (or $u^{0(2)}$, $v^{0(2)}$ if necessary), we are done.

Now, we assume for all $i \in [n-1]$, $v_i \neq 0$. Write $\mathrm{TriDiag}(u,v)$ as Y and let Y_i represent the ith leading square submatrix of Y. Then consider $0 \neq w \in \mathrm{Null}(Y)$, and we have

$$u_1w_1 + w_2v_1 = 0 \Rightarrow w_2 = \frac{u_1w_1}{v_1} = \frac{\det(Y_1)}{v_1}w_1$$

$$v_1w_1 + u_2w_2 + v_2w_3 = 0 \Rightarrow w_3 = \frac{u_1u_2}{v_1v_2}w_1 \quad \frac{v_1}{v_2}w_1 = \frac{\det(Y_2)}{v_1v_2}w_1$$

The following is well known.

Claim 2.3.7.1.

$$w_i = (-1)^{i+1}\frac{\det(Y_{i-1})}{\prod_{j=1}^{i-1} v_j}w_1, \forall i \quad 2$$

Proof. We prove it by induction. The cases for $i = 2,3$ are shown above. Given $i > 3$, assume the

equation holds for all $k \in [i]$, then

$$
\begin{aligned}
& w_{i-1}v_{i-1} + w_i u_i + w_{i+1}v_i = 0 \\
\Longrightarrow\ & w_{i+1} = -\frac{u_i}{v_i}w_i - \frac{v_{i-1}}{v_i}w_{i-1} \\
\Longrightarrow\ & w_{i+1} = (-1)^{i+2}\frac{u_i}{v_i}\frac{\det(Y_{i-1})}{\prod_{j=1}^{i-1} v_j}w_1 + (-1)^{i+1}\frac{v_{i-1}}{v_i}\frac{\det(Y_{i-2})}{\prod_{j=1}^{i-2} v_j}w_1 \\
\Longrightarrow\ & w_{i+1} = (-1)^{i+2}\frac{u_i \det(Y_{i-1})}{\prod_{j=1}^{i} v_j}w_1 + (-1)^{i+1}\frac{v_{i-1}^2\det(Y_{i-2})}{\prod_{j=1}^{i} v_j}w_1 \\
\Longrightarrow\ & w_{i+1} = \frac{(-1)^{i+2}}{\prod_{j=1}^{i} v_j}\left(u_i\det(Y_{i-1}) - v_{i-1}^2\det(Y_{i-2})\right)w_1 \\
\Longrightarrow\ & w_{i+1} = (-1)^{i+2}\frac{\det(Y_i)}{\prod_{j=1}^{i} v_j}w_1,
\end{aligned}
$$

as required. □

That is, w can be written in the form $w_1 h$ where $h_1 = 1$ and h only depends on Y. Since w is arbitrary, we know $\mathrm{Null}(Y) = \mathrm{span}\{w\}$, so $\mathrm{rank}(Y) = n - 1$. In this case, if w has a zero entry, say $w_k = 0$, then we consider $M = \mathrm{Diag}(e_k)$ and let $(u', 2v') = \mathrm{tridiag}(Y + M)$, then $w \in \mathrm{Null}(\mathrm{TriDiag}(u', v'))$ which implies $\mathrm{Null}(Y) \neq \mathrm{Null}(\mathrm{TriDiag}(u', v'))$ and $\langle \mathbb{1}, u\rangle' = \langle \mathbb{1}, u\rangle + 1 = 1$, then we find the required (u', v'). If w has no zero entries, then we add Y_{11} by $\beta > 0$, add Y_{12}, Y_{21} by $\frac{w_1}{w_2}\beta$ and add Y_{22} by $\frac{w_1^2}{w_2^2}\beta$ and call the new matrix Y'. In this way, $w \in \mathrm{Null}(Y')$ and let $(u', 2v') = \mathrm{tridiag}(Y')$, we have $\langle \mathbb{1}, u\rangle' = \langle \mathbb{1}, u\rangle + \beta + \frac{w_1^2}{w_2^2}\beta > 0$ and $\mathrm{Null}(Y) \neq \mathrm{Null}(\mathrm{TriDiag}(u', v'))$. Hence, by induction, we finish the proof. □

Proposition 2.3.8. [6, Proposition 1] Consider the primal-dual pair of SDPs in the form

$$\max\{\langle C, X\rangle : \mathcal{A}(X) = b, X \succeq 0\} \tag{2.3.4}$$

$$\min\{\langle b, y\rangle : S = \mathcal{A}^*(y) - C, S \succeq 0\} \tag{2.3.5}$$

where $\mathcal{A} : \mathbb{S}^n \to \mathbb{R}^m$ is surjective. Assume there exists $\hat{X} \succ 0$ and $\hat{y} \in \mathbb{R}^m$ such that $\mathcal{A}(\hat{X}) = b$ and $\mathcal{A}^*(\hat{y}) \succ 0$. Suppose for every $0 \neq y \in \mathbb{R}^m$, there exists $z \in \mathbb{R}^m$ such that $b^\top z \geq 0$ and $\mathrm{Null}(\mathcal{A}^*(y)) \subseteq \mathrm{Null}(\mathcal{A}^*(z))$. Then (2.3.5) has a unique optimal solution.

Theorem 2.3.9. The problem (2.3.2) always has a unique optimal solution.

We provide two proofs for this theorem.

First Proof. By Proposition 2.3.7 and Prop. 2.3.8, the result is clear. □

Second Proof. Suppose (2.3.2) has two different optimal solutions $(S, (u, v))$ and $(S', (u', v'))$. By Theorem 1.1.3 we can consider an optimal solution X of the dual and since both S, S' are

optimal to (2.3.2), we have $X\ S\ = 0 = X\ S^0$ which implies $X\ (S\ \ S^0) = 0$. That is,

$$\begin{aligned} X\ (S\ \ S^0) &= X\ [(A + \mathrm{TriDiag}(u\ , v\ ))\ \ (A + \mathrm{TriDiag}(u^0, v^0))] \\ &= X\ (\mathrm{TriDiag}(u\ , v\ )\ \ \mathrm{TriDiag}(u^0, v^0)) \\ &= 0. \end{aligned}$$

We write $\mathrm{TriDiag}(\overline{u}, \overline{v}) = \mathrm{TriDiag}(u\ , v\ )\ \ \mathrm{TriDiag}(u^0, v^0)$. Then, let X_i denote the ith column of X, and we have $\overline{u}_1 X_1 + \overline{v}_1 X_2 = 0$ which implies $\overline{u}_1 X_{11} + \overline{v}_1 X_{12} = 0$ and $\overline{u}_1 X_{21} + \overline{v}_1 X_{22} = 0$. Since $X_{11} = X_{22} = 1$ and $X_{12} = X_{21} = 0$, we have $\overline{u}_1 = \overline{v}_1 = 0$. Repeating this process, we see that $(\overline{u}, \overline{v}) = (0, 0)$. That is, $(u\ , v\ ) = (u^0, v^0)$, which implies $S\ = S^0$. Thus, we reach a contradiction. □

We now give an example showing that the relaxation problem (2.3.2) sometimes gives an optimal solution to the original problem (2.3.1).

Example 2.3.10. Consider

$$A = \begin{bmatrix} 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \end{bmatrix}.$$

The optimal solution of (2.3.2) with respect to this A is

$$(u\ , v\ ) = (\begin{bmatrix} \sqrt{\frac{8}{5}} & \sqrt{\frac{1}{10}} + \sqrt{\frac{5}{8}} & \sqrt{\frac{5}{2}} & \sqrt{\frac{1}{10}} + \sqrt{\frac{5}{8}} & \sqrt{\frac{8}{5}} \end{bmatrix}, \begin{bmatrix} \sqrt{\frac{2}{5}} & \frac{1}{2} + \sqrt{\frac{5}{8}} & \frac{1}{2} + \sqrt{\frac{5}{8}} & \sqrt{\frac{2}{5}} \end{bmatrix})^>$$

where

$$A + \mathrm{TriDiag}(u\ , v\ ) = \begin{bmatrix} \sqrt{\frac{8}{5}} & \sqrt{\frac{2}{5}} & 1 & 1 & 0 \\ \sqrt{\frac{2}{5}} & \sqrt{\frac{1}{10}} + \sqrt{\frac{5}{8}} & \frac{1}{2} + \sqrt{\frac{5}{8}} & 1 & 1 \\ 1 & \frac{1}{2} + \sqrt{\frac{5}{8}} & \sqrt{\frac{5}{2}} & \frac{1}{2} + \sqrt{\frac{5}{8}} & 1 \\ 1 & 1 & \frac{1}{2} + \sqrt{\frac{5}{8}} & \sqrt{\frac{1}{10}} + \sqrt{\frac{5}{8}} & \sqrt{\frac{2}{5}} \\ 0 & 1 & 1 & \sqrt{\frac{2}{5}} & \sqrt{\frac{8}{5}} \end{bmatrix}$$

is positive semidefinite and has rank 2, we can see that by observing $a := [1/2,\ \ 1, 0, 0, \sqrt{5/8}]^>$, $b := [\sqrt{5/8}, 0, 0,\ \ 1, 1/2]^>$ and $c := [0, 1,\ \ (1 + \sqrt{\frac{2}{5}}), 1, 0]^>$, in $\mathrm{Null}(A + \mathrm{TriDiag}(u, v))$. Then

let $\alpha := \frac{1+\sqrt{1/10}+\sqrt{5/8}}{1/2+\sqrt{5/8}} = (1+\sqrt{\frac{2}{5}}), \beta = \frac{1}{2}+\sqrt{5/8}$, we have

$$X := \frac{1}{\alpha^2}(a+b+c)(a+b+c)^\top + \frac{1}{2 \quad 1/\beta}\left[\frac{1}{\beta}(a+b)(a+b)^\top + 2(aa^\top + bb^\top)\right]$$

$$= \begin{bmatrix} 1 & 0 & \sqrt{\frac{5}{8}} & \frac{3}{2}\sqrt{\frac{1}{10}} & \frac{\beta^2}{\alpha^2}+\frac{3}{4}\sqrt{\frac{1}{10}} \\ 0 & 1 & 0 & \sqrt{\frac{2}{5}} & \frac{3}{2}\sqrt{\frac{1}{10}} \\ \sqrt{\frac{5}{8}} & 0 & 1 & 0 & \sqrt{\frac{5}{8}} \\ \frac{3}{2}\sqrt{\frac{1}{10}} & \sqrt{\frac{2}{5}} & 0 & 1 & 0 \\ \frac{\beta^2}{\alpha^2}+\frac{3}{4}\sqrt{\frac{1}{10}} & \frac{3}{2}\sqrt{\frac{1}{10}} & \sqrt{\frac{5}{8}} & 0 & 1 \end{bmatrix} \quad 0$$

Then X is feasible for the dual problem of (2.3.2), problem (2.3.3). And by the definition of X, we know $(A + \mathrm{TriDiag}(u, v))X = 0$ and $\langle A, X\rangle = 4\sqrt{\frac{5}{8}} + 6\sqrt{\frac{1}{10}} + 2\sqrt{\frac{2}{5}} = \sqrt{\frac{5}{2}} + 2\sqrt{\frac{5}{8}} + 2\sqrt{\frac{1}{10}} + 2\sqrt{\frac{8}{5}} = \mathbb{1}^\top u$. Thus, both $(u, v), X$ are optimal for (2.3.2) and (2.3.3) respectively.

Also, for every (u, v), the matrix

$$A + \mathrm{TriDiag}(u, v) \quad 0$$

has rank at least 2 because its first and second columns are always linearly independent. Also, a nonoptimal solution of the relaxation might be optimal for the original problem. For example, for (2.3.1),

$$Y := \begin{bmatrix} 2 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 2 \end{bmatrix}$$

is an optimal solution with rank 2 because

$$A + Y = \mathbb{1}\mathbb{1}^\top + \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

but $2 + 1 + 1 + 1 + 2 = 7 > \mathbb{1}^\top u \quad 6.32$.

2.4 Tridiagonal Perturbation with Regularization

One application (factor analysis) of low-rank plus diagonal decomposition is that we want to write a given (observed or forecasted) numerical covariance matrix as a sum of a low-rank positive

semidefinite matrix (representing the main factors in the phenomenon at hand) and a diagonal matrix (representing some noises which are assumed to be independent and identically distributed, possibly with another set of independent random variables). In this context, our tridiagonal generalization allows additional random variables where the ith variable is independent of all variables with index $1, \ldots, i \quad 2, i + 2, \ldots, n$. This structure arises in many applications including time dependent models (when the index i represents the ith time period) or physical models like truss design (when the closeness of indices represents physical proximity). The latter application or its approximations may lead to banded structure in the matrix of which tridiagonal structure is the simplest (after diagonal matrices). In these types of applications, it may be useful to apply different weights or functions to the variables in the bidiagonal part of the matrix compared to those on the diagonal.

Now, we consider a tridiagonal problem whose constraints are the same as the ones of the problem in the previous section. However, we now put a regularization term on the absolute values of bidiagonal entries, that is, we want to give them different weights.

Given a $\lambda \quad 0$, consider the problem,

$$\begin{array}{ll} \min & \mathbb{1}^{>} u + \lambda \|v\|_1 \\ s.t. & \mathrm{TriDiag}(u, v) \qquad A \\ & u \; 2 \; \mathsf{R}^n \\ & v \; 2 \; \mathsf{R}^{n \; 1}, \end{array} \tag{2.4.1}$$

where $\|x\|_1 := \sum_{i=1}^{n} |x_i|$ represents the 1-norm.

Remark 2.4.1. Notice that, when $\lambda = 0$, this problem is the same as the problem (2.3.2). Hence, all the results we have in this section can be applied to the problem (2.3.2).

This problem is equivalent to

$$\begin{array}{ll} \min & \mathbb{1}^{>} u + \lambda \mathbb{1}^{>} t \\ s.t. & \mathrm{TriDiag}(u, v) \qquad A \\ & t \quad v \quad 0 \\ & t + v \quad 0 \\ & u \; 2 \; \mathsf{R}^n \\ & t, v \; 2 \; \mathsf{R}^{n \; 1}, \end{array} \tag{TriRegP}$$

and we define the linear map on the left-hand side of the constraints as $A \; (u, v, t) : \mathsf{R}^n \quad \mathsf{R}^{n \; 1}$ $\mathsf{R}^{n \; 1} \; ! \; \mathsf{S}^n \quad \mathsf{R}^{n \; 1} \quad \mathsf{R}^{n \; 1}$.

Its dual is defined as

$$\begin{aligned} \max \quad & \langle A, X \rangle \\ & \mathrm{diag}(X) = \mathbb{1} \\ & \mathrm{bidiag}(X) + w \quad \xi = 0 \\ & w + \xi = \lambda \mathbb{1} \\ & X \quad 0, w \quad 0, \xi \quad 0, \end{aligned} \tag{TriRegD}$$

where the linear map on the left-hand side of the constraints is $\mathcal{A}(X, \xi, \omega) : \mathbb{S}^n \quad \mathbb{R}^{n-1} \quad \mathbb{R}^{n-1} \to \mathbb{R}^n \quad \mathbb{R}^{n-1} \quad \mathbb{R}^{n-1}$ and let $\hat{b} = [\mathbb{1}^\top, 0, \lambda \mathbb{1}]^\top$ denote the right-hand side for the rest of this section.

Corollary 2.4.2. The left-hand side linear map $\mathcal{A}\ (u, v, t)$ of (TriRegP) is the adjoint of the left-hand side linear map $\mathcal{A}(X, \xi, \omega)$ of (TriRegD).

Proof. Consider any (u, v, t) and (X, ω, ξ), we have

$$\begin{aligned} & \langle \mathcal{A}\ (u, v, t), (X, \xi, \omega) \rangle \\ =& \langle (\mathrm{TriDiag}(u, v), t \quad v, t + v), (X, \xi, \omega) \rangle \\ =& \langle \mathrm{TriDiag}(u, v), X \rangle + \langle t \quad v, \xi \rangle + \langle t + v, \omega \rangle \\ =& \langle u, \mathrm{diag}(X) \rangle + \langle v, \mathrm{bidiag}(X) \rangle + \langle v, \omega \quad \xi \rangle + \langle t, \omega + \xi \rangle \\ =& \langle u, \mathrm{diag}(X) \rangle + \langle v, \mathrm{bidiag}(X) + \omega \quad \xi \rangle + \langle t, \omega + \xi \rangle \\ =& \langle (u, v, t), \mathcal{A}(X, \xi, \omega) \rangle \end{aligned}$$

□

We now prove that the primal problem has a unique solution when $\lambda < 2$. We start with a lemma.

Lemma 2.4.3. For every $0 \neq (u, v, t) \in \mathbb{R}^n \quad \mathbb{R}^{n-1} \quad \mathbb{R}^{n-1}$, there is $z = (u^0, v^0, t^0) \in \mathbb{R}^n \quad \mathbb{R}^{n-1} \quad \mathbb{R}^{n-1}$ such that $\mathbb{1}^\top u^0 + \lambda \mathbb{1}^\top t^0 \neq 0$ and $\mathrm{Null}(\mathcal{A}\ (u, v, t)) \quad \mathrm{Null}(\mathcal{A}\ (u^0, v^0, t^0))$.

Proof. First assume that $\mathbb{1}^\top u + \lambda \mathbb{1}^\top t = 0$, otherwise let $(u^0, v^0, t^0) = (u, v, t)$ and we are done.

We prove the claim by doing induction on n. When $n = 1$, the lemma clearly holds. When $n = 2$, we write

$$\mathcal{A}\ (u, v, t) = \begin{bmatrix} u_1 & v_1 & 0 & 0 \\ v_1 & u_2 & 0 & 0 \\ 0 & 0 & t \quad v_1 & 0 \\ 0 & 0 & 0 & t + v_1 \end{bmatrix}$$

and let $0 \neq w \in \mathrm{Null}(\mathcal{A}\ (u, v, t))$. If exactly one of w_1, w_2 is zero, without loss of generality, say $w_1 = 0, w_2 \neq 0$, then $u_2 = v_1 = 0$. And $\mathbb{1}^\top u + \lambda t = 0$ implies $t = \quad u_1/\lambda \neq 0$ because $(u, v, t) \neq 0$. Hence $\mathrm{Null}(\mathcal{A}\ (u, v, t)) = \mathit{span}\{[0\ 1\ 0\ 0]^\top\}$, we can find the required u^0, v^0, t^0 easily.

We then consider the case that both w_1, w_2 are nonzero. We have

$$\begin{aligned} w_1u_1 + w_2v_1 = 0 \implies u_1 &= \frac{w_2}{w_1}v_1 \\ w_1v_1 + w_2u_2 = 0 \implies u_2 &= \frac{w_1}{w_2}v_1 \\ w_3(t - v_1) &= 0 \\ w_4(t + v_1) &= 0. \end{aligned}$$

If $t - v_1 \neq 0$ and $t + v_1 \neq 0$, then consider $(u, v, \alpha t)$ and $\alpha \neq 1$, we find a required (u^0, v^0, t^0). Hence, assume $t \in \{v_1, -v_1\}$. Then $\mathbb{1}^T u + \lambda t = u_1 + u_2 + \lambda t = 0 = \begin{cases} (\frac{w_2}{w_1} + \frac{w_1}{w_2} - \lambda)v_1, & t = v_1; \\ (\frac{w_2}{w_1} + \frac{w_1}{w_2} + \lambda)v_1, & t = -v_1. \end{cases}$

If $v_1 = 0$, then $u_1 = u_2 = t = 0$, so $(u, v, t) = 0$, contradiction. Hence, we assume $v_1 \neq 0$, and by dividing $-v_1$ and multiply w_1w_2 on both side, we have

$$0 = \begin{cases} (\frac{w_2}{w_1} + \frac{w_1}{w_2} - \lambda), & t = v_1; \\ (\frac{w_2}{w_1} + \frac{w_1}{w_2} + \lambda), & t = -v_1 \end{cases}$$

$$(\ ) \quad 0 = \begin{cases} w_1^2 + w_2^2 - \lambda w_1w_2, & t = v_1; \\ w_1^2 + w_2^2 + \lambda w_1w_2, & t = -v_1, \end{cases}$$

then we have

$$\begin{cases} \left.\begin{array}{l} (w_1 - w_2)^2 + (-\lambda + 2)w_1w_2 = 0 \\ (w_1 + w_2)^2 + (-\lambda - 2)w_1w_2 = 0 \end{array}\right\}, & t = v_1; \\ \left.\begin{array}{l} (w_1 - w_2)^2 + (\lambda + 2)w_1w_2 = 0 \\ (w_1 + w_2)^2 + (\lambda - 2)w_1w_2 = 0 \end{array}\right\}, & t = -v_1 \end{cases}$$

which implies $w_1w_2 = 0$ when $\lambda < 2$, contradiction. That is, both of w_1, w_2 are zeros. Then pick $z = [u_1^0\ u_2^0\ v_1\ t]$ where $u_1^0 + u_2^0 + t \neq 0$, we are done. Hence, for the $n = 2$ case, we can always find a required (u^0, v^0, t).

Now, assume that given an $n > 2$, the lemma holds for all $3, \ldots, n - 1$ and we prove by strong induction. If v has a zero entry, say $v_i = 0$, then let $u^0 = [u_1, \ldots, u_i]$, $v^0 = [v_1, \ldots, v_{i-1}]$, $t^0 = [t_1, \ldots t_{i-1}]$ and let $u^{00} = [u_{i+1}, \ldots, u_n]$, $v^{00} = [v_{i+1}, \ldots, v_{n-1}]$, $t^{00} = [t_{i+1}, \ldots t_{n-1}]$. So

$$A(u, v, t) = \begin{bmatrix} \mathrm{TriDiag}(u^0,v^0) & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \mathrm{TriDiag}(u^{00},v^{00}) & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \mathrm{Diag}(t^0 - v^0) & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & t_i & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \mathrm{Diag}(t^{00} - v^{00}) & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \mathrm{Diag}(t^0 + v^0) & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & t_i & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathrm{Diag}(t^{00} + v^{00}) \end{bmatrix}.$$

Notice that for $(u^0, v^0, t^0) \in \mathbb{R}^{i-1} \times \mathbb{R}^{i-1}$ and $(u^{00}, v^{00}, t^{00}) \in \mathbb{R}^{n-i-1} \times \mathbb{R}^{n-i-1}$, there

exists z', z'' satisfying the lemma with respect to them by the induction assumption. Hence, $z = [u_{z'}\ u_{z''}\ v_{z'}\ 0\ v_{z''}\ t_{z'}\ t_i\ t_{z''}]$ is a vector satisfying the lemma with respect to (u, v, t) (scale z'' if necessary).

Now, we can assume that v has no zero entries. In this way, by the proof of Proposition 2.3.7, we know $\mathsf{Null}(\mathsf{TriDiag}(u, v)) = span\{w\}$. Then if $w_i = 0$ for some $i \in [n]$, we perturb u_i by 1 to get a required z. If $w_i \neq 0$ for $i \in [n]$, for every $\beta > 0$, add u_1 by β, v_1 by $\frac{w_1}{w_2}\beta$ and u_2 by $\frac{w_1^2}{w_2^2}\beta$ and if $t_1 \neq v_1 = 0$, then add $\frac{w_1}{w_2}\beta$ to t_1, else, add $\frac{w_1}{w_2}\beta$ to t_1. We let the resulting vector be z. For every $w \in \mathsf{Null}(A \neq (u, v, t))$, if $t_1 \neq v_1 = 0$, then assume $w_{n+1} \neq 0$, and $t_1' \neq v_1' = 0$, so $w_{n+1}(t_1' \neq v_1') = 0$. Then $A \neq (u_z, v_z, t_z)w = A \neq (u, v, t)w + [\beta w_1 \quad \frac{w_1}{w_2}\beta w_2, \quad \frac{w_1}{w_2}\beta w_1 + \frac{w_1^2}{w_2^2}\beta w_2, 0, \ldots, 0, w_{n+1}(t_1' \neq v_1'), 0, \ldots, 0]^\top = 0 + 0 = 0$. The same arguments apply if $t_1 + v_1 = 0$ or $t_1 \neq v_1 \neq 0, t_1 + v_1 \neq 0$. Hence, $w \in \mathsf{Null}(A \neq (u_z, v_z, t_z)$, and $\mathsf{Null}(A \neq (u, v, t)) \neq \mathsf{Null}(A \neq (u_z, v_z, t_z))$.

Also, $\mathbb{1}^\top u_z + \lambda \mathbb{1}^\top t_z = \mathbb{1}^\top u + \lambda \mathbb{1}^\top t + (1 + \frac{w_1^2}{w_2^2} \quad \frac{w_1}{w_2})\beta > 0$ because $(1 + \frac{w_1^2}{w_2^2} \quad \frac{w_1}{w_2}) > 0$ and $\beta > 0$. We are done.

By induction, for every n, the claim holds. □

Theorem 2.4.4. The tridiagonal perturbation problem (TriRegP) has a unique optimal solution when $\lambda < 2$.

Proof. For problem (TriRegP), we can pick $v = 0$ and let every entry of t and u goes to 1, then we have a Slater point. For the dual, pick $X = I$ with $w = \xi = \frac{\lambda}{2}\mathbb{1}$, which gives a Slater point. Also, clearly, the linear map $\mathcal{A}(X, \xi, \omega)$ in the dual problem is a surjective map. Then, by Prop. 2.3.8 and Lemma 2.4.3, we know the problem (TriRegP) has a unique solution. □

Remark 2.4.5. Notice that when $\lambda \quad 2$, the proof of Lemma 2.4.3 does not hold. Consider

$$u = \begin{bmatrix} \frac{p^2}{\lambda \quad p\sqrt{\lambda^2 \quad 4}} \\ \frac{\lambda \quad p\sqrt{\lambda^2 \quad 4}}{2} \end{bmatrix}, v = 1, t = \quad 1$$

such that

$$A \neq (u, v, t) = \begin{bmatrix} \frac{p^2}{\lambda \quad p\sqrt{\lambda^2 \quad 4}} & \frac{1}{\lambda \quad p\sqrt{\lambda^2 \quad 4}} & 0 & 0 \\ 1 & \frac{2}{0} & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

whose null space is defined as

$$span\left\{ \begin{bmatrix} \frac{p\sqrt{\lambda^2 \quad 4} \quad \lambda}{2} \\ 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \right\}.$$

Then for every z such that $\mathsf{Null}(A\ (u,v,t))\quad \mathsf{Null}(A\ (u_z,v_z,t_z))$, we have

$$A\ (z) = \begin{bmatrix} \frac{\lambda\quad p^2}{\lambda^2\ 4} & 1 & 0 & 0 \\ 1 & \frac{\lambda\quad p\overline{\lambda^2\ 4}}{2} & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} v_z$$

where $\mathbb{1}^{>} u_z + \lambda t_z = (\frac{\lambda\quad p^2}{\lambda^2\ 4} + \frac{\lambda\quad p\overline{\lambda^2\ 4}}{2}\quad \lambda) v_z = 0$.

Now we also prove that when $\lambda > 2$, the optimal solution is still unique.

Theorem 2.4.6. The problem (TriRegP) has a unique optimal solution when $\lambda > 2$.

Proof. In the proof of Theorem 2.4.4, we show both (TriRegP) and (TriRegD) have Slater points. Since they are both in Standard SDP form, we know both of them attain their optimal values and their optimal values are the same. Let $(X\ ,\xi\ ,\omega\ )$ be an optimal solution of (TriRegD), and $(u\ ,v\ ) = \mathsf{tridiag}(X^0)$. Now, consider any two optimal solutions $(u^0,v^0,t^0),(u^{00},v^{00},t^{00})$ of (TriRegP), and let $z = \begin{bmatrix} u_z \\ v_z \\ t_z \end{bmatrix} = \begin{bmatrix} u^0 \\ v^0 \\ t^0 \end{bmatrix} \quad \begin{bmatrix} u^{00} \\ v^{00} \\ t^{00} \end{bmatrix}$ be their difference. By the properties of optimal solutions of the primal and dual problems, we know $X\ (\mathsf{TriDiag}(u^0,v^0) + A\quad \mathsf{TriDiag}(u^{00},v^{00})\ A) = X\ \mathsf{TriDiag}(u_z,v_z) = 0, \mathsf{Diag}(\xi\ )\,\mathsf{Diag}(t_z\quad v_z) = 0 = \mathsf{Diag}(w\ )\,\mathsf{Diag}(t_z + v_z)$. Let $K_1 := \{ij\xi_i = 0\}$ and $K_2 := \{ijw_i = 0\}$.

Claim 2.4.6.1. $K_1 \cup K_2 = ;$.

Proof. First, since X is feasible for the dual, we have $v = \frac{1}{2}(\xi\quad w\ )$ and $w + \xi = \lambda\mathbb{1}, w\ ,\xi\ 0$. Suppose $K_1 \cup K_2 \neq ;$, without loss of generality, pick $i \in K_1$, then $w_i = \lambda$ and $v_i = \frac{\lambda}{2} < 1$. Then, $X_{ii}X_{(i+1)(i+1)}\quad X_{i(i+1)}X_{(i+1)i} = 1^2\quad \frac{\lambda^2}{4} < 0$, contradicts to the fact the $X\quad 0$. Hence, $K_1 = ;$, similarly, we can prove $K_2 = ;$, so we finish the proof of the claim.

Notice that by

$$\mathsf{Diag}(\xi\ )\,\mathsf{Diag}(t_z\quad v_z) = 0 = \mathsf{Diag}(w\ )\,\mathsf{Diag}(t_z + v_z),$$

and by the claim above, $w\ ,\xi\ > 0$, so $t_z\quad v_z = 0$ and $t_z + v_z = 0$, which implies that $t_z = 0$ and $v_z = 0$. Then $X\ \mathsf{TriDiag}(u_z,v_z) = 0$ is equivalent to $X\ \mathsf{Diag}(u_z) = 0$, but $\mathsf{diag}(X\ ) = u\ = \mathbb{1}$, so $u_z = 0$. That is, $z = 0$. So (TriRegP) has a unique optimal solution. □

However, when $\lambda = 2$, (TriRegP) might have infinitely many optimal solutions. We show this by the following example.

Example 2.4.7. Let $A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \in \mathsf{S}^2$, $\lambda = 2$. Then for every feasible solution X of (TriRegD), we have $X\quad 0$ and $\mathsf{tridiag}(X) = (u, 2v) = (\mathbb{1}, 2v), \xi + w = 2$, then $\langle A, X\rangle = \quad 2v$. Since

$X \quad 0$, we have $1 = u_1 u_2 \quad v^2$, hence $h \quad A, X i = \quad 2v \quad 2$. Then,

$$X = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, w = 2, \xi = 0$$

is an optimal solution for (TriRegD). And we consider an arbitrary optimal solution (u, v, t) of (TriRegP). By Theorem 1.1.3, we know $X(\mathsf{TriDiag}(u, v) + A) = 0, \mathsf{Diag}(t \quad v)\,\mathsf{Diag}(\xi) = 0 = \mathsf{Diag}(t + v)\,\mathsf{Diag}(w)$ which implies $t + v = 0$ and $u_1 = u_2 = 1 + v = 1 \quad t$, so $\mathbb{1}^{>} u + \lambda t = 1 \quad t + 1 \quad t + 0 + 2t = 2$. That is, given any $1 \quad t \quad 0$, $((1 \quad t, 1 \quad t), \quad t, t)$ is an optimal solution of (TriRegP).

Remark 2.4.8. When we force $\lambda \quad ! \quad 1$, for (TriRegP), clearly we force t and hence v to go to zero, so the problem becomes the MTFA problem. Besides, we can see from the proof of Theorem 2.4.6 above and complementary slackness conditions that when $\lambda > 2$, we have $w \ , \xi \ > 0$ for every optimal solution of (TriRegD) which, by Theorem 1.1.3, implies that any optimal solution of the primal satisfies $t \quad v = 0$ and $t + v = 0$, so $t = v = 0$. That is, the optimal solution of (TriRegP) only perturbs the diagonal entries, so we are back to the MTFA problem. The same argument causes an issue when $\lambda = 2$, because the optimal solution might not be strictly feasible and (TriRegP) might have optimal solutions perturbing the bidiagonal entries, for example, Ex. 2.4.7. Similar results are observed when we consider the dual. Notice that when $\lambda > 2$, the dual constraints

$$\begin{aligned} &\mathsf{diag}(X) = \mathbb{1} \\ &\mathsf{bidiag}(X) + w \quad \xi = 0 \\ &w + \xi = \lambda \mathbb{1} \\ &X \quad 0, w \quad 0, \xi \quad 0, \end{aligned}$$

are equivalent to

$$\begin{aligned} &\mathsf{diag}(X) = \mathbb{1} \\ &\quad \lambda \mathbb{1} \quad \mathsf{bidiag}(X) \quad \lambda \mathbb{1} \\ &X \quad 0, \end{aligned}$$

where the second constraint is always satisfied when $\mathsf{diag}(X) = \mathbb{1}$ and $X \quad 0$, so the problem is equivalent to the dual of MTFA.

Let us consider (MTFAD):

$$\begin{aligned} &\min \ hA, X i \\ &s.t. \ \mathsf{diag}(X) = \mathbb{1} \\ &\quad X \quad 0. \end{aligned}$$

By adding the redundant constraints, $2\mathbb{1} \quad \mathsf{bidiag}(X) \quad 2\mathbb{1}$, the primal we get is (TriRegP) with $\lambda = 2$, instead of (MTFA). However, by Theorem 1.1.3, the optimal values of (MTFA),(MTFAD) and (TriRegP) with $\lambda = 2$ are equal. For example, for Ex. 2.4.7, if we consider $A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ as an

input for (MTFA), then the optimal solution of it is $u = [1,1]^\top$, with optimal value being 2. So we can find the optimal value of (TriRegP) with $\lambda = 2$ by solving the corresponding (MTFA).

With the tridiagonal perturbation problem with a given λ, we consider three problems:

1. Suppose X can be written in the form of $X = S + \mathrm{TriDiag}(u, v)$, where S is positive semidefinite. What properties or conditions of (S, u, v) will ensure that (S, u, v) is the unique optimal solution of (TriRegP) with the input $A = X$?

2. We first consider a closed convex set

$$K := \{X \in \mathbb{S}^n : \mathrm{diag}(X) = \mathbb{1},\quad \lambda\mathbb{1} \leq \mathrm{bidiag}(X) \leq \lambda\mathbb{1}\}.$$

Claim 2.4.8.1. Every face F of K can be written as:

$$F = K \text{ or} \tag{2.4.2}$$
$$F = \{X \in \mathbb{S}^n : \mathrm{diag}(X) = \mathbb{1}, k \in [n-1] \text{ entries of } \mathrm{bidiag}(X) \text{ is fixed as } \lambda \text{ or } -\lambda\}. \tag{2.4.3}$$

Proof. Clearly, $F = K$ is a face of K. So we consider the second form above.

First, we show that F in the form (2.4.3) is a face of K. Without loss of generality, say there exists $k \in [n-1]$, such that

$$F = \{X \in \mathbb{S}^n : \mathrm{diag}(X) = \mathbb{1}, \mathrm{bidiag}(X)_i = \lambda, \forall i \in [k],\quad \lambda\mathbb{1} \leq \mathrm{bidiag}(X) \leq \lambda\mathbb{1}\}.$$

Suppose $X, Y \in K$, and $\alpha \in (0,1)$ such that $\alpha X + (1-\alpha)Y \in F$. Then since $X, Y \in K$, we know $\mathrm{diag}(X) = \mathrm{diag}(Y) = \mathbb{1}$ and $-\lambda\mathbb{1} \leq \mathrm{bidiag}(X) \leq \lambda\mathbb{1}$, $-\lambda\mathbb{1} \leq \mathrm{bidiag}(Y) \leq \lambda\mathbb{1}$. Since $\alpha(X) + (1-\alpha)Y \in F$, we know for $i \in [k]$, $\alpha\,\mathrm{bidiag}(X)_i + (1-\alpha)\,\mathrm{bidiag}(Y)_i = \lambda$. Since $\alpha \in (0,1)$ and $\mathrm{bidiag}(X)_i, \mathrm{bidiag}(Y)_i \leq \lambda$, we have $\mathrm{bidiag}(X)_i = \mathrm{bidiag}(Y)_i = \lambda$. Then $X, Y \in F$ because $i \in [k]$ is arbitrary. That is, F is a face of K.

Now, suppose there is a face F that cannot be written in the form (2.4.3). Then, there exists $X \in F$ that cannot be written in the form of elements of (2.4.3), so for every $i \in [n-1]$, we have $-\lambda < \mathrm{bidiag}(X)_i < \lambda$. We see that X is a relative interior point of K, and the only face of a convex set containing its relative interior points is the convex set itself. That is, $F = K$. □

Notice that there exists $1 + \sum_{k=1}^{n-1} \binom{n}{k} 2^k$ different faces of K.

Consider a set

$$E_n^0 := \{X \in E_n : -\lambda\mathbb{1} \leq \mathrm{bidiag}(X) \leq \lambda\mathbb{1}\} = \mathbb{S}_+^n \setminus K.$$

Since every face of the intersection of two closed convex sets C_1, C_2 is an intersection of a face F_1 of C_1 and a face F_2 of C_2, we know every face of E_n^0 can be written as $F \setminus F_U$, where F is a face of K and F_U is a face of $\mathbb{S}_+^n$ uniquely defined by a subspace U of $\mathbb{R}^n$. For example, given a subspace U of $\mathbb{R}^n$, by Claim 2.4.8.1, the set

$$E_n^0 \setminus F_U = \{X \succeq 0 : \mathrm{Null}(X) \supseteq U, \mathrm{diag}(X) = \mathbb{1},\quad -\lambda\mathbb{1} \leq \mathrm{bidiag}(X) \leq \lambda\mathbb{1}\},$$

is a face of $E_n^{\emptyset}$. However, recall that there exist some subspaces of $\mathbb{R}^n$ (e.g. $U = \text{span}\{e_1\}$) such that there is no $X \succeq K$ with $\text{Null}(X) \subseteq U$, so $F \setminus F_U$ is empty for every face F of K. Then this brings another question, which subspaces define a nonempty face of $E_n^{\emptyset}$?

3. Then we consider the problem: for which subspaces V of $\mathbb{R}^n$, do there exist a positive integer k and a $k \times n$ matrix V with row space V and a centered ellipsoid passing through all its columns such that when the points are projected onto the unit ball corresponding to the ellipsoid, the absolute value of the cosine value of the angle between the projected ith and $(i+1)$th columns is upper bounded by $\lambda/2$?

Similar to the diagonal perturbation problem, being motivated by the three problems above, we have the following definitions:

Definition 2.4.9. Consider a subspace U of $\mathbb{R}^n$ and $\lambda \in \mathbb{R}_+ \setminus \{2\}$.

- We say U is λ-*tridiagonal recoverable* if there exists $u \in \mathbb{R}^n, v \in \mathbb{R}^{n-1}, S \in \mathbb{S}_+^n$ such that $\text{col}(S) = U$, $(u, v, |v|)$ is the unique solution of (TriRegP) given $A = S + \text{TriDiag}(u, v)$.
- We say U is λ-*tridiagonal realizable* if there exists $(Q, \omega, \xi) \in \mathbb{S}_+^n \times \mathbb{R}^{n-1} \times \mathbb{R}^{n-1}$ such that $U \subseteq \text{Null}(Q)$, $\mathcal{A}(Q, \omega, \xi) = \hat{b}, \omega \in \mathbb{R}_+^{n-1}, \xi \in \mathbb{R}_+^{n-1}$.
- We say U has the λ-*tridiagonal ellipsoid fitting property* if
  1. there is a $k \times n$ matrix V with row space V and $\omega \in \mathbb{R}_+^{n-1}, \xi \in \mathbb{R}_+^{n-1}$ such that there is a centered ellipsoid in $\mathbb{R}^k$ passing through each column of V.
  2. Let $M \in \mathbb{S}_+^k$ represent the ellipsoid, and write $M = BB^\top$, and let $B := \{Bv : v^\top Mv = 1\}$, which is the projected unit ball corresponding to the ellipsoid. And the angle θ_i between projections of the ith and $(i+1)$th column of V onto the ball satisfies $\theta_i = \cos^{-1}(\frac{\xi_i - \omega_i}{2})$, while $\xi_i + \omega_i = \lambda$.

For the rest of the thesis, unless we state specifically, we assume that $\lambda < 2$ for (TriRegP).

Proposition 2.4.10. Consider subspaces U of $\mathbb{R}^n$ with $\lambda \in \mathbb{R}_+ \setminus \{2\}$, the following are equivalent:

1. U is λ-tridiagonally recoverable.
2. U is λ-tridiagonally realizable.
3. $U^\perp$ has the λ-tridiagonal ellipsoid fitting property.

Proof. Notice that both (TriRegP) and (TriRegD) are in the standard form of SDPs, and they both have Slater points by considering $[\alpha u, 0, t], u \neq 0, t > 0, \alpha \neq 1$ for (TriRegP) and considering $(I, \frac{\lambda}{2}\mathbb{1}, \frac{\lambda}{2}\mathbb{1})$ for (TriRegD).

Prove 1) $\Rightarrow$ 2). Now, by Theorem 1.1.3, we know $(u, v, |v|)$ and S is optimal if and only if there exists (Q, ξ, ω) being feasible for (TriRegD) such that $QS = 0$, $(|v|+v)^\top \omega = 0$, $(|v| - v)^\top \xi = 0$. Hence, $U = \text{col}(S) \subseteq \text{Null}(Q)$.

Conversely, if U is λ-tridiagonal realizable, then there exists (Q, ω, ξ) being feasible for (TriRegD) such that $QS = 0$ for every S with column space U. Consider any $S \succeq 0$ with $\mathsf{col}(S) = U$, and $u = \mathsf{diag}(S)$, $v = 0$. Then $QS = 0$, $(v + jvj)^\top \omega = 0$, and $(v - jvj)^\top \xi = 0$. And (u, v, jvj) is feasible for (TriRegP) with input $A = S - \mathsf{TriDiag}(u, v)$. By Theorem 1.1.3 and the uniqueness of optimal solution of (TriRegP) when $\lambda < 2$, we know U is λ-tridiagonal recoverable.

Prove 2) $\Rightarrow$ 3). For a given realizable U, consider $V \in \mathbb{R}^{k \times (3n-2)}$ such that $\mathsf{Null}(V) = U$, so $\mathsf{row}(V) = U^\perp$. Then by the realizability, there exists $(Q, \omega, \xi) \in \mathbb{S}^n_+ \times \mathbb{R}^{n-1} \times \mathbb{R}^{n-1}$ which is feasible for (TriRegD) and $\mathsf{Null}(Q) \supseteq U$. Then, there exists $M \in \mathbb{S}^k_+$ such that $Q = V^\top M V \succeq 0$, and $\mathcal{A}(Q, \omega, \xi) = [\mathbb{1}, 0, \lambda\mathbb{1}]^\top$ is equivalent to $v_i M v_i = 1$, $v_i M v_{i+1} = \frac{\xi_i - \omega_i}{2}$ and $\xi_i + \omega_i = \lambda$. The converse holds trivially by considering $(V^\top M V, \omega, \xi)$. □

Since we know that when $\lambda \neq 1$, the (TriRegP) is equivalent to (MTFA). For the definitions of realizability, recoverability and λ-tridiagonal ellipsoid fitting property, we also expect them to converge to the definitions of the diagonal case when $\lambda \neq 1$.

Remark 2.4.11.

1. When the $\lambda \neq 1$, optimal t and v are forced to be 0. In this case, we only consider the feasible solutions in the form $\mathcal{A}(u, 0, 0)$. Then the λ-tridiagonal recoverability of U is equivalent to that there exists $u \in \mathbb{R}^n$, $S \in \mathbb{S}^n_+$ such that $\mathsf{col}(S) = U$, u is the unique optimal solution of (TriRegP) given $A = S - \mathsf{Diag}(u)$, and by Prop. 2.1.5, we know it is equivalent to U being diagonally recoverable.

2. Similar arguments apply to realizability. Since the constraints

$$\begin{aligned} \mathsf{bidiag}(X) + \omega - \xi &= 0 \\ \omega + \xi &= \lambda\mathbb{1} \end{aligned}$$

   are equivalent to

$$\lambda\mathbb{1} \geq \mathsf{bidiag}(X) \geq \lambda\mathbb{1},$$

   when $\lambda \neq 1$, the above constraint is redundant. Hence, the λ-tridiagonal realizability condition becomes there exists $A \in \mathbb{S}^n$ such that $\mathsf{diag}(A) = \mathbb{1}$, $A \succeq 0$ and $\mathsf{Null}(A) \supseteq U$, which is the diagonal realizability condition by Proposition 2.1.5.

3. Consider the λ-tridiagonal ellipsoid fitting property. Notice that it requires a centred ellipsoid passing through each column of V, which is equivalent to requiring V satisfying the ellipsoid fitting property as defined in Definition 2.1.2. Also, since $v_i M v_{i+1} = v_i B B^\top v_{i+1}$, we have $v_i M v_{i+1} = \|B^\top v_i\| \|B^\top v_{i+1}\| \cos\theta_i$. And by the ellipsoid fitting property, we know $\|B^\top v_i\| = 1, \forall i \in [n]$. That is, the λ-tridiagonal ellipsoid fitting property requires $v_i M v_{i+1} = \cos\theta_i = \frac{\xi_i - \omega_i}{2}$. Given that $\xi_i + \omega_i = \lambda$, we have $|\cos\theta_i| \leq \frac{\lambda}{2}$. Hence, when $\lambda \geq 2$, $\cos\theta_i$ is free.

   We now give an example verifying the significance of the condition $|\cos\theta_i| \leq \lambda/2$.

   Example 2.4.12. For every $\lambda < 2$, we can find a triple V not satisfying the λ-tridiagonal ellipsoid fitting property. Consider $\lambda < 2$, $V = \mathsf{span}\{[1 \quad 1 \quad 0]^\top, [0 \quad 0 \quad 1]^\top\}$, then for

every $V \ 2 \ \mathsf{R}^{k \ 3}$ with row space V, we have $V_1 = \ V_2$. Hence, for every ellipsoid M satisfying $V_i^{>} M V_i = 1$, we have $V_1^{>} M V_2 = \ 1 < \ \lambda/2$, so this triple does not have the λ-tridiagonal ellipsoid fitting property.

Now we can apply similar arguments as above to the λ-tridiagonal ellipsoid fitting property. Let θ_i be the angle between the ith and $(i + 1)$th column of V. As above, as $\lambda \ ! \ 1$, $j\cos\theta_i j \ \frac{\lambda}{2}$ becomes a redundant constraint, so we are back to the conditions for ellipsoid fitting property in Definition 2.1.2.

The proposition above characterizes stricter properties than [22, Prop. 3.1], notice that when U is λ-tridiagonal recoverable for $\lambda < 2$ with (S, u, v), then $(S, u, 0)$ also verifies the property, and we have U is diagonally recoverable, similarly, the other two λ-tridiagonal properties also imply the corresponding diagonal ones.

A clear difference between the diagonal ellipsoid fitting property and λ-tridiagonal ellipsoid fitting property is that, for the diagonal ellipsoid fitting property, the order of columns of V does not affect the property. Hence if V has the ellipsoid fitting property, given any permutation matrix P, the matrix VP also satisfy the property with the same ellipsoid. However, this does not apply to λ-tridiagonal ellipsoid fitting property since it specifically constrains the angle between the ith and $(i + 1)$th columns.

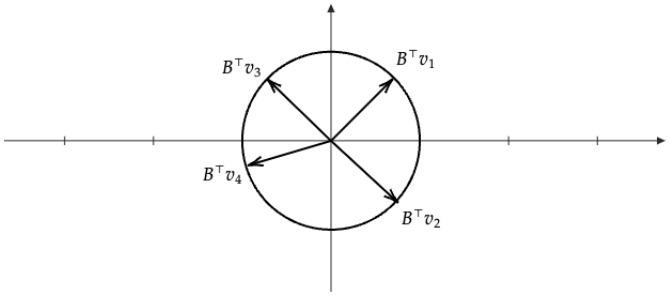


Figure 2.1: ellipsoid fitting

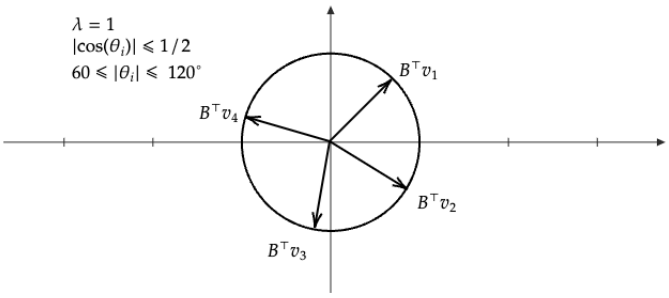


Figure 2.2: 1-tridiagonal ellipsoid fitting

Notice that from the figure (2.2) above, if we switch the column order of v_2 and v_3, then $j\cos(\theta_i)j \ 1/2$ is not satisfied, so the resulting matrix does not have the 1-tridiagonal ellipsoid fitting property.

Chapter 3

Coherence of a Subspace and Computational Examples

In this chapter, we discuss the *coherence* of a subspace. It is an indicator of how close a subspace of $\mathbb{R}^n$ is to containing any e_i, where e_i is in the standard basis in $\mathbb{R}^n$. We present some conditions on the coherence of a subspace which ensure the subspace is diagonally realizable [22]. Then, we provide conditions on the coherence of a subspace that are sufficient for the subspace to be λ-tridiagonally realizable with $\lambda \in [0, 2)$. In the end, we provide numerical examples to test the sufficiency of our conditions and to prove those conditions are not necessary.

3.1 Coherence of a Subspace

In this section, we give the definition of the *coherence* of a subspace of $\mathbb{R}^n$ and provide different ways of interpreting it. Then, we provide a result which characterizes the diagonal realizability of a subspace with its coherence, which also characterizes the recoverability and the ellipsoid fitting property by Proposition 2.1.3.

First, consider the definition:

Definition 3.1.1. [3] Let U be a subspace of $\mathbb{R}^n$ of dimension r and $\mathbf{P}_U \in \mathbb{S}_+^n$ be the orthogonal projection matrix onto U. Then the *coherence* of U (with respect to the standard basis e_i) is defined to be

$$\mu(U) := \max_{1 \le i \le n} \|\mathbf{P}_U e_i\|^2.$$

Notice that $\|\mathbf{P}_U e_i\|^2 = [\mathbf{P}_U]_{ii}$. Then let $A \in \mathbb{R}^{n \times r}$ be the matrix with rank r and columns being an orthonormal basis of U. We have

$$\mathbf{P}_U = AA^\top$$

and $\mathrm{tr}(\mathbf{P}_U) = \sum_{i=1}^n [\mathbf{P}_U]_{ii} = \mathrm{tr}(AA^\top) = \mathrm{tr}(A^\top A) = r$ by the orthonormality. That is,

$$\max_{1 \le i \le n} \|\mathbf{P}_U e_i\|^2 = \max_{1 \le i \le n} [\mathbf{P}_U]_{ii} \quad \frac{r}{n}.$$

In addition, notice that the largest eigenvalue of $\mathbf{P}_U$ is $1 = \max_{\|v\|=1} v^\top \mathbf{P}_U v \quad \|\mathbf{P}_U e_i\|^2$. Hence,

$$\frac{r}{n} \quad \max_{1 \ i \ n} \|\mathbf{P}_U e_i\|^2 \quad 1.$$

There are some other ways of interpreting the coherence of a subspace. Since for every subspace U of $\mathbb{R}^n$, we have $\mathbf{P}_U = AA^\top$ as above. Then $\mu(U) = \max_i \|A(i,:)\|^2$. With this expression, we can try to find a simple orthonormal basis of U and compute the coherence easily.

We can also view the coherence of U as an indicator of how close it is to containing any e_i. If $\mu(U) = 1$, then $\|\mathbf{P}_U e_i\|_2^2 = |\langle e_i, \mathbf{P}_U e_i\rangle| = 1$ for some i, which implies that e_i is in U and the ith column of $\mathbf{P}_U$ is e_i. Conversely, if $\mu(U) \neq 1$, then none of e_i is in U. In the view of geometry, we have $\|\mathbf{P}_U e_i\|^2 = |\langle e_i, \mathbf{P}_U e_i\rangle| = \cos\theta_i \|e_i\| \|\mathbf{P}_U e_i\| = \cos\theta_i \|\mathbf{P}_U e_i\|$, where θ_i is the angle between e_i and $\mathbf{P}_U e_i$. Then we can see that, unless $\|\mathbf{P}_U e_i\| = 0$, we have $\cos\theta_i = \|\mathbf{P}_U e_i\|$. Then $\mu(U) = \max_i \cos^2\theta_i$. Notice since $\mu(U)$ is the largest $\cos^2\theta_i$, we are actually looking for the θ_i that is the closest to 0. That is, $\mu(U)$ is actually defined by the smallest angle between the ith column of $\mathbf{P}_U$ and e_i.

Example 3.1.2. Consider two subspaces $U_1 := \text{span}\{[1/\sqrt{2}, 1/\sqrt{2}]^\top\}$, and $U_2 := \text{span}\{[\sqrt{3}/2, 1/2]^\top\}$. Then we have $\mathbf{P}_{U_1} = \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix}$ and $\mathbf{P}_{U_1} = \begin{bmatrix} 3/4 & \sqrt{3}/4 \\ \sqrt{3}/4 & 1/4 \end{bmatrix}$. Note for U_1, the smallest angle of angles between the ith column of its projection matrix and e_i is 45 , while for U_2, it is 30 . And $\mu(U_1) = 1/2 < \mu(U_2) = 3/4$. We can see this from Figure 3.1.

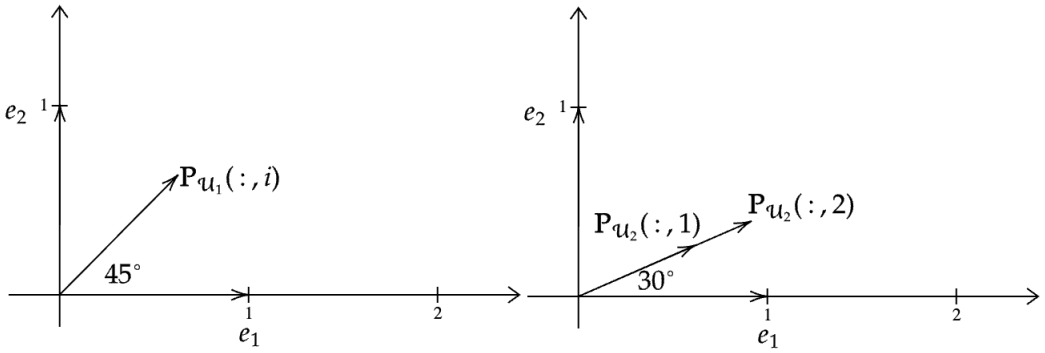


Figure 3.1: Coherence w.r.t. angles

The following theorem gives a sufficient condition with the sharpest possible constant for the diagonal realizability of a subspace.

Theorem 3.1.3. [22] If a subspace U of $\mathbb{R}^n$ has $\mu(U) < 1/2$ then U is diagonally realizable. On the other hand, for every $\alpha > 1/2$, there exists a diagonally unrealizable subspace U with $\mu(U) = \alpha$.

This condition is sufficient for a subspace U to be realizable, and then by Proposition 2.1.3, it is also sufficient for a "yes" answer to the other two problems motivating the Definition 2.1.2.

3.2 Sufficient Conditions for λ-tridiagonal Realizability

From Theorem 3.1.3, we see a condition on the coherence of subspace being sufficient to prove that the subspace is diagonally realizable. In addition to that, there exist conditions for one-dimensional subspaces which completely characterize the diagonal realizability. We will see such conditions later in Theorem 3.3.2, which shows that a one-dimensional subspace is realizable if and only if its basis vector is balanced. For $0 \quad \lambda < 2$, we provide a sufficient condition for the λ-tridiagonal realizability.

Assume $0 \quad \lambda < 2$ for the rest of this section. We start with looking for sufficient conditions implying that there exists

$$Q \; 2 \; f\mathbf{P}_{U^?} \; \mathrm{TriDiag}(u,v)\mathbf{P}_{U^?} : \mathrm{TriDiag}(u,v) \quad 0g$$

and $\omega, \xi \; 2 \; \mathsf{R}_+^{n \quad 1}$ such that (Q,ω,ξ) is feasible for (TriRegD). If we find such a (Q,ω,ξ), then this means we find a $Q \; 2 \; \mathsf{S}_+^n, \omega, \xi \; 2 \; \mathsf{R}_+^{n \quad 1}$ such that $\mathsf{Null}(Q) \quad U$ and $A(Q,\omega,\xi) = \hat{b}$, which shows U is λ-tridiagonally realizable. Note

$$\begin{aligned}
&\mathbf{P}_{U^?} \; \mathrm{TriDiag}(u,v)\mathbf{P}_{U^?} \\
=&\mathbf{P}_{U^?} \; \mathrm{Diag}(u)\mathbf{P}_{U^?} + \mathbf{P}_{U^?} \; \mathrm{BiDiag}(v)\mathbf{P}_{U^?} \\
=&\mathbf{P}_{U^?} \; \mathrm{Diag}(u)\mathbf{P}_{U^?} + \mathbf{P}_{U^?}(:,1:(n \quad 1)) \; \mathrm{Diag}(v)\mathbf{P}_{U^?}(2:n,:) \\
&+ [\mathbf{P}_{U^?}(:,1:(n \quad 1)) \; \mathrm{Diag}(v)\mathbf{P}_{U^?}(2:n,:)]^{>}.
\end{aligned}$$

Now we provide a lemma which shows that $Q = \mathbf{P}_{U^?} \; \mathrm{TriDiag}(u,v)\mathbf{P}_{U^?}, \omega, \xi$ verify the λ-tridiagonal realizability of a subspace if an SDP system has a solution. Then, we provide a condition on the subspace U and its projection matrix $\mathbf{P}_U$, which is sufficient for the SDP system to have a solution, and thus it is a sufficient condition for U to be λ-tridiagonally realizable.

Lemma 3.2.1. Let $G = (\mathbf{P}_{U^?} \quad \mathbf{P}_{U^?})$ and $H = \mathbf{P}_{U^?}(:,1:(n \quad 1)) \quad \mathbf{P}_{U^?}(:,2:n)$, then $Q = \mathbf{P}_{U^?} \; \mathrm{TriDiag}(u,v)\mathbf{P}_{U^?}, \omega, \xi$ verify the λ-tridiagonal realizability of U if

$$\begin{aligned}
&\mathrm{diag}(Q) = Gu + 2Hv = \mathbb{1} \\
&\mathrm{bidiag}(Q) = 2[H^{>}u + 2\underbrace{\mathbf{P}_{U^?}(1:(n \quad 1),1:(n \quad 1)) \quad \mathbf{P}_{U^?}(2:n,2:n)}_{:=M}v] \\
&\mathrm{bidiag}(Q) + \omega \quad \xi = 0 \qquad (3.2.1) \\
&\omega + \xi = \lambda\mathbb{1} \\
&\mathrm{TriDiag}(u,v) \quad 0 \\
&\omega \quad 0, \xi \quad 0
\end{aligned}$$

are satisfied.

Proof. Suppose there exists $Q = \mathbf{P}_{U^?} \; \mathrm{TriDiag}(u,v)\mathbf{P}_{U^?}, \omega, \xi$ satisfying system (3.2.1). First,

since $\mathbf{P}_{U^?}$ is symmetric, we have

$$\begin{aligned}\mathrm{diag}(Q) &= \mathrm{diag}(\mathbf{P}_{U^?}\,\mathrm{Diag}(u)\mathbf{P}_{U^?}) + \mathrm{diag}(2\mathbf{P}_{U^?}(:,1:(n\quad 1))\,\mathrm{Diag}(v)\mathbf{P}_{U^?}(2:n,:))\\ &= (\mathbf{P}_{U^?}\quad \mathbf{P}_{U^?}^{>})u + 2(\mathbf{P}_{U^?}(:,1:(n\quad 1))\quad \mathbf{P}_{U^?}(2:n,:)^{>})v\\ &= Gu + 2Hv\end{aligned}$$

Also, since $\mathrm{bidiag}(Q) = 2\,\mathrm{diag}\left([I_{n\ \ 1}\ 0]Q\begin{bmatrix}0\\ I_{n\ \ 1}\end{bmatrix}\right)$, we have

$$\begin{aligned}\mathrm{bidiag}(Q) =& 2\mathbf{P}_{U^?}(1:(n\quad 1),:)\,\mathrm{Diag}(u)\mathbf{P}_{U^?}(:,2:n)\\ &+ 4\mathbf{P}_{U^?}(1:(n\quad 1),1:(n\quad 1))\,\mathrm{Diag}(v)\mathbf{P}_{U^?}(2:n,2:n)\\ =& 2(\mathbf{P}_{U^?}(1:(n\quad 1),:)\quad \mathbf{P}_{U^?}(:,2:n)^{>})u\\ &+ 4(\mathbf{P}_{U^?}(1:(n\quad 1),1:(n\quad 1))\quad \mathbf{P}_{U^?}(2:n,2:n)^{>})v\\ =& 2(\mathbf{P}_{U^?}(1:(n\quad 1),:)\quad \mathbf{P}_{U^?}(2:n,:))u\\ &+ 4(\mathbf{P}_{U^?}(1:(n\quad 1),1:(n\quad 1))\quad \mathbf{P}_{U^?}(2:n,2:n))v\\ =& 2H^{>}u + 4Mv.\end{aligned}$$

Then since $Q = \mathbf{P}_{U^?}\,\mathrm{TriDiag}(u,v)\mathbf{P}_{U^?}$ and $\mathrm{TriDiag}(u,v)\quad 0$, we have $U\quad \mathrm{Null}(Q)$, $A(Q,\omega,\xi) = \hat{b}$ and $Q\quad 0$. We prove what is required. □

Notice that for such subspaces, we may use the equations $\omega + \xi = \lambda\mathbb{1}$, and upon defining new variables $h := \frac{1}{2}(\xi\quad \omega)$, we can eliminate ω and ξ from the system. Then, the system (3.2.1) is equivalent to

$$\begin{aligned}&Q = \mathbf{P}_{U^?}\,\mathrm{TriDiag}(u,v)\mathbf{P}_{U^?}\\ &\mathrm{diag}(Q) = Gu + 2Hv = \mathbb{1}\\ &\mathrm{bidiag}(Q) = 2[H^{>}u + 2Mv] = 2h\\ &\quad\frac{\lambda}{2}\quad h\quad \frac{\lambda}{2}\\ &\mathrm{TriDiag}(u,v)\quad 0.\end{aligned} \tag{3.2.2}$$

Now we want a condition which ensures the system (3.2.2) has a solution (u,v), then we have U is λ-tridiagonally realizable by Lemma 3.2.1. First, we provide a theorem which shows the positive semidefiniteness of $\mathbf{P}_{U^?}\quad \mathbf{P}_{U^?}$.

Theorem 3.2.2. (Schur's Product Theorem [23]) The Hadamard product of two positive semidefinite matrices is positive semidefinite. The Hadamard product of two positive definite matrices is positive definite.

We now provide a sufficient condition for (3.2.2) having a solution, which involves the largest column norm of $\mathbf{P}_{U^?}(:,1:(n\quad 1))\quad \mathbf{P}_{U^?}(:,2:n)$ and the coherence of U.

Proposition 3.2.3. Consider subspace U of $\mathbb{R}^n$ with coherence $\mu(U) < 1/2$ and let $p > 0$ be a constant such that $\|\mathbf{P}_{U^?}(:,i)\quad \mathbf{P}_{U^?}(:,i+1)\| \quad p$ for every $i \in [n\quad 1]$. If $p\frac{\overline{n}}{(1\quad \mu)^2} < \frac{\lambda}{2}$, there exists

infinitely many Q satisfying Q 0, U Null(Q) and

$$\lambda\mathbb{1} \quad \text{bidiag}(Q) \quad \lambda\mathbb{1}.$$

Proof. Assume U is a subspace of $\mathbb{R}^n$ with coherence $\mu(U) < 1/2$ and let $p > 0$ be a constant such that $k\mathbf{P}_{U^?}(:, i) \quad \mathbf{P}_{U^?}(:, i+1)k \quad p$ for every i 2 $[n \quad 1]$. Since $\mu < 1/2$, by [22, proof of Lemma A.1], we know $G := (\mathbf{P}_{U^?} \quad \mathbf{P}_{U^?})$ is invertible. Let D be a diagonal matrix with $D_{ii} = G_{ii}$. By the definition of G and Schur's Product Theorem, we know G is positive semidefinite, and since G is invertible, G is positive definite, so $G_{ii} > 0$ for every i 2 $[n]$. That is, for every i 2 $[n]$, $D_{ii} = G_{ii} > 0$, so D is positive definite and invertible. By the definition of coherence, we have $[\mathbf{P}_U]_{ii} \quad \mu$, then $\mathbf{P}_{U^?} = I \quad \mathbf{P}_U$ implies that $[\mathbf{P}_{U^?}]_{ii} \quad (1 \quad \mu)$, so $D_{ii} = G_{ii} \quad (1 \quad \mu)^2$ and $D_{ii}^{1} \quad 1/(1 \quad \mu)^2$. By the system (3.2.2) and G being invertible, we want to solve

$$\begin{aligned} &u = G^{1}\mathbb{1} \quad 2G^{1}Hv \\ &jH^{>}G^{1}\mathbb{1} + 2(M \quad H^TG^{1}H)vj \quad \lambda\mathbb{1} \\ &\text{TriDiag}(u, v) \quad 0 \end{aligned}$$

for u, v.

Then, by [32], we have $0 < G^{1}\mathbb{1} \quad D^{1}\mathbb{1}$. With the definition of H and $k\mathbf{P}_{U^?}(:, i) \quad \mathbf{P}_{U^?}(:, i+1)k \quad p$, we have

$$[H^{>}G^{1}\mathbb{1}]_i \quad k\mathbf{P}_{U^?}(:, i) \quad \mathbf{P}_{U^?}(:, i+1)kkG^{1}\mathbb{1}k \quad pkG^{1}\mathbb{1}k \quad pkD^{1}\mathbb{1}k \quad p\frac{p\frac{}{n}}{(1 \quad \mu)^2}.$$

That is, when $p\frac{p\frac{}{n}}{(1 \quad \mu)^2} < \lambda/2$, we have

$$\frac{\lambda}{2}\mathbb{1} < H^TG^{1}\mathbb{1} < \frac{\lambda}{2}\mathbb{1}.$$

By considering v close enough to 0, we have $jH^{>}G^{1}\mathbb{1} + 2(M \quad H^TG^{1}H)vj \quad \lambda\mathbb{1}$, and $u = G^{1}\mathbb{1} \quad 2G^{1}Hv > 0$ by $G^{1}\mathbb{1} > 0$. Also, when v is close enough to zero, we know the matrix TriDiag(u, v) is diagonally dominant, hence it is positive semidefinite. Then, with any v (there are infinitely many) satisfying the conditions above,

$$Q := \mathbf{P}_{U^?}\,\text{TriDiag}(u, v)\mathbf{P}_{U^?}$$

is a matrix we are looking for. □

Example 3.2.4. The condition $p\frac{p\frac{}{n}}{(1 \quad \mu)^2} < \frac{\lambda}{2}\mathbb{1}$ is satisfied by some subspaces. With $\lambda = 1.5$,

consider the subspace

$$U := \text{span}\left\{\begin{bmatrix} 0.3062 \\ 0.1621 \\ 0.2206 \\ 0.3856 \\ 0.1468 \\ 0.2595 \\ 0.3378 \\ 0.1198 \\ 0.0382 \\ 0.4121 \\ 0.0286 \\ 0.2369 \\ 0.1618 \\ 0.1586 \\ 0.4309 \end{bmatrix}\right\}, \mu = 0.1857, p = 0.1162, p\frac{p}{(1-\mu)^2} = 0.6787 < 0.75 = \lambda/2.$$

With

$$u = G^{-1}\mathbb{1} = \begin{bmatrix} 1.0985 \\ 1.0237 \\ 1.0460 \\ 1.1806 \\ 1.0192 \\ 1.0664 \\ 1.1262 \\ 1.0126 \\ 1.0012 \\ 1.2194 \\ 1.0007 \\ 1.0539 \\ 1.0236 \\ 1.0226 \\ 1.2521 \end{bmatrix}, v = 0, Q = \mathbf{P}_{U^\perp}\,\text{TriDiag}(u, v)\mathbf{P}_{U^\perp},$$

one can easily verify that $\max_{i2[n-1]}|Q_{i(i+1)}| = 0.0918$, and for every $i \in [n]$, $Q_{ii} = 1$, and clearly, $U \subseteq \text{Null}(Q)$.

Now, we provide an improved condition which requires knowing the dimension of U. Before that, we recall a theorem of convex analysis.

Theorem 3.2.5. Given a closed convex function f, when its domain contains a nonempty compact convex set C, such a function f attains its maximum over C and there exists an extreme point of C being a maximizer of f over C.

Proposition 3.2.6. Consider a subspace U with dimension r, coherence $\mu < 1/2$ and let $p > 0$ be a constant such that $\|\mathbf{P}_{U^\perp}(:,i) \quad \mathbf{P}_{U^\perp}(:,i+1)\| \quad p$ for every $i \in [n \quad 1]$. If $\kappa(p,r,\mu) := p\sqrt{n + \left(\frac{1}{(1 \quad \mu)^4} \quad 1\right)\frac{r}{\mu}} < \frac{\lambda}{2}$, then there exists infinitely many Q satisfying $Q \quad 0, U \quad \mathsf{Null}(Q)$ and

$$\lambda\mathbb{1} \quad \mathsf{bidiag}(Q) \quad \lambda\mathbb{1}.$$

Proof. Assume U is a subspace of $\mathbb{R}^n$ with dimension r, coherence $\mu(U) < 1/2$ and let $p > 0$ be a constant such that $\|\mathbf{P}_{U^\perp}(:,i) \quad \mathbf{P}_{U^\perp}(:,i+1)\| \quad p$ for every $i \in [n \quad 1]$. All the other steps are the same as the proof of Proposition 3.2.3, but the upper bound on the norm of $\|D^{-1}\mathbb{1}\|$ is improved. Let $g = \mathsf{diag}(\mathbf{P}_U)$, then $D_{ii}^{-1} = \frac{1}{(1 \quad g_i)^2}$. Hence $\|D^{-1}\mathbb{1}\|$ is always bounded from above by the optimal value of the following problem:

$$\begin{aligned} \max \quad & \sqrt{\sum_{i=1}^{n} \frac{1}{(1 \quad g_i)^4}} \\ s.t \quad & \sum_{i=1}^{n} g_i = r \\ & 0 \quad g_i \quad \mu, \forall i \in [n]. \end{aligned}$$

Since the square root function is strictly monotone on $\mathbb{R}_+$, g maximizes $\sqrt{\sum_{i=1}^n \frac{1}{(1 \quad g_i)^4}}$ if and only if it maximizes $\sum_{i=1}^n \frac{1}{(1 \quad g_i)^4}$ over the same set. Hence, we consider the problem with the same constraints but with the objective function $f(g) = \sum_{i=1}^n \frac{1}{(1 \quad g_i)^4}$, and the set of optimal solutions does not change. Since the feasible set is a polytope, it is a nonempty compact convex set. Then by Theorem 3.2.5, there exists an extreme point of the feasible set being optimal.

Claim 3.2.6.1. Any extreme point $g(k)$ of the feasible set can be written as:

$$\begin{aligned} & \text{for some } k \in [n \quad 1] \setminus \{0\} \\ & k \text{ of } g_i \text{ is } 0 \\ & (n \quad k \quad 1) \text{ of } g_i \text{ is } \mu \\ & 1 \text{ of } g_i \text{ is } (r \quad (n \quad k \quad 1)\mu). \end{aligned}$$

Proof. Suppose there exists an extreme point g of the feasible set with at least two entries strictly less than μ and strictly positive. Without loss of generality, assume $0 < g_1 < \mu$ and $0 < g_2 < \mu$. Consider an $\epsilon > 0$ such that $0 + \epsilon < g_1 < \mu \quad \epsilon$ and $0 + \epsilon < g_2 < \mu \quad \epsilon$. Then $g' = g + \epsilon e_1 \quad \epsilon e_2$ and $g'' = g \quad \epsilon e_1 + \epsilon e_2$ are still in the feasible set, and $g' \neq g''$, $\frac{1}{2}g' + \frac{1}{2}g'' = g$, so g is not an extreme point, we reach a contradiction. Then consider a feasible g with at most one entry, say g_1, strictly less than μ and strictly positive, we show it is an extreme point. Suppose not, then there exists $g' \neq g''$ that are both in the feasible set and $\frac{1}{2}g' + \frac{1}{2}g'' = g$, then for all $i \neq 1$, $g'_i = g''_i = g_i \in \{0, \mu\}$, otherwise, we have either one of g'_i, g''_i larger than μ or less than 0, then g', g'' are not feasible. Then since $g' \neq g''$, we know $g'_1 \neq g''_1$, and $\frac{1}{2}g'_1 + \frac{1}{2}g''_1 = g_1$. Without loss of generality, say $g'_1 < g_1 < g''_1$, then $\sum_{i=1}^n g'_i < \sum_{i=1}^n g_i = r$, so g' is not feasible, contradiction. □

Notice the last condition above requires 0 $(r \quad (n \quad k \quad 1)\mu) \quad \mu < \frac{1}{2}$, which is equivalent to $n \quad \frac{r}{\mu} \quad k \quad n \quad \frac{r}{\mu} \quad 1$.

Consider writing extreme points with k entries being 0 as $g(k)$, instead of using k 2 $[n \quad 1]$ [$f0g$, we extend it to k 2 $[0, n \quad 1]$, and consider $f(g(\ )) : \mathsf{R} \ ! \ \mathsf{R}^n$ which is defined as

$$f(g(k)) := \frac{n \quad 1}{(1 \quad \mu)^4} + \left(1 \quad \frac{1}{(1 \quad \mu)^4}\right) k + \frac{1}{(1 \quad r + (n \quad 1)\mu \quad k\mu)^4}.$$

Notice that when k is a nonnegative integer, $f(g(k)) = \sum_{i=1}^{k} 1 + \sum_{k+1}^{n} \frac{1}{(1 \quad \mu)^4} + \frac{1}{(1 \quad r+(n \quad 1)\mu \quad k\mu)^4}$, and $g(k)$ is a valid extreme point of the feasible set.

It is easy to see that $\frac{\partial^2 f(g(k))}{\partial k^2} = \frac{20\mu^2}{(1 \quad r+(n \quad 1)\mu \quad k\mu)^6} > 0$. Thus, $f(g(k))$ is strictly convex over $[n \quad \frac{r}{\mu} \quad 1, n \quad \frac{r}{\mu}]$. By Theorem 3.2.5, it is maximized at one of the extreme points of $[n \quad \frac{r}{\mu} \quad 1, n \quad \frac{r}{\mu}]$. So there exists a maximizer k of $f(g(k))$ where k 2 $fn \quad \frac{r}{\mu}, n \quad \frac{r}{\mu} \quad 1g$. Since $f(g(n \quad \frac{r}{\mu})) = f(g(n \quad \frac{r}{\mu} \quad 1)) = n + \left(\frac{1}{(1 \quad \mu)^4} \quad 1\right)\frac{r}{\mu}$, we have $f(g) \quad n + \left(\frac{1}{(1 \quad \mu)^4} \quad 1\right)\frac{r}{\mu}$, which implies $kD \ ^1\mathbb{1}k \quad \max_{g \text{ feasible}} \sqrt{f(g)} \quad \sqrt{n + \left(\frac{1}{(1 \quad \mu)^4} \quad 1\right)\frac{r}{\mu}}$. □

The bound above might not be tight, because when $n \quad \frac{r}{\mu}$ is not an integer, then the maximum of $f(g(k))$ can only be attained at $bn \quad \frac{r}{\mu}c = dn \quad \frac{r}{\mu} \quad 1e$, otherwise, k is not an integer. Notice the above constant is smaller than $p\frac{}{n}\frac{}{(1 \quad \mu)^2}$ for every subspaces. By the definition of coherence, we have $\mu \quad \frac{r}{n}$, then

$$n\left(\frac{1}{(1 \quad \mu)^4} \quad 1\right) \quad \frac{r}{\mu}\left(\frac{1}{(1 \quad \mu)^4} \quad 1\right).$$

After rearranging the terms and take the square root of both sides, we have

$$\sqrt{\frac{n}{(1 \quad \mu)^4}} \quad \sqrt{\frac{r}{\mu}\left(\frac{1}{(1 \quad \mu)^4} \quad 1\right) + n}.$$

3.3 Computational Examples

In this section, we show some computational examples of verifying the tridiagonal realizability of subspaces. We first introduce how we can verify if a subspace is tridiagonally realizable by solving an SDP, then we introduce the tool we use for solving it. After that, we provide some examples verifying our theoretical results and some computational experiments about one-dimensional subspaces.

The realizability of a given subspace (e.g. via a basis) can be tested by solving an SDP. Given 0 $\lambda < 2$ and a subspace U with an orthonormal basis $u_1, \ldots, u_r$, consider the matrix $U =$

$[u_1 \ \dots \ u_r]$ and $P := UU^T$. Then U is λ- tridiagonally realizable if and only if the SDP:

$$\begin{aligned}
&QP = 0\\
&\mathrm{diag}(Q) = \mathbb{1}\\
&\mathrm{diag}(Q[2:n, 1:(n-1)]) \quad \lambda/2 \quad \mathbb{1}_{n-1}\\
&\mathrm{diag}(Q[2:n, 1:(n-1)]) \quad \lambda/2 \quad \mathbb{1}_{n-1}\\
&Q \quad 0
\end{aligned}$$

has a feasible solution, which is an SDP that can be solved quickly by regular convex solvers. We use Domain Driven Solver (DDS) to solve the corresponding SDP for verifying tridiagonal-realizability of a subspace [17], which gives robust and accurate results. We consider a subspace realizable if DDS solves the SDP, and unrealizable if not.

For each U, consider $\beta := \min_{i2[n]}[\mathbf{P}_U]_{ii}$, and $p = \max k\mathbf{P}_{U^?}(:, i) \quad \mathbf{P}_{U^?}(:, i+1)k$.

Here are some examples of tridiagonally realizable and tridiagonally unrealizbale subspaces. Consider a subspace that is not 1-tridiagonally realizable:

$$U_1 = \mathrm{span}\left\{\begin{bmatrix} 0.1084 & 0.4777\\ 0.1493 & 0.6351\\ 0.0409 & 0.5983\\ 0.2702 & 0.0428\\ 0.6878 & 0.0690\\ 0.6467 & 0.0625 \end{bmatrix}\right\} \qquad \begin{aligned} \mu &= 0.4778\\ \beta &= 0.0748\\ p &= 0.3447. \end{aligned}$$

We set the solver tolerance as the default 10^{-8}. For $\lambda = 1$ and U_1, we run DDS to solve for Q, the software outputs $jA^{>}yj = 3.48 \quad 10^{-9}, \max fhy, Axi : x \text{ feasible } g \qquad 90.7 < 0$. For DDS, it means that for the implicit dual problem it creates for the given optimization problem, there exists a line in the feasible region, such that by following the line, the dual problem is unbounded, which is an indicator saying the optimization problem we input is infeasible.

Here is a 1-tridiagonally realizable subspace:

$$U_2 = \left\{\begin{bmatrix} 0.3881 & 0.2960\\ 0.0091 & 0.6386\\ 0.5600 & 0.3899\\ 0.3621 & 0.2987\\ 0.6357 & 0.3092\\ 0.0224 & 0.4096 \end{bmatrix}\right\}$$

$$\begin{aligned} \mu &= 0.4997\\ \beta &= 0.1682\\ p &= 0.3074. \end{aligned}$$

For this subspace, after solving the problem with DDS, we have primal feasibility being $5.60 \quad 10^{-15}$ dual feasibility being $2.95 \quad 10^{-16}$ and relative duality gap being $3.79 \quad 10^{-14}$. That is, with the default tolerance 10^{-8}, we consider our input problem (the primal) to be feasible and the implicit

dual created by DDS is also feasible. Also, we consider the duality gap to be zero, so the input problem has an optimal solution (with the objective value always being zero). That is, we consider the subspace U_2 to be tridiagonally realizable.

Now we consider a rational example,

$$U_3 = \text{span}\left\{ \frac{1}{6} \begin{bmatrix} 3 \\ 4 \\ 3 \\ 1 \\ 1 \end{bmatrix} \right\}, Q = \frac{1}{50} \begin{bmatrix} 50 & 23 & 27 & 14 & 9 \\ 23 & 50 & 23 & 28 & 34 \\ 27 & 23 & 50 & 5 & 18 \\ 14 & 28 & 5 & 50 & 5 \\ 9 & 34 & 18 & 5 & 50 \end{bmatrix}.$$

Notice $\text{diag}(Q) = \mathbb{1}$, $Q \quad 0$ and $U_3 \quad \text{Null}(Q)$, and $\mu = 4/9, \beta = 1/36, p = \sqrt{174528}/648$ 0.728. Then for this subspace, we have

$$p\frac{\sqrt{n}}{(1 \quad \mu)^2} \quad 5.274$$

$$\kappa(p, r, \mu) = p\sqrt{n + \left(\frac{1}{(1 \quad \mu)^4} \quad 1\right)\frac{r}{\mu}} \quad 1.9122.$$

Since $|Q_{i(i+1)}| < 0.5$ for every $i \; 2$ [4], we know U_3 is 1-tridiagonally realizable. However,

$$\min\left\{ p\frac{\sqrt{n}}{(1 \quad \mu)^2}, \kappa(p, r, \mu) \right\} > 0.5,$$

which shows that the conditions in Proposition 3.2.3 and Proposition 3.2.6 are not necessary for a subspace to be 1-tridiagonally realizable. Hence, in general, they are not necessary for a subspace to be λ-tridiagonally realizable, where $0 \quad \lambda < 2$.

We consider 50 of 1.5-tridiagonally realizable subspaces of $\mathbb{R}^{15}$ with dimension 1, and their corresponding $p\frac{\sqrt{n}}{(1 \quad \mu)^2}, \kappa(p, r, \mu)$:

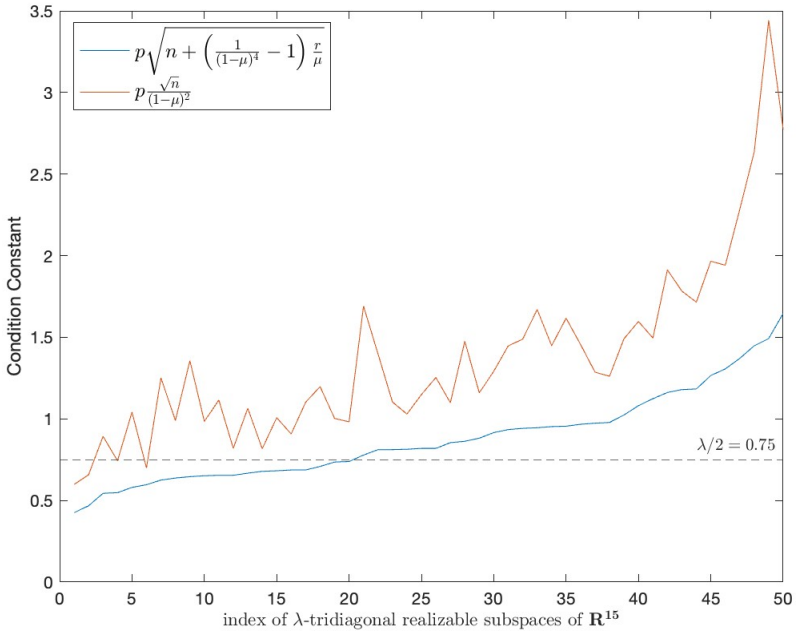


Figure 3.2: Sufficient Conditions of 1.5-tridiagonally realizable subspaces of R^{15}

We have $p\frac{\sqrt{n}}{(1-\mu)^2} \geq \kappa(p,r,\mu)$ for every subspace, and the condition $\kappa(p,r,\mu) < \lambda/2$ works much better than $p\frac{\sqrt{n}}{(1-\mu)^2} < \lambda/2$ in practice. The figure 3.2 shows that there are only a few subspaces satisfying the condition $p\frac{\sqrt{n}}{(1-\mu)^2} < \lambda/2$ but 20 of them satisfy $\kappa(p,r,\mu) < \lambda/2$.

We also does the same test for 50 subspaces of R^5. From figure 3.3, we see every 1.8-tridiagonally unrealizable subspace fails to satisfy $\kappa(p,r,\mu) \geq p\frac{\sqrt{n}}{(1-\mu)^2} < \lambda/2$. While every 1.8-tridiagonally realizable subspace fails to satisfy $p\frac{\sqrt{n}}{(1-\mu)^2} < \lambda/2$, there are still few of them satisfying $\kappa(p,r,\mu) < \lambda/2$.

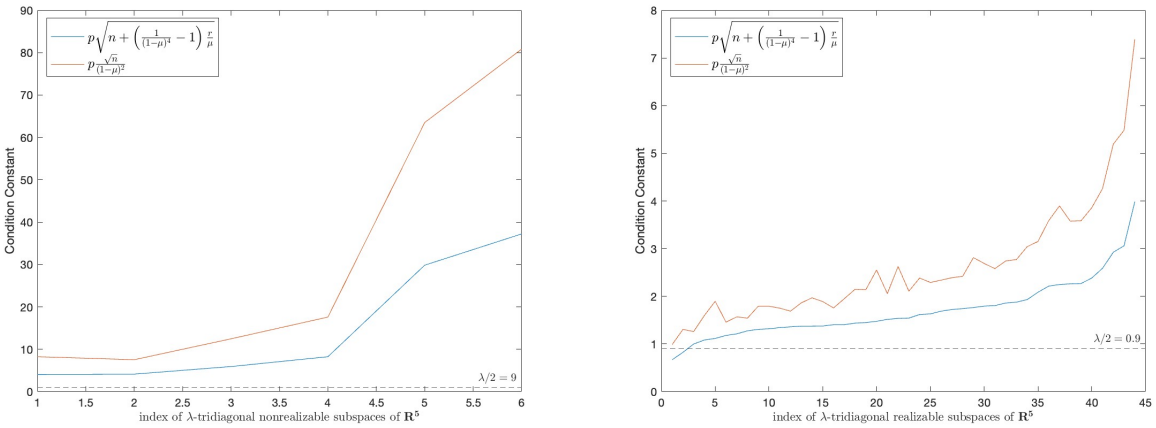


Figure 3.3: Sufficient Conditions of 1.8-tridiagonally realizable and nonrealizable subspaces of R^5

Perturb One-Dimensional Subspaces

For the diagonal case, the one-dimensional subspaces are fully characterized by a property of their basis.

Definition 3.3.1. [8] A vector $u \in \mathbb{R}^n$ is *balanced* if for all $i \in [n]$,

$$|u_i| \leq \sum_{j \neq i} |u_j|.$$

We call the vector is *strictly balanced* if the inequalities are strict for all $i \in [n]$.

The following theorem characterizes the diagonal realizability of one-dimensional subspaces by their balancedness.

Theorem 3.3.2. [8, 22, 15, 18] If a subspace U of $\mathbb{R}^n$ is realizable then every $u \in U$ is balanced. If $U = \text{span}\{u\}$ is one-dimensional then U is realizable if and only if u is balanced.

We give a sufficient condition for the basis vector of a one-dimensional subspace to be balanced.

Corollary 3.3.3. When $U = \text{span}\{u\}$, $\mu < 0.5$, we have u being balanced.

Proof. Without loss of generality, assume $\|u\|_2 = 1$. When $\mu < 0.5$, for every $i \in [n]$, we have

$$u_i^2 < 0.5 < \sum_{j \in [n] \setminus \{i\}} u_j^2 \implies |u_i| < \sqrt{\sum_{j \in [n] \setminus \{i\}} u_j^2} \leq \sum_{j \in [n] \setminus \{i\}} |u_j|,$$

hence u is balanced. □

Now we immediately extend the theorem above to the λ-tridiagonal realizability.

Corollary 3.3.4. If a subspace $U = \text{span}\{u\}$ is not balanced, then it is not λ-tridiagonal realizable for every $\lambda \neq 0$.

We also give a necessary condition on μ for a one-dimensional subspace to be tridiagonally realizable.

Corollary 3.3.5. Given $0 \leq \lambda < 2$, every one-dimensional λ-tridiagonally realizable subspace has $\mu \geq (n - 1)/n$.

Proof. Consider a realizable subspace $U = \text{span}\{u\}$. Since it is tridiagonally-realizable, it is diagonally-realizable and then the vector u is balanced. That is, for every $i \in [n]$, we have

$$|u_i| \leq \sum_{j \in [n] \setminus \{i\}} |u_j|,$$

then we have

$$\begin{aligned} |u_i|^2 &\geq 1 - u_i^2 + \sum_{j,k \in [n] \setminus \{i\}, j \neq k} |u_j||u_k| \\ &\geq 1 - u_i^2 + 2\binom{n-1}{2} \frac{1-u_i^2}{n-1} \\ &= 1 - u_i^2 + (n-2)(1-u_i^2) \end{aligned}$$

which implies $u_i^2 \leq \frac{n-1}{n}$, that is, $\mu \leq \frac{n-1}{n}$. □

Given a one dimensional subspace $U = \text{span}\{u\}$ of $\mathbb{R}^n$, we use

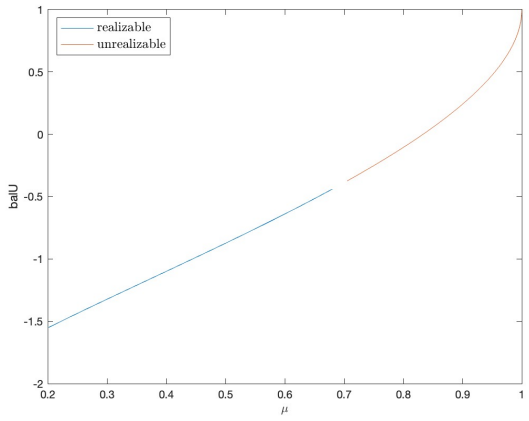
$$\text{bal}_U := \max_{j \in [n]} |u_i| \sum_{j \in [n], j \neq i} |u_j|$$

to measure how balanced U is. The smaller bal_U is, the more balanced U is. When $\text{bal}_U \leq 0$, U is balanced, and it is strictly balanced if $\text{bal}_U < 0$. Also, suppose for some i, $|u_i| = \max_{j \in [n]} |u_j|$, then $\text{bal}_U = |u_i| - \sum_{j \in [n], j \neq i} |u_j|$, because

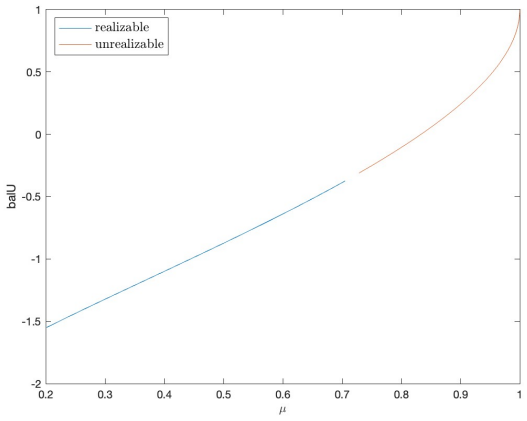
$$|u_i| - \sum_{j \in [n], j \neq i} |u_j| = |u_i| - |u_k| - \sum_{j \in [n], j \neq i,k} |u_j| \leq |u_i| - |u_k| - |u_i| + \sum_{j \in [n], j \neq i,k} |u_j|, \forall k \in [n].$$

Being motivated by Theorem 3.3.2, we want to know how the λ-tridiagonal realizability of one-dimensional subspaces is related to how balanced its basis vector is.

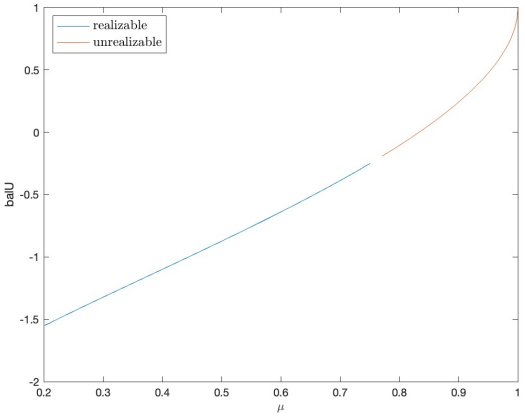
We consider the one-dimensional subspaces and test if they are realizable while perturbing one entry of its basis. We tried different λ with values $0.25, 0.5, 1, 1.5$, and for each λ, we compute the measurement of being balanced and test the realizability of subspaces of $\mathbb{R}^6$ after adding the first entry of the basis by 0.05 and normalizing it for 150 times, then the following plots show the relations between the measurement of being balanced and μ for the λ-tridiagnally realizable and unrealizable subspaces.



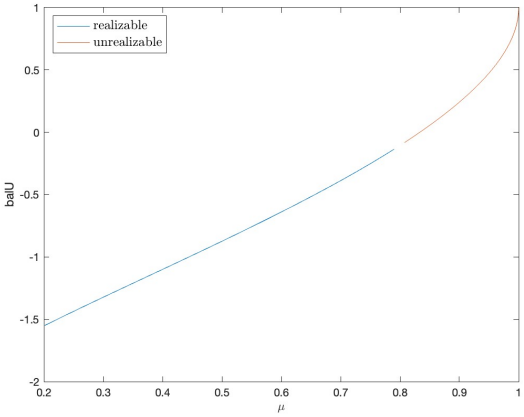
(a) $\lambda = 0.25$



(b) $\lambda = 0.5$



(c) $\lambda = 1$



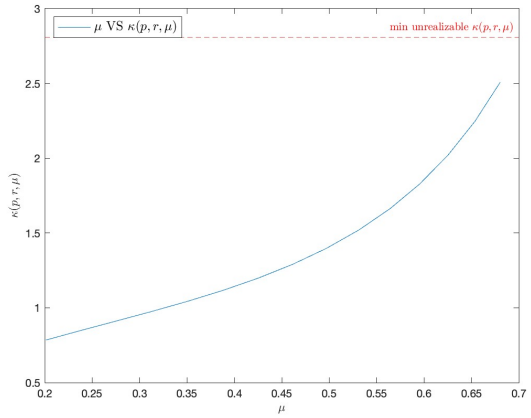
(d) $\lambda = 1.5$

Figure 3.4: bal_U VS μ for different λ

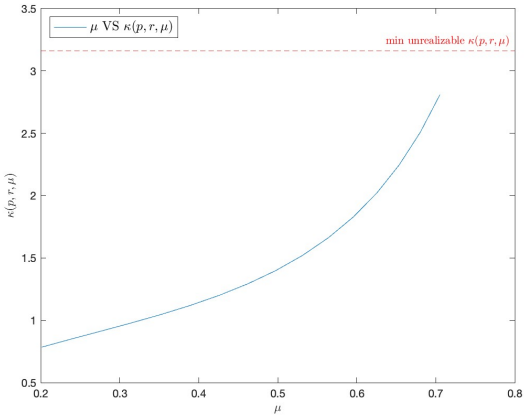
From the plots, we observe that there are thresholds determining if a one-dimensional subspace is λ-tridiagonally realizable or not. Also, as λ increases, the threshold increases as well. Recall Theorem 3.3.2 shows that when $\lambda > 2$, a one-dimensional subspace is realizable if and only if $\mathsf{bal}_U \quad 0$. That is, we can expect that when λ increases, the threshold converges to 0 from below. We can see that all the subspaces with $\mu < 0.5$ are balanced which agrees with Corollary 3.3.3. Also, the plots show that, for $0 < \lambda < 2$, the basis vector being balanced is not sufficient for the one-dimensional subspace to be λ-tridiagonally realizable.

Similar results can be obtained by observing the $\kappa(p, r, \mu)$ of the λ-tridiagonally realizable one-dimensional subspaces of R^6 and the minimum $\kappa(p, r, \mu)$ of the λ-tridiagonally unrealizable ones. We can see from the plots below that the minimum $\kappa(p, r, \mu)$ of the λ-tridiagonally unrealizable subspaces is greater than all $\kappa(p, r, \mu)$ of the λ-tridiagonally realizable subspaces for each λ. The minimum $\kappa(p, r, \mu)$ increases as λ increases. We can expect that a subspace is more likely to be λ-tridiagonally realizable when λ is larger. Also, by $\kappa(p, r, \mu)$, we expect that the subspaces with larger μ are less likely to be λ-tridiagonally realizable. That is, when λ increase, the subspaces with large μ which were unrealizable might become λ-tridiagonally realizable for the increased

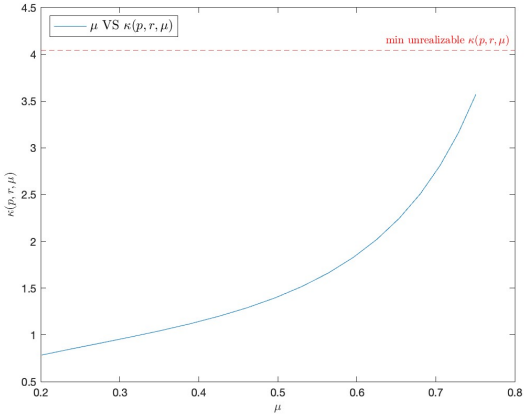
λ. Also we see that none of the μ of the λ-tridiagnally realizable subspaces exceed $\frac{n-1}{n} = \frac{5}{6}$ as Corollary 3.3.5 claims.



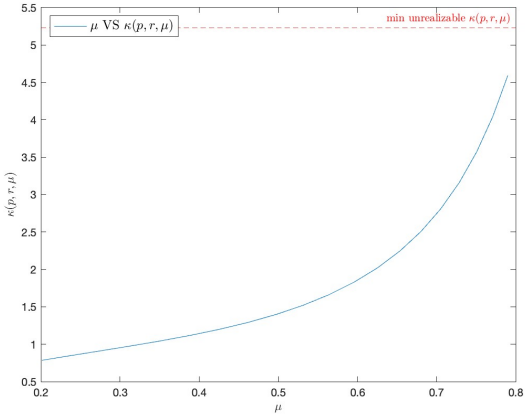
(a) $\lambda = 0.25$



(b) $\lambda = 0.5$



(c) $\lambda = 1$



(d) $\lambda = 1.5$

Figure 3.5: $\kappa(p, r, \mu)$ of λ-tri. real. VS $\min \kappa(p, r, \mu)$ of λ-tri. unreal. for different λ

Chapter 4

Algorithms for Low-Rank Plus Sparse Matrices Decomposition of Symmetric Matrices

In this chapter, we study algorithms for low-rank plus sparse matrices decomposition. In the first section, we analyze the *low-rank plus diagonal decomposition problem*, and show that this problem is NP-hard in general. However, when the optimal value of this problem is bounded above by an absolute constant, we can develop an algorithm to solve it in polynomial time. This section is an exposition of results in [31].

Then in the second section of this chapter, we show that when the optimal value of *low-rank plus tridiagonal matrices decomposition problem* is bounded by an absolute constant, the previous algorithm can be extended and applied to the problem and solves it in polynomial time.

4.1 Low-Rank Plus Diagonal Decomposition

Recall the *low-rank plus diagonal decomposition problem*, which is in the form: given $A \in \mathbf{S}^n$,

$$\begin{aligned} \min_{x \in \mathbf{R}^n, L \in \mathbf{S}^n} \ & \operatorname{rank}(L) \\ s.t. \ & A = L + \operatorname{Diag}(x) \\ & L \succeq 0, \end{aligned} \tag{LD1}$$

which can also be written as

$$\begin{aligned} \min_{x \in \mathbf{R}^n, L \in \mathbf{S}^n} \ & \operatorname{rank}(A + \operatorname{Diag}(x)) \\ s.t. \ & A + \operatorname{Diag}(x) \succeq 0 \end{aligned} \tag{LD2}$$

and we may assume $\operatorname{diag}(A) = 0$ for it.

Recall that for every simple undirected graph $G = ([n], E)$, $|E| = p$, we have a sparsity pattern and a linear map $\mathsf{SparseMat} : \mathbf{R}^n \times \mathbf{R}^p \to \mathbf{S}^n$. With that, we consider another optimization

problem with respect to the sparsity pattern:

$$\begin{aligned} \min_{u,v \in \mathbb{R}^n \ \mathbb{R}^p} \ & \mathrm{rank}(A + \mathrm{SparseMat}_G(u,v)) \\ s.t. \ & A + \mathrm{SparseMat}_G(u,v) \quad 0 \end{aligned} \tag{LS}$$

As shown in [31, Section 3], all these three problems, (LD1), (LD2) and (LS), are NP-hard.

Now we give an algorithm provided in [31], which can be implemented in polynomial time if the optimal value of an instance of (LD2) is $r = O(1)$. First, we have that there exists d such that $A + \mathrm{Diag}(d) \quad 0$ and $\mathrm{rank}(A + \mathrm{Diag}(d)) = r$ if and only if there exists $U \in \mathbb{R}^{n \ r}$ such that $rank(U) = r$ and $A + \mathrm{Diag}(d) = UU^{>}$. Notice that $\mathrm{rank}(U) = r$ if and only if there exists $J \quad [n]$ such that $|J| = r$ and $U_J := U(J,:)$ is invertible. Suppose such J exists, then we may assume $J = [r]$ and let $J := [n] \setminus J$. Then we can have

$$A + \mathrm{Diag}(d) = \begin{bmatrix} U_J U_J^{>} & U_J U_J^{>} \\ U_J U_J^{>} & U_J U_J^{>} \end{bmatrix}.$$

where $U_J U_J$ is positive definite. With this structure, we can consider the following lemma.

Lemma 4.1.1. [31] Given $n \quad 2$ be an integer, $r \in [n \quad 1]$ and $A \in \mathbb{S}^n$ with $\mathrm{diag}(A) = 0$. Then $d \in \mathbb{R}^n$ is feasible for (LD2) with objective value r if and only if there exists $J \quad [n]$ such that $|J| = r$, and with $J := [n] \setminus J$ the following system has a solution $(d, V) \in \mathbb{R}^n \quad \mathbb{S}^r$:

$$\begin{aligned} & [A(J,i) \quad A(J,j)]^{>} \mathrm{vec}(V) = A_{ij}, \forall i,j \in J, i < j \\ & [A(J,i) \quad A(J,j)]^{>} \mathrm{vec}(V) = d_i, \forall i \in J \\ & e_i^{>} V^{\ 1} e_j = A_{ij}, \forall i,j \in J, i < j (V^{\ 1})_{ii} = d_i, i \in J \\ & V \in \mathbb{S}^r_{++}. \end{aligned} \tag{4.1.1}$$

Given $J \quad [n]$ with $|J| = r$, every feasible solution (d, V) corresponding to J is completely characterized by (4.1.1). The following algorithm takes a given J and tries to solve (4.1.1) by only considering the constraints that are easy to handle.

Algorithm 1.1 [31] Linear solver with a given index set

Input: $A \in \mathbb{S}^n$, $\mathrm{diag}(A) = 0$, $J \subseteq [n]$

$\overline{J} := \{1, 2, \ldots, n\} \setminus J$, $r := |J|$, solve:

$$[A(J,i) \quad A(J,j)]^\top \mathrm{vec}(V) = A_{ij}, \forall i < j, i, j \in J \tag{4.1.2}$$

if (4.1.2) has no solution **then**
  return a certificate that either the original problem is infeasible or there may be a solution to the problem but for any solution, $U(J,:)$ is singular.
else if (4.1.2) has infinitely many solutions V and some v **then**
  return $B \in \mathbb{R}^{k \times r(r+1)/2}, b \in \mathbb{R}^k$ such that $\mathrm{rank}(B) = k$, V solves (4.1.2) if and only if $B\,\mathrm{svec}(V) = b$.
else if (4.1.2) has a unique solution V and some v **then**
  if $V \notin \mathbb{S}^r_{++}$ or $\exists i, j \in J, i < j, e_i^\top V^{-1} e_j \neq A_{ij}$ **then**
    return the corresponding certificate that either the original problem is infeasible ($\exists i, j \in J, i < j, e_i^\top V^{-1} e_j \neq A_{ij}$) or there may be a solution to the problem but for any solution, $U(J,:)$ is singular ($V \notin \mathbb{S}^r_{++}$).
  else
    Compute $u \in \mathbb{R}^n$ by $u_i := \begin{cases} (V^{-1})_{ii}, \forall i \in J \\ A(J,i)^\top V A(J,i), \forall i \in \overline{J} \end{cases}$
    return u
  end if
end if

The $B\,\mathrm{svec}(V) = b$ in the algorithm can be written as $[A(J,i) \overset{s}{\otimes} A(J,j)]\,\mathrm{svec}(V) = A_{ij}$ where $\overset{s}{\otimes} : \mathbb{R}^r \times \mathbb{R}^r \to \mathbb{R}^{r(r+1)/2}$ is symmetric Kronecker product and $\mathrm{svec} : \mathbb{S}^r \to \mathbb{R}^{r(r+1)/2}$ is defined as

$$\mathrm{svec}(V) = (V_{11}, \sqrt{2}V_{21}, \ldots, \sqrt{2}V_{r1}, V_{22}, \sqrt{2}V_{32}, \ldots, \sqrt{2}V_{n2}, \ldots, V_{nn})^\top$$

which returns the vectorized lower-triangular part of a symmetric matrix. When Algorithm 1.1 fails to solve (LD2) and returns a linear system $B\,\mathrm{svec}(V) = b$, the solution set of this returned linear system contains all solutions to (LD2) with the given J. And this guides us to the following algorithm.

Algorithm 1.2 [31] Nonlinear solver with a given index set

Input: $A \in \mathbb{S}^n$, $\mathrm{diag}(A) = 0$, $J := \{j_1, j_2, \ldots, j_r\} \subseteq [n]$, $B \in \mathbb{R}^{k \times r(r+1)/2}$, $b \in \mathbb{R}^k$ and $\mathrm{rank}(B) = k$.
solve:

$$\begin{cases} B\,\mathrm{svec}(V) = b \\ A_{ij}\det(V) \quad \mathrm{adj}(V)_{ij} = 0, \forall i < j, i, j \in J \\ \det(V_{J_k}) z_k^2 = 1, \forall k \in \{1, 2, \ldots, r\} \text{ where } J_k := \{j_1, j_2, \ldots, j_k\} \\ V \in \mathbb{S}^r, z \in \mathbb{R}^r \end{cases} \tag{4.1.3}$$

if (4.1.3) does not have a solution **then**
  return an infeasibility certificate
else
  given the solution (V, z) of (4.1.3)
  compute $u \in \mathbb{R}^n$ where $u_i := \begin{cases} (V^{-1})_{ii}, \forall i \in J \\ A(J,i)^\top V^{-1} A(J,i), \forall i \in J \end{cases}$
  return u
end if

where $A_{ij}\det(V) \quad \mathrm{adj}(V)_{ij} = 0$ is equivalent to $e_i^\top V^{-1} e_j = A_{ij}$ given $\det(V) \neq 0$.

With both Algorithm 1.1 and Algorithm 1.2, we can test every $J \subseteq [n]$ with $|J| = r$ to determine if (LD2) has a feasible solution with objective value r.

Algorithm 1.3 [31] Solver without given index sets

Input: $A \in \mathbb{S}^n, r \in [n-1]$
for every $J \subseteq [n]$ such that $|J| = r$ **do**
  Run Algorithm 1.1 with A and J, r. If Algorithm 1.1 returns a feasible solution, return the solution. Otherwise, if Algorithm 1.1 returns a linear system of equations, run Algorithm 1.2 with A, J, r and the returned linear system. If Algorithm 1.2 returns a feasible solution, return the solution.
end for

Since Algorithm 1.3 determines if (LD2) has a feasible solution with objective value r, when the optimal value of an instance is $r = O(1)$, we can solve (LD2) by enumerating $[r]$ with Algorithm 1.3.

Theorem 4.1.2. [31] If $r = O(1)$, then Algorithm 1.3 can be implemented in polynomial time. Hence, if an instance of (LD2) has optimal value $r = O(1)$, then it can be solved by calling Algorithm 1.3 with each possible rank from 1 to r.

4.2 Low-Rank Plus Tridiagonal Decomposition

In this section, we provide an algorithm which solves the low-rank plus tridiagonal decomposition problems in polynomial time when its optimal value is bounded by an absolute constant.

Consider the low-Rank and tridiagonal decomposition problem:

$$\begin{array}{l} \min \ \mathrm{rank}(L) \\ s.t.\ A = L + Y \\ \quad L \quad 0 \\ \quad Y \ 2 \ \mathsf{T}^n \setminus \mathsf{S}^n, \end{array} \tag{4.2.1}$$

which can be rewrite as the problem (2.3.1):

$$\begin{array}{l} \min \ \mathrm{rank}(A + Y) \\ s.t.\ A + Y \quad 0 \\ \quad Y \ 2 \ \mathsf{T}^n \setminus \mathsf{S}^n, \end{array}$$

and we may assume $\mathrm{diag}(A) = 0$ for it.

We extend the algorithms in the previous section to algorithms that solve (2.3.1) in polynomial time if the optimal value is bounded above by an absolute constant. Notice that for any $A \ 2 \ \mathsf{S}^n$ with $\mathrm{diag}(A) = 0$ and $A = U \ U^{>}$, we can set $u := \lambda_n(A)\mathbb{1} = \mathrm{diag}(\lambda_n UU^{>})$, $v := 0$ and have $\mathrm{rank}(A + \mathrm{TriDiag}(u,v)) \ n \ 1$. Notice (2.3.1) has a solution with objective value r if and only if there exists $U \ 2 \ \mathsf{R}^{n \ r}$ such that $\mathrm{rank}(U) = r$ and $UU^{>} = A + \mathrm{TriDiag}(u,v)$. Also, $\mathrm{rank}(U) = r$ if and only if there exists $J \ [n]$ such that $jJj = r$ and the submatrix $U_J := U(J,:) \ 2 \ \mathsf{R}^{r \ r}$ is nonsingular. If such J exists, without loss of generality, we can assume $J = [r]$ and $\overline{J} := [n] \setminus J$. Then we have the equation

$$A + \mathrm{TriDiag}(u,v) = \begin{bmatrix} U_J U_J^{>} & U_J U_{\overline{J}}^{>} \\ U_{\overline{J}}^{>} U_J & U_{\overline{J}} U_{\overline{J}}^{>} \end{bmatrix}.$$

Notice that, we have $[A + \mathrm{TriDiag}(u,v)]_{J,J} = U_J U_J^{>}$, which shows that $U_J^{>} = U_J^{\ 1}[A + \mathrm{TriDiag}(u,v)]_{J,J}$.

Lemma 4.2.1. [31] Let $n \ 2$ be an integer, $A \ 2 \ \mathsf{S}^n$ with $\mathrm{diag}(A) = 0$ and $r \ 2 \ [n \ 1]$ be given. $(u,v) \ 2 \ \mathsf{R}^n \ \mathsf{R}^{n \ 1}$ is a feasible solution of (2.3.1) with objective value r if and only if there exists $J \ [n]$, $\overline{J} := [n] \setminus J$ with $jJj = r$, such that the following system has a solution

$(u, v, V) \in \mathbb{R}^n \times \mathbb{R}^{n-1} \times \mathbb{S}^r$:

$$
\begin{aligned}
& K = A + \mathrm{TriDiag}(0, v) \\
& [K(J,i) \quad K(J,j)]^\top \mathrm{vec}(V) = A_{ij}, \forall i, j \in \overline{J}, i < j \quad -1 \\
& [K(J,i) \quad K(J,i+1)]^\top \mathrm{vec}(V) = v_i + A_{i(i+1)}, \forall i, i+1 \in \overline{J} \\
& [K(J,i) \quad K(J,i)]^\top \mathrm{vec}(V) = u_i, \forall i \in \overline{J} \\
& e_i^\top V^{-1} e_j = A_{ij}, \forall i, j \in J, i < j \quad -1 \\
& e_i^\top V^{-1} e_{i+1} = A_{ij} + v_i, \forall i, i+1 \in J \\
& (V^{-1})_{ii} = u_i, i \in J \\
& V \in \mathbb{S}^r_{++}.
\end{aligned}
\tag{4.2.2}
$$

Proof. Consider a feasible solution $(u, v) \in \mathbb{R}^n \times \mathbb{R}^{n-1}$ for the problem (2.3.1) with objective value $r \in [n-1]$. Then, there exists $U \in \mathbb{R}^{n \times r}$ with $\mathrm{rank}(U) = r$ such that $UU^\top - \mathrm{TriDiag}(u, v) = A$. Consider $J \subseteq [n]$ such that $\mathrm{rank}(U_J) = r = |J|$. Let $V := (U_J U_J^\top)^{-1} \in \mathbb{S}^r_{++}$. Consider $i, j \in \overline{J}, i < j \le n-1$, then $[K(J,i) \quad K(J,j)]^\top \mathrm{vec}(V) = \mathrm{tr}(V, K(J,j)K(J,i)^\top) = K(J,i)^\top U_J^{\top} U_J^{-1} K(J,j) = U(i,:)U(j,:)^\top = A_{ij}$. Other equations are satisfied similarly.

Conversely, for some given J as in the statement and (u, v, V) feasible for (4.2.2), without loss of generality, we assume that $J = [r]$. Compute $\hat{U} \in \mathbb{R}^{r \times r}$ from $V^{-1} = \hat{U}\hat{U}^\top \succ 0$. Let $K := A + \mathrm{TriDiag}(0, v)$ and $U \in \mathbb{R}^{n \times r}$ defined as

$$
U(i,:) := \begin{cases} \hat{U}(i,:), & \text{if } i \in J \\ [\hat{U}^{-1} K(J,i)]^\top, & \text{if } i \in \overline{J} \end{cases}
$$

where $\overline{J} := [n] \setminus J$. For $i, j \in \overline{J}, i < j \le n-1$, we have $(UU^\top)_{ij} = [\hat{U}^{-1} K(J,i)]^\top [\hat{U}^{-1} K(J,j)] = K(J,i)^\top \hat{U}^{-\top} \hat{U}^{-1} K(J,j) = K(J,i)^\top V K(J,j) = [K(J,i) \quad K(J,j)]^\top \mathrm{vec}(V) = K_{ij} = A_{ij}$. Apply similar arguments for other equations, we find $\succ \quad UU^\top = A + \mathrm{TriDiag}(u, v)$. That is, (u, v) is feasible for (2.3.1) with objective value r. □

Similar to the diagonal case, now we give an algorithm which solves (4.2.2) by only considering the constraints that are easy to handle.

Algorithm 2.1 quadratic solver with a given index set

Input: $A \in \mathbb{S}^n$, $\mathrm{diag}(A) = 0$, $J \subseteq [n]$, $r := |J|$
$\overline{J} := [n] \setminus J$, $J_+ := \{i \in [n] : i+1 \notin J \text{ and } i \neq 1 \notin J\}$, $J_- := [n] \setminus J$, solve:

$$\begin{aligned} &K = A + \mathrm{TriDiag}(0, v) \\ &[K(J,i) \quad K(J,j)]^\top \mathrm{vec}(V) = K_{ij} = A_{ij}, \forall i,j \in \overline{J} \setminus J, i < j \quad 1 \end{aligned} \tag{4.2.3}$$

if (4.2.3) has no solution **then**
  return a certificate that either the original problem is infeasible or there may be a solution to the problem but for any solution, $U(J,:)$ is singular.
else if (4.2.3) has infinitely many solutions V and some v **then**
  return $B \in \mathbb{R}^{k \times r(r+1)/2}, b \in \mathbb{R}^k$ such that $\mathrm{rank}(B) = k$, V solves (4.2.3) if and only if $B\,\mathrm{svec}(V) = b$.
else if (4.2.3) has a unique solution V and some v **then**
  if $V \notin \mathbb{S}^r_{++}$ or $\exists i,j \in J, i < j \quad 1$, $e_i^\top V^{-1} e_j < K_{ij} = A_{ij}$ **then**
    return the corresponding certificate that either the original problem is infeasible ($\exists i,j \in J, i < j \quad 1$, $e_i^\top V^{-1} e_j < A_{ij}$) or there may be a solution to the problem but for any solution, $U(J,:)$ is singular ($V \notin \mathbb{S}^r_{++}$).
  else
    Solve:

$$\begin{aligned} &[K(J,i) \quad K(J,j)]^\top \mathrm{vec}(V) = A_{ij}, \forall i,j \in J, i < j \quad 1 \text{ and at least one of } i,j \text{ is in } J \\ &[K(J,i) \quad K(J,i+1)]^\top \mathrm{vec}(V) = v_i + A_{i(i+1)}, \forall i \in J, i+1 \in \overline{J} \\ &[K(J,i) \quad K(J,i \quad 1)]^\top \mathrm{vec}(V) = v_i + A_{i(i+1)}, \forall i \in J, i \quad 1 \in \overline{J}. \end{aligned} \tag{4.2.4}$$

    if the quadratic system above has no solution v **then**
      return Infeasibility Certificate
    else
      Compute $u \in \mathbb{R}^n$ by $u_i := \begin{cases} (V^{-1})_{ii}, \forall i \in J \\ K(J,i)^\top V K(J,i), \forall i \in J \end{cases}$
      and compute the other entries of $v \in \mathbb{R}^{n-1}$ by

$$v_i := \begin{cases} (V^{-1})_{i(i+1)} \quad A_{i(i+1)}, \forall i, i+1 \in J \\ K(J,i)^\top V K(J,i+1) \quad A_{i(i+1)}, \forall i, i+1 \in J \end{cases}$$

    end if
  end if
  return (u, v)
end if

The algorithm first solves a linear system of V. By considering $i,j \in J \setminus J_-$, none of $K(J,i), K(J,j)$ contains v_i, hence given $i,j \in \overline{J} \setminus J_-, i < j \quad 1$, the equation $[K(J,i) \quad K(J,j)]^\top \mathrm{vec}(V) = K_{ij} = A_{ij}$ is a linear system.

Depending on the uniqueness of V, the solution to the original problem is computed. For the case where infinitely many V exist, the algorithm returns a quadratic system whose solution set contains all solutions of (2.3.1), (4.2.3) with respect to the given index set J, so we proceed to another phase to solve a system of polynomial equations. When the V is unique, the algorithm tests if the V is feasible for the other linear conditions and finds a feasible v by first determining v_i, v_j where $i \in J, i+1 \in J, i \in J, i-1 \in J$ and $i, j \in J, i < j-1$ with at least one of i, j being in J . Since there are at most $2r$ of i such that $i \in J$ and $i-1 \in J$ and $|J| \leq 2r$, the cubic system (4.2.4) has at most $2r + (2r)^2 \in O(r^2)$ equations. Notice that there are at most $2r$ of (i, j) such that $i \in J, j \in J$, and $\mathsf{vec}(V)$ has r^2 entries, so (4.2.4) has at most $2r + r^2 \in O(r^2)$ variables. The motivation is that if we write a feasible solution as

$$A + \mathsf{TriDiag}(u, v) = UU^\top,$$

then V determines U_J and those entries of v determines U_J. Thus, $UU^\top$ is determined and so are u and the rest entries of v.

If there are infinitely many V, then we apply similar steps as the diagonal cases. $[K(J, i) \; K(J, j)]^\top \mathsf{vec}(V) = K_{ij}$ is equivalent to $[K(J, i) \overset{s}{-} K(J, j)]\,\mathsf{svec}(V) = K_{ij}$. Then $B\,\mathsf{svec}(V) = b$ may be written as a quadratic system of symmetric Kronecker products and $\mathsf{svec}(V)$. Given $\det(V) \neq 0$, $K_{ij} \det(V) - \mathsf{adj}(V(i, j)) = 0$ can be written as $V_{ij}^{-1} = K_{ij}$. Also $\det(V_{J_k}) z_k^2 = 1$ and $V \in \mathsf{S}^r, z \in \mathsf{R}^r$ is equivalent to $V \in \mathsf{S}^r_{++}$.

Algorithm 2.2 Nonlinear solver with a given index set

Input: $K \in \mathbb{S}^n$, $\mathrm{diag}(K) = 0$, $J := \{j_1, j_2, \ldots, j_r\} \subseteq [n]$, $B \in \mathbb{R}^{k \times r(r+1)/2}$, $b \in \mathbb{R}^k$ and $\mathrm{rank}(B) = k$.
solve:

$$\begin{cases} B\,\mathrm{svec}(V) = b \\ K_{ij} \bullet \det(V) \cdot \mathrm{adj}(V)_{ij} = 0, \forall i < j, i, j \in J \\ \det(V_{J_k}) z_k^2 = 1, \forall k \in [r] \text{ where } J_k := \{j_1, j_2, \ldots, j_k\} \\ [K(J,i) \quad K(J,j)]^\top \mathrm{vec}(V) = A_{ij}, \forall i, j \in J, i < j \quad 1 \text{ and at least one of } i, j \text{ is in } J \\ [K(J,i) \quad K(J,i+1)]^\top \mathrm{vec}(V) = v_i + A_{i(i+1)}, \forall i \in J, i+1 \in \overline{J} \\ [K(J,i) \quad K(J,i \quad 1)]^\top \mathrm{vec}(V) = v_i + A_{i(i+1)}, \forall i \in J, i \quad 1 \in \overline{J} \\ V \in \mathbb{S}^r, z \in \mathbb{R}^r \end{cases} \tag{4.2.5}$$

if (4.2.5) does not have a solution **then**
  return an infeasibility certificate
else
  given the solution (V, z) of (4.2.5)
  compute $u \in \mathbb{R}^n$ where $u_i := \begin{cases} (V \quad ^1)_{ii}, \forall i \in J \\ K(J,i)^\top V K(J,i), \forall i \in J \end{cases}$
  compute the other entries of $v \in \mathbb{R}^{n \quad 1}$ where $v \in \mathbb{R}^{n \quad 1}$ by
  $v_i := \begin{cases} (V \quad ^1)_{i(i+1)} \quad A_{i(i+1)}, \forall i, i+1 \in J \\ K(J,i)^\top V K(J,i+1) \quad A_{i(i+1)}, \forall i, i+1 \in J \end{cases}$
  return u
end if

Similar to Algorithm 2.1, we first determine U_J, then $U_{\overline{J}}$, then we know u and some entries of v. For the rest of v, we determine them by considering $U_J U_{\overline{J}}^\top$. Also, since the system (4.2.4) contains $O(r^2)$ cubic equations with $O(r^2)$ variables, we know (4.2.5) also contains $O(r^2)$ cubic equations with $O(r^2)$ variables.

If we assume $r \in O(1)$, then for both Algorithm 2.1 and Algorithm 2.2, we are solving systems of equations with a degree at most 3 while the number of non-linear equations and the number of variables are $O(1)$. For most applications, Algorithm 2.1 can be implemented efficiently, because except for $O(1)$ cubic equations, it only solves linear systems. Thus, in some applications, running Algorithm 2.1 for different J might be worthwhile before moving to Algorithm 2.2 as if Algorithm 2.1 returns a solution with $|J| = r$, then there is no need to run Algorithm 2.2 for this r.

Since when $r = O(1)$, Algorithm 2.2 involves at most $O(1)$ number of polynomial equations, Algorithm 2.3 below can determine whether (2.3.1) has a solution with objective value r in polynomial time. And if the optimal objective value is $r = O(1)$, we can solve the problem by enumerating all possible values for $r \in [r]$.

Algorithm 2.3 Solver without given index sets

Input: $A \in \mathbb{S}^n, r \in [n \quad 1]$

for every $J \quad [n]$ such that $|J| = r$ **do**

Run Algorithm 2.1 with A and J, r. If Algorithm 2.1 returns a feasible solution, return the solution. Otherwise, if Algorithm 2.1 returns a linear system of equations, run Algorithm 2.2 with A, J, r and the returned linear system. If Algorithm 2.2 returns a feasible solution, return the solution.

end for

Now we prove that we can solve a low-rank plus tridiagonal decomposition instance when it has an optimal value $r = O(1)$.

Theorem 4.2.2. If $r = O(1)$, then Algorithm 2.3 can be implemented to run in polynomial time. Thus, if the optimal objective value of (2.3.1) is $r = O(1)$, then such instances can be solved in polynomial time by trying all possible ranks from 1 to r with Algorithm 2.3.

Proof. Algorithm 2.1 can be run in polynomial time, because it only involves solving a quadratic system of equations whose number of equations is bounded by a $poly(n)$ number and performing a Cholesky decomposition on a matrix $V \in \mathbb{S}^r_{++}$. If $r = O(1)$, then the system (4.2.5) is a system of polynomial equations with $O(r^2) = O(1)$ variables and $O(r^2) + O(r) + O(2r) = O(1)$ equations, so Algorithm 2.2 can be run in $O(1)$ time (e.g. by cylindrical algebraic decomposition, for instance, see [2] and [5]). By running with all possible ranks $[r]$, Algorithm 2.3 calls Algorithm 2.1 and Algorithm 2.2 at most

$$\sum_{r=1}^{r} \binom{n}{r} = O(n^r) = O(n^{O(1)})$$

times. Hence, if $r = O(1)$, we can solve (2.3.1) in polynomial time. □

Example 4.2.3. Consider the matrix

$$A := \begin{bmatrix} 0 & 0 & 1 & 2 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 2 & 1 \\ 2 & 1 & 2 & 0 & 2 \\ 1 & 1 & 1 & 2 & 0 \end{bmatrix},$$

and consider the corresponding instance of (TriRegP). Notice that, we consider the submatrix $A(1:2, 4:5)$ of A, whose entries are not affected by the tridiagonal perturbation. This submatrix has rank 2, so $\mathrm{rank}(A + \mathrm{TriDiag}(u, v)) \quad 2$ for every $(u, v) \in \mathbb{R}^n \quad \mathbb{R}^{n-1}$. Consider $u = [1, 1, 1, 5, 2]^\top, v = [0, 0, 0, 1]^\top$, then $\mathrm{rank}(A + \mathrm{TriDiag}(u, v)) = 2$, because

$$A + \mathrm{TriDiag}(u, v) = \begin{bmatrix} 1 & 0 & 1 & 2 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 2 & 1 \\ 2 & 1 & 2 & 5 & 3 \\ 1 & 1 & 1 & 3 & 2 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 2 \\ 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 1 & 2 & 1 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 1 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 & 1 & 1 \end{bmatrix},$$

so (u, v) is an optimal solution.

Consider $J = \{1, 2\}$, $J = \{3, 4, 5\}$ and run Algorithm 2.1. The linear system (4.2.3) considers $K = A + \text{TriDiag}(0, v)$. Thus there are infinitely many feasible V because the system does not have a full column rank. The system (4.2.5) considers $[K(J, 3) \quad K(J, 5)]^\top \text{vec}(V) = A_{35}$ and $[K(J, 2) \quad K(J, 3)]^\top \text{vec}(V) = v_2 + A_{23}$, and the system can be written as: replacing K_{12} by v_1 since $A_{12} = 0$, we have

$$
\begin{aligned}
& v_1 \det(V) \quad \text{adj}(V)_{12} = 0 \\
& V_{11} z_1^2 = 1 \\
& (V_{11} V_{22} \quad V_{12}^2) z_2^2 = 1 \\
& [1 \; v_1 \; 1 \; v_1] \, \text{vec}(V) = 1 \\
& [v_1 \; v_1 v_2 \; 0 \; 0] \, \text{vec}(V) = v_2 + A_{23} = v_2 \\
& V \in \mathbb{S}^r, z \in \mathbb{R}^r.
\end{aligned}
$$

One solution to (4.2.3) is

$$
V = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, z = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, v_1 = v_2 = v_3 = 0, v_4 = 1.
$$

In this way, the u and v we compute is

$$
u = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 5 \\ 2 \end{bmatrix}, v = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix},
$$

which is exactly the optimal solution we had before.

Beyond such special cases above, in general, we do not know how to construct a polynomial time algorithm because we have seen that such problems are NP-hard.

Chapter 5

Generalization, Conclusion and Future Research

In this chapter, we first write our relaxation of the low-rank plus tridiagonal problem, (TriRegP), in conic form and discuss how the optimal conditions we developed in the previous chapters can be related to the conic form. Then, we discuss the general low-rank plus sparsity pattern decomposition problem. Finally, we give a conclusion for the thesis and give some future research directions.

5.1 Convex Programming in Conic Form and General Low-Rank plus Sparsity Pattern Decomposition

In previous chapters, we present the low-rank plus tridiagonal problems defined over positive semidefinite cone. In this section, we write low-rank plus tridiagonal problems as general convex programming in conic form. Then, we show the optimality conditions of the low-rank plus tridiagonal problem are equivalent to the optimality conditions of the ones in conic form.

Consider $\ell(\hat{t}, v) = b^\top v + c\hat{t}$: given $b \in \mathbb{R}^n, A \in \mathbb{S}^n$,

$$\begin{aligned} \min\ & a^\top u + \lambda(b^\top v + c\hat{t}) \\ & \mathrm{TriDiag}(u, v) + A \succeq 0 \\ & (u, (v, \hat{t})) \in \mathcal{K} \end{aligned} \tag{ConePrimal}$$

where $\mathcal{K} := \mathbb{R}^n \times \mathcal{K}$ and $\mathcal{K} := \left\{ \begin{pmatrix} v \\ t \end{pmatrix} : \|v\|_p \le \hat{t} \right\}, p \in [1, \infty]$. And its dual problem is defined as

$$\begin{aligned} \max\ & \langle A, X \rangle \\ s.t.\ & \mathrm{diag}(X) = a \\ & (\lambda b - \mathrm{bidiag}(X), \lambda c) \in \mathcal{K}^* \\ & X \succeq 0. \end{aligned} \tag{ConeDual}$$

Notice in the primal problem, the linear map we consider is $\mathcal{A}(u,(v,\hat{t})) := \mathrm{TriDiag}(u,v)$. So we have $\langle \mathcal{A}^*(X),(u,(v,\hat{t}))\rangle = \langle X,\mathcal{A}(u,(v,\hat{t}))\rangle = \langle \mathrm{tridiag}(X),(u,v)\rangle$. Thus, $\mathcal{A}^*(X) = (\mathrm{tridiag}(X),0) = (\mathrm{diag}(X),(\mathrm{bidiag}(X),0))$. Hence, when we consider the cone programming, (ConePrimal) can be written as

$$\begin{aligned} \min\ & a^\top u + \lambda(b^\top v + c\hat{t}) \\ & \mathcal{A}(u,(v,\hat{t})) + A \in \mathbb{S}^n_+ \\ & (u,(v,\hat{t})) \in \mathbb{R}^n \times \mathcal{K} \end{aligned}$$

and the dual is

$$\begin{aligned} \max\ & \langle A, X\rangle \\ s.t.\ & \mathcal{A}^*(X) \in \{0\}^* \times \mathcal{K}^* - (\lambda b, \lambda c) \\ & X \in \mathbb{S}^n_+. \end{aligned}$$

Thus, if we consider the tridiagonal perturbation problem, we have $b = 0$, $c = 1$ and $\mathcal{K} = \left\{ \begin{pmatrix} v \\ t \end{pmatrix} : \|v\|_1 \le \hat{t} \right\}$. Then, the dual problem can be written as

$$\begin{aligned} \max\ & \langle A, X\rangle \\ s.t.\ & \mathrm{diag}(X) = \mathbb{1} \\ & (\mathrm{bidiag}(X),\lambda) \in \mathcal{K}^* \\ & X \succeq 0. \end{aligned}$$

where $\mathcal{K}^* = \left\{ \begin{pmatrix} v \\ t \end{pmatrix} : \|v\|_\infty \le \hat{t} \right\}$.

Notice here we require ℓ to be a linear function but it can be replaced by any function satisfying $f(\beta\hat{t},\beta v) \le \beta f(\hat{t},v)$ for $\beta \ge 0$, because we can bound $f(\hat{t},v)$ by a new variable α and modify the cone $\mathcal{K}$ as $\left\{ \begin{pmatrix} v \\ \hat{t} \\ \alpha \end{pmatrix} : \|v\|_p \le \hat{t}, f(\hat{t},v) \le \alpha \right\}$.

The optimal conditions from conic programming still hold, for our tridiagonal perturbation setting, we can write the optimality condition as $(u,(v,\hat{t})),(X,\lambda)$ being both feasible and

$$\langle A + \mathrm{TriDiag}(u,v), X\rangle + \lambda t + v^\top(b - \mathrm{bidiag}(X)) = 0$$

which is equivalent to the system

$$\begin{aligned} & \langle A + \mathrm{TriDiag}(u,v), X\rangle = 0 \\ & \lambda\hat{t} + v^\top(0 - \mathrm{bidiag}(X)) = 0. \end{aligned} \tag{5.1.1}$$

Corollary 5.1.1. The conditions (5.1.1) are equivalent to the optimality conditions of (TriRegP) from the Theorem 1.1.3.

Proof. For (u, v, t) feasible for (TriRegP) and (X, ξ, ω) feasible (TriRegD), the optimality conditions from the Theorem 1.1.3 are

$$\begin{aligned} &\langle A + \mathrm{TriDiag}(u, v), X \rangle = 0 \\ &(t \quad v)^\top \xi = 0 \\ &(t + v)^\top \omega = 0. \end{aligned}$$

We can sum up the last two equations and get

$$t^\top(\xi + \omega) + v^\top(\omega \quad \xi) = \lambda t^\top \mathbb{1} \quad v^\top \mathrm{bidiag}(X) = 0$$

and for our settings, $\hat{t}$ is equivalent to $t^\top \mathbb{1}$. □

In addition to generalizing the cone the problems are defined over, we can also generalize the sparse matrix in the decomposition. We now consider the low-rank plus sparse decomposition problem with a general sparsity pattern. Given a sparsity pattern $G = ([n], E)$, $|E| = m$, we have a general low-rank plus sparse G decomposition problem:

$$\begin{aligned} &\inf \mathrm{rank}(L) \\ &s.t.\ L + \mathrm{SparseMat}_G(u, v) = A \\ &\quad L \quad 0. \end{aligned}$$

Similar to the low-rank plus tridiagonal problem, we may relax the rank function as the nuclear norm, and put regularizations on the entries v_i representing the edges. Instead of bounding $|v_i|$ by t_i, we consider a constraint $(v, \hat{t}) \in \mathcal{K}$, where $\mathcal{K}$ is a general cone defined by $\mathcal{K} := \left\{ \begin{pmatrix} v \\ t \end{pmatrix} : \|v\|_p \quad \hat{t} \right\}$, $p \in [1, \infty]$. After replacing $L = A \quad \mathrm{SparseMat}_G(u, v) \quad 0$ by

$$\mathrm{SparseMat}_G(u, v) \quad A$$

and replacing $\mathrm{rank}(L) = \mathrm{rank}(A \quad \mathrm{SparseMat}_G(u, v))$ by

$$\|A \quad \mathrm{SparseMat}_G(u, v)\|_* = \mathrm{tr}(A \quad \mathrm{SparseMat}_G(u, v)) = \mathrm{tr}(\quad \mathrm{SparseMat}_G(u, v)) = \quad \mathbb{1}^\top u,$$

we replace u, v by $\quad u, \quad v$ and add regularizations on $\quad u, \quad v$. Then, we have the following relaxation: given $a \in \mathbb{R}^n$, $b \in \mathbb{R}^m$, $\lambda \in \mathbb{R}$,

$$\begin{aligned} &\inf\ a^\top u + b^\top v + \lambda \hat{t} \\ &s.t.\ \mathrm{SparseMat}_G(u, v) \quad A \\ &\quad (u, (v, \hat{t})) \in \mathbb{R}^n \quad \mathcal{K}. \end{aligned} \tag{SparseMat}$$

If we assume $\mathrm{diag}(A) = 0$, then $\mathbb{R}^n$ can be changed to $\mathbb{R}^n_+$.

Definition 5.1.2. We say a simple graph G is *chordal* if every cycle of length at least four has a chord. And we say a simple graph G is *homogeneous chordal* if it is chordal and it does not contain a path of length four as an induced subgraph.

When we have $p := n \quad 1$, and $\mathsf{SparseMat}_G(\ ,\ ) := \mathsf{TriDiag}(\ ,\ )$, G is a chordal graph because it is a graph which is a path. A more detailed discussion about chordal sparsity patterns and the optimization problem related to it can be seen in [30].

5.2 Conclusion and Future Research

In this thesis, we studied the low-rank plus sparse matrices decomposition. We have seen the low-rank plus diagonal matrices decomposition problem and how one of its semidefinite programming relaxations describes the diagonal recoverability, realizability and ellipsoid fitting property of a subspace of R^n [22]. We have also seen that low-rank plus diagonal matrices decomposition problem is NP-hard, but when it has an optimal objective value $r = O(1)$, there exists an algorithm which solves the problem in polynomial time [31].

We introduced the low-rank plus tridiagonal matrices decomposition problem and one of its semidefinite programming relaxations. We proposed relaxations with and without a linear regularization on the bidiagonal entries. We showed that when there is no regularizations or the penalty parameter $\lambda \not< 2$, the relaxed problems have unique optimal solutions. In particular, when $\lambda > 2$, we showed that the optimal solution is equivalent to the optimal solution of a low-rank plus diagonal decomposition with the same input matrix. We also proposed λ-tridiagonal recoverability, realizability and ellipsoid fitting property and showed that they are equivalent to diagonal recoverability, realizability, and ellipsoid fitting property respectively when $\lambda > 2$. By considering the coherence of a subspace, we gave a sufficient condition for λ-tridiagonal realizability. Although we did not prove the NP-hardness of the general low-rank plus tridiagonal matrices decomposition problem, we developed an algorithm which solved the problem in polynomial time when it has an optimal value $r = O(1)$.

There are some open questions relating to our results and would be of interest for future research:

1. For problem (TriRegP), we used a linear objective function with a penalty parameter on the absolute values of bidiagonal entries. Can we replace this objective function with more general functions? In particular, can we replace it with other norms (like in (SparseMat)) or other general convex functions and extend the properties like realizability, and uniqueness of optimal solutions to those cases?

2. In this thesis, we introduced the low-rank plus tridiagonal decomposition problem, and analyzed its optimality conditions and different properties. Can we apply similar analyses and expect results from more general sparsity patterns? For example, if we change tridiagonal matrices to matrices with a chordal sparsity pattern, we would expect more general results because tridiagonal matrices represent a chordal sparsity pattern, but what are we gaining by having more freedom on the sparsity pattern?

References

[1] P.M Bentler. A lower-bound method for the dimension-free measurement of internal consistency. *Social Science Research*, 1(4):343–357, 1972.

[2] Christopher W. Brown and James H. Davenport. The complexity of quantifier elimination and cylindrical algebraic decomposition. In *Proceedings of the 2007 International Symposium on Symbolic and Algebraic Computation*, ISSAC '07, page 54–60, New York, NY, USA, 2007. Association for Computing Machinery.

[3] Emmanuel J. Candès and Benjamin Recht. Exact matrix completion via convex optimization. *CoRR*, abs/0805.4471, 2008.

[4] Venkat Chandrasekaran, Sujay Sanghavi, Pablo A. Parrilo, and Alan S. Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011.

[5] George E. Collins. Quantifier elimination for real closed fields by cylindrical algebraic decompostion. In H. Brakhage, editor, *Automata Theory and Formal Languages*, pages 134–183, Berlin, Heidelberg, 1975. Springer Berlin Heidelberg.

[6] Marcel K. de Carli Silva and Levent Tunçel. Strict complementarity in semidefinite optimization with elliptopes including the MaxCut SDP. *SIAM Journal on Optimization*, 29(4):2650–2676, 2019.

[7] Giacomo Della Riccia and Alexander Shapiro. Minimum rank and minimum trace of covariance matrices. *Psychometrika*, 47(4):443–448, December 1982.

[8] Charles Delorme and Svatopluk Poljak. Combinatorial properties and the complexity of a max-cut approximation. *European Journal of Combinatorics*, 14(4):313–333, 1993.

[9] Maryam Fazel. *Matrix Rank Minimization with Applications*. PhD thesis, Stanford University, 2002.

[10] Delbert Ray Fulkerson and Oliver Alfred Gross. Incidence matrices and interval graphs. *Pacific Journal of Mathematics*, 15:835–855, 1965.

[11] Michael R. Garey and David S. Johnson. *Computers and Intractability; A Guide to the Theory of NP-Completeness*. Series of books in the mathematical science. W. H. Freeman & Co., USA, 1990.

[12] Michel X. Goemans and David P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM*, 42(6):1115–1145, November 1995.

[13] Gene H. Golub and Charles F. Van Loan. *Matrix Computations (3rd Ed.)*. Johns Hopkins University Press, USA, 1996.

[14] Robert Grone, Charles R. Johnson, Eduardo M. Sá, and Henry Wolkowicz. Positive definite completions of partial Hermitian matrices. *Linear Algebra and its Applications*, 58:109–124, 1984.

[15] Robert Grone, Stephen Pierce, and William Watkins. Extremal correlation matrices. *Linear Algebra and its Applications*, 134:63–70, 1990.

[16] Bernd Gärtner and Jiří Matoušek. *Approximation algorithms and semidefinite programming*. Springer, Heidelberg, 2012.

[17] Mehdi Karimi and Levent Tunçel. Domain-driven solver (DDS) version 2.0: a matlab-based software package for convex optimization problems in domain-driven form, 2019. arxiv:1908.03075.

[18] Walter Ledermann. I.—On a problem concerning matrices with variable diagonal elements. *Proceedings of the Royal Society of Edinburgh*, 60(1):1–17, 1940.

[19] Nathan Linial, Eran London, and Yuri Rabinovich. The geometry of graphs and some of its algorithmic applications. *Combinatorica*, 15(2):215–245, June 1995.

[20] B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24(2):227–234, April 1995.

[21] Benjamin Recht, Maryam Fazel, and Pablo A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, Jan 2010.

[22] J. Saunderson, V. Chandrasekaran, P. A. Parrilo, and A. S. Willsky. Diagonal and low-rank matrix decompositions, correlation matrices, and ellipsoid fitting. *SIAM Journal on Matrix Analysis and Applications*, 33(4):1395–1416, 2012.

[23] J. Schur. Bemerkungen zur theorie der beschränkten bilinearformen mit unendlich vielen veränderlichen. *Journal für die reine und angewandte Mathematik*, 140:1–28, 1911.

[24] Alexander Shapiro. Statistical inference of semidefinite programming. *Mathematical Programming*, 174(1):77–97, March 2019.

[25] Yaroslav Shitov. How hard is the tensor rank?, 2021. viXra:2107.0049.

[26] Anthony Man-Cho So. *A Semidefinite Programming Approach to the Graph Realization Problem: Theory, Applications and Extensions*. PhD thesis, Stanford University, 2007.

[27] Anthony Man-Cho So and Yinyu Ye. Theory of semidefinite programming for Sensor Network Localization. *Mathematical Programming*, 109(2):367–384, March 2007.

[28] C. Spearman. "General intelligence," objectively determined and measured. *The American Journal of Psychology*, 15(2):201–292, 1904.

[29] Levent Tunçel. *Polyhedral and semidefinite programming methods in combinatorial optimization*, volume 27 of *Fields Institute monographs*. American Mathematical Society, Providence, Rhode Island, 2010.

[30] Levent Tunçel and Lieven Vandenberghe. Linear optimization over homogeneous matrix cones. *Acta Numerica*, 32:675–747, 2023.

[31] Levent Tunçel, Stephen A. Vavasis, and Jingye Xu. Computational complexity of decomposing a symmetric matrix as a sum of positive semidefinite and diagonal matrices, 2022. arXiv:2209.05678.

[32] Jerry A. Walters. Nonnegative matrix equations having positive solutions. *Mathematics of Computation*, 23(108):827–827, 1969.

[33] Wikipedia contributors. Netflix prize — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.phptitle=Netflix_Prize&oldid=1140748928, 2023. [Online; accessed 20-April-2023].

[34] Alp Yurtsever, Joel A. Tropp, Olivier Fercoq, Madeleine Udell, and Volkan Cevher. Scalable semidefinite programming. *SIAM Journal on Mathematics of Data Science*, 3(1):171–200, 2021.